

# Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games

Jobst Heitzig\*, Kai Lessmann\*, and Yong Zou\*

\*Potsdam Institute for Climate Impact Research, P. O. Box 60 12 03, 14412 Potsdam, Germany

Final authors' draft as accepted for publication in PNAS on July 25, 2011. Published version: DOI:10.1073/pnas.1106265108

As the Copenhagen Accord indicates, most of the international community agrees that global mean temperature should not be allowed to rise more than two degrees Celsius above pre-industrial levels to avoid unacceptable damages from climate change. The scientific evidence distilled in the IPCC's 4th Assessment Report and recent reports by the U.S. National Academies shows that this can only be achieved by vast reductions of greenhouse gas (GHG) emissions.

Still, international cooperation on GHG emissions reductions suffers from incentives to free-ride and to renegotiate agreements in case of non-compliance, and the same is true for other so-called 'public good games.' Using game theory, we show how one might overcome these problems with a simple dynamic strategy of Linear Compensation (LinC) when the parameters of the problem fulfill some general conditions and players can be considered to be sufficiently rational.

The proposed strategy redistributes liabilities according to past compliance levels in a proportionate and timely way. It can be used to implement any given allocation of target contributions, and we prove that it has several strong stability properties.

greenhouse gas emissions | free-riding | compliance | renegotiation | strategy | compensation

In many situations of decision-making under conflicting interests, including the management of natural resources (1), game theory – the study of rational behaviour in situations of conflict – proves to be a useful analysis tool. Using its methods, we provide in this article a partial solution for the cooperation problem in a class of so-called public good games: If a number of players repeatedly contribute some quantity of a public good, how can they make sure everyone cooperates to achieve a given optimal level of contributions? The main application we have in mind are international efforts to mitigate climate change. There the players are countries and the corresponding public good is the amount of GHG emissions they abate as compared to a reference scenario (e.g., a 'business as usual' emissions path). The existing literature on the emissions problem stresses the fact that only international agreements which contain sufficient incentives for participation and compliance can lead to substantive cooperation (2; 3), and game theory is a standard way of analysing the strategic behavior of sovereign countries under such complex incentive structures. While earlier game-theoretic studies have been mainly pessimistic about the likelihood of cooperation (4–19), our results show that with emissions trading and a suitable strategy of choosing individual emissions, high levels of cooperation might be achieved.

The general situation is modeled here as a repeated game played in a sequence of periods, with a continuous control variable (e.g., emissions reductions) that can take on any value in principle. We focus on the case where the marginal costs of contributing to the public good are the same for all players. This is, e.g., the case if there is an efficient market for contributions (24; 25).

We show that players can ensure compliance with a given initially negotiated target allocation of contributions by adopting a certain simple dynamic strategy to choose their actual contributions over time. In each period, the allocation of liabilities is redistributed in-

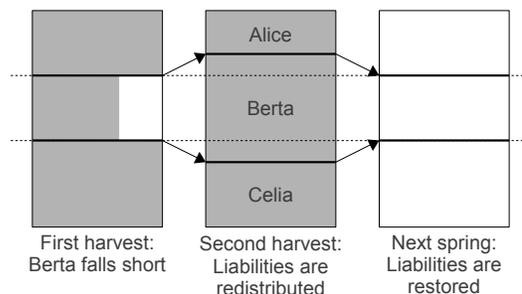


Fig. 1. Illustration of Linear Compensation in a simple public good game. Alice, Berta, and Celia farm their back-yard for carrots. Each has her individual farming liability (thick separators) but harvests are divided equally. In the first year, Berta falls short of her target by some amount (white area). Thus in the second year her share of the total liabilities is temporarily increased by some multiple of this amount, while those of the other two are decreased accordingly. Since in year two, all comply with this completely, liabilities are then restored to their target values (dashed separators).

action to the preceding compliance levels. The redistributions are basically proportional to shortfalls, i.e., to the amount by which players have failed to comply in the previous period, but with a strategically important adjustment to keep total liabilities constant. This strategy will be called 'Linear Compensation' (LinC), and its basic idea is illustrated in Fig. 1 in a fictitious community gardening example. In the emissions game, these liabilities to reduce emissions then translate into emissions allowances via the formula  $allowance = reference\ emissions - liability$ . Our assumptions and the proposed strategy are summarized in Table 1.

We prove that under certain conditions, an agreement to use the strategy LinC is *self-enforcing* in that no player or group of players has a rational incentive to ever deviate from this strategy or can ever convince the other players to switch to a different strategy by renegotiating with them. In game-theoretic terms, it is both strongly renegotiation-proof (26; 27) and a Pareto-efficient and strong Nash-equilibrium in each subgame if all players use LinC. Moreover, applying LinC requires only little knowledge of costs, benefits, and discounting, and is robust with regard to implementation errors such as inadvertent shortfalls since it reacts in a proportionate way and restores full cooperation soon afterwards. Since the strategy LinC can in principle stabilize an agreement to meet any given target allocation, it does not solve the problem of selecting these targets themselves. However, it indicates that players can focus on 'first-best' outcomes, negotiating an allocation of the highest achievable total payoff and then implementing that allocation by using LinC.

Before presenting our results in detail, we give a short literature review and define our formal framework. Regarding the emissions game, we will then discuss the validity of our assumptions and what implications the results might have for real-world climate politics.

**Existing literature on the emissions game.** A commonly used approach to strategic interaction on mitigating pollution is the theory of International Environmental Agreements, recently surveyed in (4). In this branch of the literature, cooperation has usually been modeled as a one-shot game. Players join or stay out of a long-term coalition for selfish (or rational) reasons, and within such ‘stable’ coalitions, players act to the best of the group. When this group includes all players, the cooperation dilemma is overcome. Early insights of this theory were that large stable coalitions tend to be unlikely, particularly when they would actually benefit players (5; 6), and that additional ingredients to the international agreement are needed in order to entice more players to join, e.g., side payments (7). More elaborate schemes have been conceived and explored, e.g., optimized transfers, linking with research cooperation, or endogenously determined minimum participation clauses (28–30), suggesting that higher participation levels may well be reached, but at the price of added complexity in the agreement.

A different route is taken by authors who include the time dimension in the game by modeling it as a repeated game (8–10), thus introducing a way for players to react to others’ shortfalls. In analogy to the Prisoners’ Dilemma, players have the discrete choice to ‘defect’ (emit the individual optimum) or ‘cooperate’ (emit only what is optimal globally) in most of these models. The conclusion is mostly that cooperation among more than a few players is unlikely because the threat to punish defection by universal defection is not credible. In (10), it is shown that in such a discrete model, defection by smaller numbers of players can be a credible threat deterring unilateral defections. But in a model where countries choose emissions levels from a continuum of choices, a similar strategy only works if players value the future high enough (11). We will improve upon these mixed results and show that in such a continuous model and with the ability to emit more than the individual optimal, one can even deter multi-lateral deviations from the global optimum by reacting in proportion to the size of the deviation, avoiding harsh punishments for small errors. While the above models focus primarily on analytical results, some authors also apply numerical models based on empirical data (12). Although their analysis is made difficult by the fact that numerical solution requires specifying a finite number of time periods, they

are able to show that the option to retaliate improves the prospect of cooperation.

Finally, the models in (13–18) describe the climate change game as a dynamic game with a stock pollutant, thus improving on both the repeated game model and the static one-shot game model. In (15; 18), it is shown that some intermediate amount of cooperation can be stabilized against unilateral deviations by harsh punishments. A similar model is also used in (19), the work most similar to ours: it introduces the idea of keeping total contributions at the optimal level also during punishments, but again using harsh instead of proportionate punishments. We will show that a proportionate version of their redistribution idea will even lead to renegotiation-proofness when marginal costs are equal for all players. This is in line with some real-world policy proposals that suggest a similar redistribution, although of direct financial transfers, to make threats credible and thus ensure compliance with emissions caps (3).

## Framework

**The public good game.** Assume that there are infinitely many periods, numbered  $1, 2, \dots$ , and finitely many players, numbered  $1, \dots, n$ . In each period,  $t$ , each player,  $i$ , has to choose a quantity  $q_i(t)$  as her individual contribution to the public good in that period. The resulting total contributions in period  $t$  are  $Q(t) = \sum_i q_i(t)$ .

In the emissions game,  $q_i(t)$  would be the difference between  $i$ ’s hypothetical amount of GHG emissions in period  $t$  in some pre-determined reference scenario (e.g., ‘business as usual’), and  $i$ ’s net emissions in period  $t$ . By ‘net emissions’ we mean the amount of real emissions caused domestically plus, if players use emissions trading, the amount of permits or certificates sold minus the amount of permits or certificates bought on the market. In other words,  $q_i(t) = 0$  corresponds to business-as-usual behaviour, and  $q_i(t) > 0$  means that  $i$  has reduced emissions in  $t$  domestically and/or by buying permits or certificates.

Depending on  $q_i(t)$  and  $Q(t)$ , player  $i$  has certain individual benefits  $b_i(t)$  and individual costs  $c_i(t)$  in period  $t$ . The typical conditions under which a problem of cooperation arises and can be approached by our results are reflected in the following somewhat idealized assumptions on these costs and benefits and on the information, commitment abilities, and rationality the players possess. For the emissions game, we discuss the validity of the following assumptions in more detail in the Discussion and in SI: Validity of assumptions.

The contributed good is called a ‘public’ good since individual benefits  $b_i(t)$  are determined by total contributions only, through an increasing function  $f_i(Q(t))$ . They are zero at  $Q = 0$ , and marginal benefits are non-increasing. A period’s total benefits  $B(t)$  are then given by  $f(Q(t)) = \sum_i f_i(Q(t))$ . On the negative side, we assume that total costs  $C(t)$  are also determined by a non-negative and non-decreasing function  $g(Q(t))$  of total contributions, start at zero, and marginal costs are non-decreasing.<sup>1</sup>

Unlike in many other models of public goods, we assume here that total costs are shared in a way that equalizes marginal costs. E.g., costs might be shared in proportion to individual contributions, giving  $c_i(t) = q_i(t)C(t)/Q(t)$ . Or, what is more realistic if there is a perfect competition market for contributions, costs might be shared according to a rule based on marginal cost pricing.<sup>2</sup> In both cases, one has the following convexity property on which our results will rely: for each  $Q$ , there is some ‘cost sensitivity’  $\gamma(Q)$  so that (i) if  $q_i$  and  $Q$  are both lowered by an amount  $x > 0$ , then  $c_i$  gets lowered by at most  $x\gamma(Q)$ , (ii) if  $q_i$  is raised by an amount  $x > 0$  but  $Q$

---

The public good game:

- Repeated game, no binding agreements or commitments
- Individual contributions are made per player and period and are publicly known after each period
- Positive, non-increasing marginal individual benefits, depending on total contributions
- Non-negative total costs with non-decreasing marginals, depending on total contributions, shared proportionally or based on marginal cost pricing
- All players discount future payoffs in the same way
- Optimal total contributions are known and an allocation into individual targets has been agreed upon

The strategy of Linear Compensation (LinC):

- Initial individual liabilities = targets
  - Shortfall per period = liability – actual contribution (if positive, otherwise zero)
  - New liability = target + [own shortfall – mean shortfall] · factor
  - The strategy is to always contribute your liability
- 

**Table 1. Main assumptions and solution for the public good game**

<sup>1</sup>Formally,  $f_i$  and  $g$  are twice differentiable,  $b_i(t) = f_i(Q(t))$ ,  $C(t) = g(Q(t)) \geq 0$ ,  $f_i(0) = g(0) = 0$ ,  $f_i'(Q) > 0$ ,  $g'(Q) \geq 0$ ,  $f_i''(Q) \leq 0$ , and  $g''(Q) \geq 0$ .

<sup>2</sup>Each player  $i$  would then actually contribute an amount  $a_i(Q)$  for which its individual pre-trade cost function  $g_i$  has marginal costs  $g_i'(a_i(Q))$  equal to the global marginal costs  $g'(Q)$ , and would buy the remaining contribution,  $q_i - a_i(Q)$ , at a price that also equals  $g'(Q)$ . Individual costs are then  $c_i = g_i(a_i(Q)) + [q_i - a_i(Q)]g'(Q)$ .

is kept constant by lowering the other values  $q_j$ , then  $c_i$  raises by at least  $x\gamma(Q)$ , and (iii)  $c_i = 0$  for  $Q \leq 0$ . In other words, lowering your contributions by  $x$  saves you at most costs of  $x\gamma(Q)$ , but if  $x$  contributions are redistributed from others to you, your costs raise by at least that same amount. It is easy to see that in the proportional cost-sharing case,  $\gamma(Q)$  equals average costs  $g(Q)/Q$ , while in the marginal cost pricing case,  $\gamma(Q)$  equals marginal costs  $g'(Q)$ .

In the emissions game, the benefits of reducing emissions by 1 Gt CO<sub>2</sub>-equivalents in period  $t$  correspond to all avoided welfare losses that would have been caused at times after  $t$  by that additional 1 Gt of emissions, properly discounted to reflect the corresponding time difference, and using any suitable welfare measure such as consumption, income, gross domestic product (GDP), etc. (20–22). The above form of the costs  $c_i$  seems justified when we assume an international emissions market between firms, similar to the European Union Emission Trading Scheme (EU ETS). A simple example cost-benefit structure is that of linear benefits and linear marginal costs (23):  $f_i(Q) = \beta_i Q$  with  $\beta_i > 0$ ,  $g(Q) = Q^2$  for  $Q > 0$ , and  $g(Q) = 0$  for  $Q \leq 0$ . For other examples, see SI: Examples.

We explicitly allow individual contributions  $q_i$  to be any real number in principle, positive or negative. However, as  $Q$  gets large, costs get prohibitively high, and as  $Q$  gets small, benefits get prohibitively negative. Hence total period payoffs,  $P(t) = B(t) - C(t)$ , are bounded from above but not from below, with  $P(t) \rightarrow -\infty$  for  $Q(t) \rightarrow \pm\infty$ . In the emissions game, large positive or negative values for some  $q_i$  can obtain if large amounts of emissions permits are traded. Although the strategy we will propose below prescribes such large values of  $q_i$  only in cases where there has already been an irrationally large earlier deviation, this might still lead to problems in practice (for an alternative model, see SI: Bounded liabilities).

Players make the choices  $q_i(t)$  individually and simultaneously in each  $t$ , and all know that no player can commit himself bindingly to some value of  $q_i(t)$  at some time earlier than  $t$ . They also know that each  $i$  has complete information about costs, benefits, and all past contributions when choosing  $q_i(t)$ . Players are assumed to be rational in that they aim at maximizing their long-term payoff, using some strategy to choose  $q_i(t)$  on the basis of this information, and expect the others to do so as well. Regarding how much the players value next period's payoffs in comparison to this period's, we assume as usual that for some constant  $\delta > 0$  and all periods  $t$ , all prefer to get one payoff unit in period  $t + 1$  to getting  $\delta$  payoff units in  $t$ .

For some optimal amount  $Q^*$  of total contributions, total (expected) payoff gets maximized, and marginal total costs equal marginal total benefits but exceed marginal individual benefits:

$$f(Q^*) - g(Q^*) = \max, \quad g'(Q^*) = f'(Q^*) > f'_i(Q^*). \quad [1]$$

$\alpha$	compensation factor
$B(t), b_i(t)$	benefits in period $t$ , total and for player $i$
$\beta_G$	marginal benefits at target, for a group of players $G$
$C(t), c_i(t)$	costs in period $t$ , total and for player $i$
$\bar{d}(t), d_i(t)$	shortfalls in period $t$ , average and of player $i$
$\delta$	lower bound for discounting factors
$f(Q), f_i(Q)$	benefit functions, total and for player $i$
$g(Q)$	total cost function
$\gamma(Q), \gamma^*$	cost sensitivity function, and value at target $Q^*$
$\ell_i(t)$	liability of player $i$ in period $t$
$Q(t), q_i(t)$	contributions in period $t$ , total and by player $i$
$Q^*, q_i^*$	target contributions, total and for player $i$
$x$	size of potential shortfall by a group of players $G$

**Table 2. Main symbols used in this article**

Optimal total payoffs are usually much larger than the total payoffs the players would end up if they do not cooperate. E.g., in the simple example with linear benefits and marginal costs, optimal total payoffs are larger than the non-cooperative equilibrium payoffs by a factor of approximately  $n^2/4$ , showing that the potential gains of cooperation can be very large and increase with the number of players (see SI: One-shot game and SI: Examples).

Finally, let us assume that players can enter no legally binding and enforceable agreements (since this is the worst case assumption when studying the possibility of cooperation) but have somehow chosen in advance (before period one) an allocation of the optimum target into individual targets  $q_i^*$ , with  $\sum_i q_i^* = Q^*$ . This allocation will be so that no group  $G$  of players has an incentive to contribute more than what was agreed as their joint target  $Q_G^* = \sum_{i \in G} q_i^*$ .<sup>3</sup>

In the emissions game, targets might be negotiated using equity criteria such as per capita emissions permits, per capita payoffs, historical responsibility, etc. (31–33) (34, p.915). In game-theoretic terms, this initial negotiation poses a problem of equilibrium selection that precedes the problem of cooperation which we are concerned with in this article (see also SI: Cooperative analysis).

**Free-riding and renegotiations.** In this kind of public good game, the problem of cooperation is now this: Although the negotiated targets provide the optimal total payoff and are often also profitable for each individual player, they constitute no binding agreement. Hence player  $i$  will hesitate to meet the target if he can hope that the others will meet it, since contributing less reduces  $i$ 's costs more than his benefits (see Eqn. [1]). If there is only one period of play, this free-rider incentive is known to make cooperation almost impossible, since rational players will then contribute a much smaller quantity, which means that the agreement is not self-enforcing (for more on this, see SI: Properties of the one-shot game).

In a repeated game, however, a player  $i$  can react to the other players' earlier actions by choosing  $q_i(t)$  according to some strategy  $s_i$  that takes into account all players' individual contributions before  $t$ . The immediate gains of free-riding might be offset by future losses if others react suitably. The announcement to react in such a way can then deter free-riding as long as that announcement is *credible* (see, e.g., Robert Aumann's Nobel Lecture (35)).

However, if those who react to free-riding would thereby reduce their own long-term payoffs, and if they cannot bindingly commit themselves beforehand to actually carry out the announced reaction despite harming themselves in doing so, then such a threat would not be credible since a potential free-rider could expect that a rational player will not harm herself but rather overlook the free-riding. After the fact, a free-rider of period  $t$  could then successfully *renegotiate* with the others between periods  $t$  and  $t + 1$ , convincing them to "let bygones be bygones". The effect is that his free-riding in  $t$  will be ignored, since in  $t + 1$  everyone benefits from doing so (26).

A famous example of such a non-credible strategy, though in a different game, is the strategy 'tit for tat', observed in various versions of the repeated Prisoners' Dilemma when players can commit themselves beforehand (36; 37). That strategy is to start with 'cooperate' and then do whatever the other player did in the previous period, thereby punishing defection with defection. But once this calls for 'defect' in some period, both would be better off at that point if they instead both continued with 'cooperate'. So the threat to defect after

<sup>3</sup>Formally:  $\sum_{i \in G} f'_i(Q^*) < h'(0)$  where  $h(x) = (Q_G^* + x)g(Q^* + x)/(Q^* + x)$ .

<sup>4</sup>Unfortunately, experimental studies of repeated games have yet been rare and inconclusive about the question of what the effect of credible threats on cooperation is. E.g., in (39) it is concluded that the existence of equilibria with credible threats is a necessary but not sufficient condition for cooperation in a certain type of game, while others, like (40), report that sometimes cooperation can also be sustained without credible threats in the laboratory. In (41, p.1502) it is concluded from the experience with existing International Environmental Agreements that only those treaties in which compliance could be enforced lead to a substantial amount of cooperation, which can also be interpreted as supporting the necessity of credible threats.

a defection is void and cannot deter free-riding under assumptions of rationality and without commitment possibilities (38).<sup>4</sup>

Another problematic strategy is to simply treat free-riding as some form of debt to be repaid with interest, as it is done, e.g., in the Kyoto protocol, in which a country falling short in one period has its liabilities in the following period increased by 1.3 times the size of its shortfalls. In our framework, such a rule would lead to inefficient contributions in  $t + 1$  that exceed the optimal value  $Q^*$ , making renegotiations likely that lower all liabilities to an efficient value. Even worse, if a player never fulfills his liabilities, he gets away with it.

Depending on the cost-benefit structure of a repeated game, there might or might not be strategies that achieve a certain level of stability against deviations such as free-riding and against incentives to renegotiate. Fortunately, we can formally prove that in our assumed framework, a rather simple, proportionate combination of the above two ideas of punishing other's and repaying one's own shortfalls is both efficient and highly stable, even when players make small errors in implementing it. Table 1 summarizes our main assumptions and the suggested solution that we present below.

## Results

**Avoiding renegotiations.** Let us deal with the question of renegotiations first. The crucial idea to avoid those in our kind of game is to keep total contributions constant and only redistribute them as a reaction to past behaviour. Consider a strategy  $s$  which, in each period  $t$ , tells all players to choose their contributions  $q_i(t)$  in a certain way which makes sure that the total target is met,  $Q(t) = Q^*$ . Then no matter the actions before  $t$ , there can be no alternative strategy  $\tilde{s}$  that achieves higher total payoffs than  $s$  from time  $t$  on. So, any alternative strategy  $\tilde{s}$  that leads to different payoffs than  $s$  would lead to a strictly smaller payoff than  $s$  for at least one player. This holds whether only payoffs in  $t$  are considered or also later payoffs with discounting. Hence there is no possible situation in the game that would cause all players to agree to change the strategy. In game-theoretic terms, such a strategy is 'strongly perfect', i.e., Pareto-efficient in all subgames. It will thus be *strongly renegotiation-proof* (26; 27) if we manage to do the redistribution of contributions in  $t + 1$  in a way that makes free-riding in  $t$  unprofitable in the long run. This we will do next.<sup>5</sup>

**Deterring simple free-riding by groups of players.** Suppose in some period  $t$ , all players contribute their targets, except that a set  $G$  of players free-rides. This means they jointly contribute only a quantity  $Q_G(t) = \sum_{i \in G} q_i(t)$  that is by some amount  $x > 0$  smaller than their joint target contribution:  $Q_G(t) = Q_G^* - x$ . Note that  $G$ 's benefits are given by  $f_G(Q) = \sum_{i \in G} f_i(Q)$ , so that  $\beta_G = f'_G(Q^*)$  is  $G$ 's target marginal benefit. Let  $\gamma^* = \gamma(Q^*)$  be the cost sensitivity at the target contributions. Then  $G$ 's shortfalls reduce their joint benefits in  $t$  by at least  $x\beta_G$ , but saves them costs of at most  $x\gamma^*$ . Hence their joint payoff increases by at most

$$x(\gamma^* - \beta_G). \quad [2]$$

How much redistribution in  $t + 1$  is now needed to make this unprofitable for  $G$ ? Suppose the contributions in  $t + 1$  are redistributed in such a way that everyone gets their target benefits but group  $G$  has additional costs, and these additional costs times  $\delta$  are no smaller than the above  $x(\gamma^* - \beta_G)$ . Then, in period  $t$ , it is not attractive for  $G$  to free-ride, since in that period, they value their resulting losses in  $t + 1$  higher than their gains in  $t$ . Such a redistribution can easily be achieved: Just raise  $G$ 's joint contributions  $Q_G(t + 1)$  from  $Q_G^*$  by at least  $x(\gamma^* - \beta_G)/\gamma^*\delta$  and reduce the other players' contributions accordingly.<sup>6</sup> This leads to additional costs for  $G$  in  $t + 1$  of at least

$$x(\gamma^* - \beta_G)/\delta. \quad [3]$$

So,  $G$ 's joint gains in  $t$  are overcompensated by these losses in  $t + 1$ . Although free-riding for one period might be profitable for some individual members of  $G$ , there is always at least one member of  $G$  for

whom it is not. Fig. 1 illustrates the basic idea. We will show next how the same kind of redistribution can be used to deter also every conceivable sequence of deviations from the target path.

**The strategy of Linear Compensation (LinC).** A simple strategy that does this assigns each player  $i$  in each period  $t$  a certain *individual liability*  $\ell_i(t)$  which that player should contribute in  $t$ . In period one, liabilities equal the negotiated targets,  $\ell_i(1) = q_i^*(1)$ . Later, they depend on the differences between last period's liabilities and actual contributions of all players. After each period  $t$ , we first compute everyone's *shortfalls* in  $t$ , which are  $d_i(t) = \ell_i(t) - q_i(t)$  if  $\ell_i(t) > q_i(t)$ , and otherwise  $d_i(t) = 0$ , that is, we do not count excesses. Then we redistribute the targets in  $t + 1$  so that these shortfalls are compensated linearly, but keeping the total target unchanged:

$$\begin{aligned} \text{new liability} &= \text{target} + [\text{own shortfall} - \text{mean shortfall}] \cdot \text{factor} \\ \ell_i(t + 1) &= q_i^* + [d_i(t) - \bar{d}(t)] \cdot \alpha. \end{aligned} \quad [4]$$

In this,  $\bar{d}(t) = \sum_i d_i(t)/n$  is the mean shortfall and  $\alpha$  is a certain positive *compensation factor* we will discuss below. Obviously, if all players comply with their liabilities by putting  $q_i(t) = \ell_i(t)$ , then all shortfalls are zero, and both liabilities and contributions stay equal to the original targets so that the optimal path is implemented.

The compensation factor  $\alpha$  has to be large enough for the argument of the previous section to apply in all possible situations, whatever the contributions have been before  $t$ . In the simple free-riding situation discussed in the previous section, the group's joint shortfall equals  $x$  and the mean shortfall is  $\bar{d}(t) = x/n$ . Hence  $G$ 's joint additional liability in  $t + 1$  is  $[x - |G|x/n] \cdot \alpha$ , where  $|G| < n$  is the number of players in  $G$ . If this is at least  $x/\delta$ , then having shortfalls of size  $x$  is not profitable, independently of what the actual liabilities in  $t$  were. Since only shortfalls but not excesses lead to a redistribution, a group can neither profit from contributing more than their liability.

In other words, to make sure no group of players has ever an incentive to deviate from their liability for one period, even if liabilities are already different from the target, it suffices if

$$\alpha > \frac{n}{\gamma^*\delta} \cdot \max_G \frac{\gamma^* - \beta_G}{n - |G|}, \quad [5]$$

where the maximum is taken over all possible groups of players  $G$ . If it is known that the benefit functions of all players are equal, then  $\beta_G = C'(Q^*)|G|/n \geq \gamma^*|G|/n$  and Eqn. [5] simplifies to  $\alpha > [n\gamma^* - C'(Q^*)]/\gamma^*\delta(n - 1)$ , so that in particular  $\alpha > 1/\delta$  suffices. Note that liabilities do not depend on costs and benefits explicitly, only via the negotiated targets  $q_i^*$  and the factor  $\alpha$ , so the information about costs and benefits one needs to apply LinC is limited to the knowledge of the optimum contribution and the marginal costs and benefits at the target. Now, a player  $i$  who complies with the liabilities defined by Eqns. [4] and [5] by putting  $q_i(t) = \ell_i(t)$  is said to apply the strategy of 'Linear Compensation' (LinC).

In game-theoretic terms, we have shown above that when all players apply LinC, this forms a 'one-shot subgame-perfect' equilibrium. It is then also never profitable to deviate from LinC for any number of successive periods. The proof for this follows a standard

<sup>5</sup>If we drop the assumption that the global target  $Q^*$  maximizes total payoff, e.g., because of uncertainty in estimating the optimum, then such redistribution strategies are no longer Pareto-efficient in all subgames. Renegotiations that improve total payoff may then happen, which is desirable. Still, the same reasoning as above shows that there is never an incentive for all players to pretend past actions were different from what they really are, hence no group of players can convince the rest to ignore their shortfalls. This is called 'weak renegotiation-proofness' (26; 27). See also SI: Renegotiations when targets are not optimal.

<sup>6</sup>If  $G$  consists of all  $n$  players, optimality of  $Q^*$  implies that shortfalls give no gains for  $G$  in period  $t$ .

argument (42).<sup>7</sup> In the Appendix, we prove that even no conceivable *infinite* sequence of deviations is profitable for any group  $G$  of players. Hence for any given set of targets  $q_i^*$ , it builds a strong Nash equilibrium in each subgame if all players apply LinC given these targets. Roughly speaking, the reason is that if  $G$  continually falls short, contributions of the other players will decrease fast enough, so that, in the long run,  $G$ 's gains from saved costs are overcompensated by their losses from decreased total contributions. Note that the others do not need to use a threat of contributing nothing forever (which would not be credible), but only threaten to respond to each period of shortfalls with a period of punishment, one at a time. This gradual escalation is credible when there is 'common knowledge of rationality,' since  $G$  knows in advance that after each individual period  $t$  of shortfalls, the others still expect them to follow their rational interest and return to compliance in  $t + 1$  instead of falling short again, no matter how many shortfalls have happened already.<sup>8</sup>

## Discussion

We have presented a simple strategy by which players in a public good game can keep each other in check in the provision of agreed target contributions. Our approach can be interpreted as a combination of a proportionate version of the punishment approach that strategies like 'tit for tat' use in the Prisoners' Dilemma, and the repayment approach that is already included in the Kyoto mechanism. This combination has been formally shown here to have strong game-theoretic stability properties in situations where some simplifying assumptions hold, a feature that is not true of strategies that use only one of the two ingredients. In Axelrod's (36) terminology, our strategy, LinC, is 'nice' in that it cooperates unless provoked, 'retaliating' when provoked, 'forgiving' when deviators repay, and uses 'contribution' to avoid the echo effect.

We believe that very similar strategies will be valuable also in contexts in which some of our assumptions are violated. E.g., future work might use an improved model of the emissions game in which the assumption of identical periods is replaced by certain path-dependencies: Real-world benefit functions  $f_i$  depend on GHG stocks and hence on time and emission history, and also the cost function  $g$  depends on time and emission history because of technological progress. Since past contributions will reduce future marginal costs, this will lead to a non-constant optimal abatement path  $Q^*(t)$ . However, these effects will probably not weaken LinC's stability when  $q_i^*$  is replaced by a time-dependent target allocation  $q_i^*(t)$  of  $Q^*(t)$  that is computed according to some initially negotiated rule (e.g., in fixed proportions). This is because then the Pareto-efficiency argument for renegotiation-proofness still holds, while shortfalls would slow down technological progress and lead to even higher marginal costs in the punishment period.

A more critical assumption is that contributions are unbounded which would make it possible in principle to punish even long sequences of large shortfalls by escalating emissions, a possible development which rational players would then avoid. If emissions can not exceed some upper bound, it would still suffice if welfare losses became prohibitively large when emissions approach that bound. Only if those losses are bounded as well, the question whether large shortfalls can be deterred depends on the actual cost-benefit structure and on the value of  $\delta$ , which is in line with general results on repeated games with bounded payoffs (42) (see also SI: Bounded liabilities, SI: Validity of assumptions).

In addition to such model refinements, future work should also (i) assess the possibility of players to "bind their hands" ahead of time by making long-term investment decisions that reduce their own ability to choose  $q_i(t)$  at  $t$ , (ii) study the influence of incomplete information due to restricted monitoring capacities, finite planning horizons and of other forms of 'bounded rationality' (43), (iii) link emissions reductions with other issues (44), (iv) include possible altruism, rep-

utation, and status effects, also using experimental approaches such as (45).

Since LinC uses a proportionate and timely measure-for-measure reaction to shortfalls, it performs well also in situations in which players cannot control their actions perfectly. It is easy to see from Eqn. [4] that random errors do not add up or lead away from the target, nor do one-time deviations initiate a long sequence of reactions.<sup>9</sup> The latter is avoided by comparing actual contributions not to the initial targets but to dynamic liabilities, which are similar to the 'standings' used in 'concrete tit-for-tat' for the repeated Prisoners' Dilemma (46). All the above stability properties of LinC hold independently of the form and amount of discounting if the compensation factor  $\alpha$  is chosen properly.<sup>10</sup> While many other games have no strong Nash equilibria, the public good game studied here somewhat surprisingly even allows players to sustain any allocation of the optimal total payoff with a strategy that is a strong Nash equilibrium even in each subgame (though leaving the coordination problem of equilibrium selection as a task for prior negotiations). Since deviations by groups have been considered before only for non-repeated 'normal-form' games, this new combination of 'strong Nash' and 'subgame-perfect' equilibrium can also be considered a contribution to game theory itself.

In real-world climate politics, redistribution mechanism such as ours could play a key role in the implementation of cap-and-trade regimes, whose importance is stressed by many authors (see, e.g., the impressively broad collection of articles in (2)). While in domestic emissions markets, caps can be issued by a central authority and compliance might be enforced legally, both is more difficult in large international markets (47). If, like in the first two periods of the EU ETS, each country in a market issues its own permit quantity  $q_i$ , a strategy like LinC might be used to ensure compliance with some agreed individual caps that realize that market's joint optimum, giving countries incentives to issue only the agreed target amount of permits and to ensure that domestic emissions are matched by permits after trading. To choose a suitable compensation factor, only a conservative estimate of the (expected) marginal costs and benefits at the target and the short-term discounting factor is needed.

In this way, one could avoid using "sticks" such as trade sanctions (48, p. 34) or tariffs (3), which are mostly considered to be difficult to push politically vis-a-vis partners, and focus on "carrots" (benefitting from other players' emissions reductions). Still, tariffs might be helpful vis-a-vis non-participants, who might prefer to avoid them by joining the market (49). Also, starting with a number of regional markets with possibly sub-optimal caps, several such markets might merge to decrease marginal costs (50; 51), eventually leading to a global cap-and-trade system with a globally optimal cap. Whenever caps need to be negotiated anew due to new participants or new cost-benefit estimates, any pre-negotiation shortfalls would still be taken into account in LinC, providing both continuity and flexibility as demanded in (48, p. 36). Likewise, compliance with the Kyoto protocol might improve if its current compensation rule was modified to keep total liabilities constant as in Eqn. [4] and if the current compensation factor of 1.3 was adjusted according to Eqn. [5]. In contrast, the harsh punishment strategies on which earlier studies have focussed are not only less strategically stable but also less practicable because of their disproportionate reactions and their strict distinction between 'normal' and 'punishment' periods.

<sup>7</sup>If  $m$  successive deviations were profitable, but no shorter sequence was, then one-shot subgame-perfectness would imply that after the first  $m - 1$  deviations, the  $m$ -th is no longer profitable. Hence already the first  $m - 1$  deviations would have been profitable — a contradiction. Infinite sequences have to be considered separately since payoffs are unbounded.

<sup>8</sup>This expectation is common to all Nash-like equilibrium concepts. The much stronger demand that compliance should be optimal regardless of the other players' behavior would require so-called 'dominant' strategies which, however, do rarely exist in repeated games.

<sup>9</sup>With implementation errors of variance  $\sigma^2$ , the mean squared deviation of  $\ell_i(t + 1)$  from the target  $q_i^*$  will be at most  $\sigma^2 \alpha^2 (n - 1)/n$ , hence the mean squared deviation between actual and target contributions is of magnitude  $\sigma^2 (1 + \alpha^2 (n - 1)/n)$ .

<sup>10</sup>The value of  $\delta$  however does play a role when, in addition to our assumptions, liabilities shall be bounded. This is further explored in SI: Bounded liabilities.

## Appendix: Why infinite sequences of deviations do not pay

Suppose all players apply LinC by putting  $q_i(t) = \ell_i(t)$  except that from some period  $t_0$  on, a group  $G$  of players play a *deviation strategy*  $s$  that leads to joint shortfalls  $\sum_{i \in G} d_i(t) = x_t$  in each period  $t \geq t_0$ . Since excess contributions never pay, we can assume that  $x_t \geq 0$ . Assume further that in each period  $t$  and for each integer  $r \geq 0$ , all players consider getting one payoff unit in period  $t+r$  as equivalent to getting  $w_{t,r}$  payoff units immediately in period  $t$ , where the discounting weights  $w_{t,r}$  fulfill the conditions

$$w_{t,0} = 1, \quad w_{t,1} > \delta, \quad w_{t,r} \geq 0, \quad \sum_{r=0}^{\infty} w_{t,r} = W_t < \infty. \quad [6]$$

E.g., players could use exponential discounting with  $w_{t,r} = \varepsilon^r$ ,  $\delta < \varepsilon < 1$ , and  $W_t = 1/(1-\varepsilon)$ .<sup>11</sup>  $G$ 's discounted long-term payoff from  $t_0$  on is then  $U_G(t_0) = \sum_{t \geq t_0} w_{t_0,t-t_0} P_G(t)$  with joint period payoffs  $P_G(t) = \sum_{i \in G} (b_i(t) - c_i(t))$ . We will show that this is no larger than if they had continued to apply LinC instead. Assume  $\Delta(s, \text{LinC}) > 0$  is the difference in  $U_G(t_0)$  between playing  $s$  and playing LinC from  $t_0$  on, and consider the following two cases.

(i) Suppose the discounted total long-term shortfalls are finite, i.e., the series  $\sum_{t \geq t_0} w_{t_0,t-t_0} x_t$  of non-negative terms converges. Now consider the truncated deviation strategy  $\tilde{s}$  that returns to compliance in some period  $t_1 > t_0$ , i.e., consists in playing  $s$  for  $t_0 \leq t < t_1$  and playing LinC for  $t \geq t_1$ . Let  $\Delta(s, \tilde{s})$  be the difference in  $U_G(t_0)$  between playing  $s$  and  $\tilde{s}$ . This is at most the costs they save in periods  $t \geq t_1$  when playing  $s$  instead of LinC, which is at most  $x_t \gamma^*$  according to Eqn. [2]. Hence  $\Delta(s, \tilde{s}) \leq \sum_{t \geq t_1} w_{t_0,t-t_0} x_t \gamma^*$ . Because of the assumed series convergence, this goes to zero for  $t_1 \rightarrow \infty$ , so it is smaller than  $\Delta(s, \text{LinC})$  if  $t_1$  is large enough. Then  $\Delta(\tilde{s}, \text{LinC}) = \Delta(s, \text{LinC}) - \Delta(s, \tilde{s}) > 0$  which means that already the truncated deviation strategy  $\tilde{s}$  is profitable. But we already proved that no finite sequence of deviations is profitable, so neither is  $s$ .

(ii) Suppose the discounted total long-term shortfalls are infinite,  $\sum_{t \geq t_0} w_{t_0,t-t_0} x_t = \infty$ . Because  $x_{t-1} \geq 0$ , the joint liability of  $G$  in period  $t$  is no smaller than the target,  $L_G(t) = \sum_{i \in G} \ell_i(t) \geq Q_G^*$ . Hence their joint costs  $C_G(t)$  are either zero if  $x_t \geq Q_G^*$ , since then total costs are zero, or they are by at most  $Q_G^* \gamma^*$  smaller than in the case where  $L_G(t) = Q_G^{\min}$ . In other words,  $C_G(t)$  is bounded from below by some value  $C_G^{\min}$ . Concerning benefits, let  $f_G(Q) = \sum_{i \in G} f_i(Q)$  and let  $\beta_G = f'_G(Q^*)$  be the target marginal benefit of  $G$ . Then  $G$ 's joint benefits are  $f_G(Q^* - x_t)$ , which is at most  $f_G(Q^*) - \beta_G x_t$  because marginal benefits are non-increasing. Thus  $G$ 's joint payoffs are at most  $(Q^* - Q_G^*) \gamma^* + f_G(Q^*) - \beta_G x_t$ , so that  $G$ 's discounted long-term payoff  $U_G(t_0)$  is then at most

$$W_{t_0} [f_G(Q^*) - C_G^{\min}] - \beta_G \sum_{t \geq t_0} w_{t_0,t-t_0} x_t. \quad [7]$$

But the latter series diverges because of our assumption, hence  $U_G(t_0) = -\infty$ . In other words, an infinite sequence of shortfalls growing this fast is infinitely bad.<sup>12</sup>

**ACKNOWLEDGMENTS.** This work was partially supported by the Federal Ministry for Education and Research (BMBF) via the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS). We thank Gunnar Luderer, the Policy Instruments Group at PIK, the SFB-TR15 seminar in Berlin, Ottmar Edenhofer, and John Schellnhuber for fruitful discussions, as well as two anonymous referees for their very helpful comments.

## References.

- Dinar A, Albiac J, & Sánchez-Soriano J (2008) *Game Theory and Policy Making in Natural Resources and the Environment* (Routledge, London).
- Aldy JE & Stavins RN, eds (2010) *Post-Kyoto International Climate Policy: Implementing Architectures for Agreement* (Cambridge University Press).
- Karp L & Zhao J. (2010) in (2), pp 530–562
- Finus M (2008) Game theoretic research on the design of international environmental agreements: insights, critical remarks, and future challenges. *Int Rev Environ Res Econ* 2:29–67.
- Carraro C & Siniscalco D (1993) Strategies for the international protection of the environment. *J Public Econ* 52:309–328.

- Barrett S (1994) Self-enforcing international environmental agreements. *Oxford Econ Pap* 46:878–894.
- Barrett S (2001) International cooperation for sale. *Eur Econ Rev* 45:1835–50.
- Barrett S (1999) A theory of full international co-operation. *J Theor Polit* 11:519–541.
- Asheim GB, Froyen CB, Hovi J, & Menz FC (2006) Regional versus global cooperation for climate control. *J Environ Econ Manag* 51:93–109.
- Froyen CB & Hovi J (2008) A climate agreement with full participation. *Econ Lett* 99:317–319.
- Asheim GB & Holtmark B (2009) Renegotiation-proof climate agreements with full participation: conditions for Pareto-efficiency. *Environ Resour Econ* 43:519–533.
- Weikard HP, Dellink R, & van Ierland E (2010) Renegotiations in the greenhouse. *Environmental and Resource Economics* 45:573–596.
- Dutta PK & Radner R (2004) Self-enforcing climate-change treaties. *P Natl Acad Sci USA* 101:5174–9
- Rubio SJ & Casino B (2005) Self-enforcing international environmental agreements with a stock pollutant. *Spanish Economic Review* 7:89–109.
- Dutta PK & Radner R (2006) A game-theoretic approach to global warming. *Adv Math Econ* 8:135–153.
- Rubio SJ & Ulph A (2007) An infinite-horizon model of dynamic membership of international environmental agreements. *J Environ Econ Manag* 54:296–310.
- de Zeeuw A (2008) Dynamic effects on the stability of international environmental agreements. *J Environ Econ Manag* 55:163–174.
- Dutta PK & Radner R (2009) A strategic analysis of global warming: theory and some numbers. *J Econ Behav Organ* 71:187–209.
- Mason CF, Polasky S, & Tarui N (2010) Cooperation on Climate-Change Mitigation. *University of Exeter Working Papers*.
- Intergovernmental Panel on Climate Change (2007) *IPCC Fourth Assessment Report: Climate Change 2007 (AR4)* (Cambridge University Press, Cambridge, UK).
- U.S. National Academies of Sciences (2010) *Advancing the Science of Climate Change* (National Academies Press, Washington, D.C.).
- U.S. National Research Council (2010) *Limiting the Magnitude of Climate Change* (National Academy Press, Washington, D.C.).
- Finus M, Ierland EV, & Dellink R (2006) Stability of climate coalitions in a cartel formation game. *Economics of Governance* 7:271–291.
- Ellerman AD & Decaux A (1998) Analysis of post-Kyoto CO<sub>2</sub> emissions trading using marginal abatement curves. *MIT Joint Program on the Science and Policy of Global Change* 40:1–33.
- Ellerman AD (2010) in (2), pp 88–150.
- Farrell J & Maskin E (1989) Renegotiation in repeated games. *Game Econ Behav* 1:327–360.
- Bergin J & MacLeod B (1993) Efficiency and renegotiation in repeated games. *J Econ Theory* 61:42–73.
- Carraro C, Eyckmans J, & Finus M (2006) Optimal transfers and participation decisions in international environmental agreements. *The Review of International Organizations* 1:379–396.
- Carraro C, Marchiori C, & Orefice S (2009) Endogenous minimum participation in international environmental treaties. *Environ Resour Econ* 42:411–425.
- Lessmann K & Edenhofer O (2010) Research cooperation and international standards in a model of coalition stability. *Resour Energy Econ* 33:36–54.
- Weikard HP, Finus M, Altamirano-Cabrera JC (2006) The impact of surplus sharing on the stability of international climate agreements. *Oxford Econ Pap* 58:209–232.
- Kverndokk S & Rose A (2008) Equity and justice in global warming policy. *Fondazione Eni Enrico Mattei Working Papers* 239:1–66.
- Knopf B, Lueken M, Bauer N, & Edenhofer O (2009) *Distributing emission allowances versus sharing mitigation burden: two contrary perspectives on climate justice among world regions* (IOP Publishing).
- Aldy JE & Stavins RN (2010) in (2), pp 899–929.
- Aumann RJ (2006) War and peace. *P Natl Acad Sci USA* 103:17075–8.
- Axelrod R (1981) The emergence of cooperation among egoists. *Am Polit Sci Rev* 75:306–318.
- Axelrod R & Keohane RO (1985) Achieving cooperation under anarchy: strategies and institutions. *World Polit* 38:226–254.
- van Damme E (1989) Renegotiation-proof equilibria in repeated prisoners' dilemma. *J Econ Theory* 47:206–217.
- Dal Bó P & Fréchette GR (2011) The evolution of cooperation in infinitely repeated games: experimental evidence. *Amer Econ Rev* 101:411–429.
- Davis DD & Holt CA (1999) Equilibrium cooperation in two-stage games: experimental evidence. *Intern J Game Theory* 28:89–109
- Barrett S (2005) in *Handbook of Environmental Economics*, Vol. 3, eds Mäler KG & Vincent JR (Elsevier), pp 1457–1516.
- Abreu D (1988) On the theory of infinitely repeated games with discounting. *Econometrica* 56:383–396.
- Jones BD (2001) *Politics and the architecture of choice: Bounded rationality and governance* (University of Chicago Press, Chicago).
- Carraro C & Marchiori C (2004) in *Endogenous strategic issue linkage in international negotiations*, eds Carraro C & Fragnelli V. (Edward Elgar), pp 65–84.
- Milinski M, Semmann D, Krambeck HJ, & Marotzke J (2006) Stabilizing the Earth's climate is not a losing game: supporting evidence from public goods experiments. *P Natl Acad Sci USA* 103:3994–8.

<sup>11</sup>Another example is hyperbolic discounting with  $w_{t,r} = (1 + \alpha_t r)^{-1 - \zeta_t}$  and certain parameters  $\alpha_t, \zeta_t > 0$  (52). For the emissions game, see the discussion in (53?). If individual players discount differently, one says they have different *time-preferences*, the analysis gets more complicated because intertemporal trade can be profitable (54), and our results concerning renegotiations might no longer hold. If efficient financial markets exist, they can be expected to equalize discount rates (54) so that our assumption would be valid in the emissions game.

<sup>12</sup>Since  $G$ 's joint payoff cannot be increased, it is in particular not possible to increase every member's individual payoff. Hence all our results concerning groups are still meaningful if there is no 'transferable utility'. In the emissions game, e.g., benefits from avoided damages might contain components related to individual well-being that cannot be considered transferable. Still, payoffs from trade must be assumed to be linear in revenues for our assumptions on the cost function to be valid.

46. Sugden R (1986) *The evolution of rights, co-operation and welfare* (Blackwell, Oxford).
47. Victor D (2010) in (2), pp 618–648.
48. Frankel J (2010) in (2), pp 31–87.
49. Lessmann K, Marschinski R, & Edenhofer O (2009) The effects of tariffs on coalition formation in a dynamic global warming game. *Econ Model* 26:641–649.
50. Flachsland C, Marschinski R, & Edenhofer O (2009) Global trading versus linking: architectures for international emissions trading. *Energy Policy* 37:1637–1647.
51. Jaffe J & Stavins RN (2010) in (2), pp 119–150.
52. Jamison DT & Jamison J (2003) *The amount and speed of discounting* (Congress of the European Economic Association, Stockholm).
53. Dasgupta P (2008) Discounting climate change. *J Risk Uncertainty* 37:141–169.
54. Lehrer E & Pauzner A (1999) Repeated games with differential time preferences. *Econometrica* 67:393–412.

# Supporting Information for article Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games

Jobst Heitzig\*, Kai Lessmann\*, and Yong Zou\*

\*Potsdam Institute for Climate Impact Research, P. O. Box 60 12 03, 14412 Potsdam, Germany

Final authors' draft as accepted for publication in PNAS on July 25, 2011. Published version: DOI:10.1073/pnas.1106265108

## SI: Validity of assumptions on the emissions game

Typical models of the emissions game used in the literature fulfill our assumptions on costs and benefits (1; 2).

Concerning *benefits*, the economic literature on climate change as distilled in the IPCC's 4th Assessment Report and recent reports by the U.S. National Academies indicates that the global society as a whole would benefit from reduced GHG concentrations (3–5). The regional distribution of the consequences of climate change is much more uncertain, but some studies (6; 7) suggest that on a suitable level of regional aggregation, most or all world regions do indeed have positive marginal benefit functions  $f'_i$ , whether in terms of GDP, consumption, or other welfare measures. If some country or region  $i$  would not profit from reduced GHG concentrations, it may still be part of a politically or economically closely integrated group of countries that would profit from reduced GHG concentrations as a group. In that case, it may be appropriate to treat that group as a single player, and indeed many models use world regions instead of countries as players (1; 2; 8). Otherwise,  $i$  has to be excluded from the scope of LinC and its contributions (if any) could be treated as an exogenous variable for our solution to be applicable. This could also be done with countries whose emissions cannot be monitored or that cannot be considered rational agents (see below) or seem to use a discounting rate much larger than the others.

The common assumption that marginal benefits are non-increasing was made mainly for simplification. Many models in the literature even assume constant marginal benefits. If actual marginal benefits can be increasing, e.g., because of certain tipping elements in the Earth system (9; 10), our analysis would still be valid if we let  $\beta_G$  denote the value  $\inf_{Q \leq Q^*} f'_G(Q)$  instead of  $f'_G(Q^*)$  and raise the compensation factor  $\alpha$  accordingly.

For *costs*, the convexity of the cost function (i.e., non-decreasing marginal costs) is more essential for our analysis but reflects the usual assumptions. A recent study (11) estimates actual marginal costs to be approximately linear, hence a model of linear benefits and marginal costs seems to be a plausible first approximation. However, we also assume that marginal costs are equalized for all countries by emissions trading and shared in a way fulfilling our convexity assumption, and whether this is justified depends on whether the market has perfect competition or whether instead prices can be influenced strategically by some countries with “market power”, as it is assumed in some models (12). Future research should investigate this issue in detail.

Another issue is that of *risk and uncertainty*. In a more accurate model, benefits of reductions in period  $t$  would be an uncertain and/or risky quantity (13; 14), e.g., due to the unknown value of future GDP and the fact that for a stock pollutant, emissions-related damages in  $t$  may depend on earlier emissions in a non-linear way. Much of the existing literature on cooperation in this game assume this non-linearity is small enough for a game-theoretic analysis to disregard it (1; 2; 8; 15). On the question of discounting in the context of the emissions game, see (16; 17), where it is also argued that

risky payoffs can be treated as risk-free by using their expected value and a lower, “risk-free” discounting rate. One might also want to use a more general payoff function of the form  $h_i(b_i - c_i)$  with concave increasing functions  $h_i$ , for which we conjecture our results will still hold.

Finally, whether the assumption of *complete information* is a reasonable approximation will have to be checked carefully, even when risk has been accounted for as indicated. Uncertain information about past contributions may be overcome by improving monitoring possibilities (18; 19), increasing the period length, or basing liabilities on earlier periods by replacing  $\ell_i(t+1)$  by  $\ell_i(t+k)$  for some  $k > 1$  in Eqn. [4]. The assumption that no country has a significant possibility to bindingly commit itself to certain future contributions has to be evaluated in light of the possibility of early investment decisions. Whether countries can be considered to be *rational players* in the sense of classical game-theory or exhibit some form of bounded rationality (20), and whether they cannot enter legally binding agreements that are not self-enforcing in the sense discussed, but can somehow be enforced by other means external to the considered game (e.g., international bodies or trade sanctions), are difficult questions of political science and international legal theory which are beyond the scope of this article. In this context, models that link emissions reductions with other issues (21), and approaches based on agent-based modeling (22), learning theory (23), or complex networks (24) are important contributions. Also, decision-makers might include criteria such as reputation and relative status in their reasoning, or might be influenced by citizens' altruistic attitudes towards public goods problems (25).

## SI: Renegotiations when targets are not optimal

Let us drop the assumption that the global target  $Q^*$  maximizes total payoff. Then LinC is no longer Pareto-efficient, hence not strongly renegotiation-proof, but is still weakly renegotiation-proof and also has the following property if  $\alpha$  is large enough: Assume some group  $G$  of players can profit from free-riding in a period  $t$  and then renegotiating a new strategy  $s$  with the others that all will follow from  $t+1$  on. Then there is another strategy  $\tilde{s}$  that all players outside  $G$  strictly prefer to play from  $t+1$  on over playing  $s$ , and so that  $G$ 's long-term payoff from  $t$  on is smaller than if all had continued to play LinC. We will prove below that this strategy  $\tilde{s}$  can be chosen so that it simply consists in continuing to play LinC, but with a new set of targets  $q_i^*$  from  $t+1$  on, and taking into account in  $t+1$  the shortfalls in  $t$ . In other words, the “meta-strategy” of *sticking to LinC and only changing the targets when necessary* deters any attempts of free-riding followed by renegotiation.

The proof is this: As all would agree to play  $s$  from  $t+1$  on, it must increase  $U_i(t+1)$  for all  $i$ , hence it must increase  $\sum_i U_i(t+1) = \sum_{r \geq 0} w_{t+1,r} P(t+1+r)$ . Thus the supremum of the new total period payoffs,  $P^+ = \sup_{r \geq 0} P(t+1+r)$ , exceeds the original target payoffs and is finite since payoffs are bounded from above. Since total payoffs  $f(Q) - g(Q)$  are a continuous function

of  $Q$ , there is a value  $Q^+$  for which they equal  $P^+$ . So any strategy  $\tilde{s}$  that has total contributions  $Q^+$  from  $t+1$  on gives at least the same value of  $\sum_i U_i(t+1)$  as  $s$  does. In particular, this is true if  $\tilde{s}$  consists in applying LinC with any targets  $q_i^+$  instead of  $q_i^*$ , as long as  $\sum_i q_i^+ = Q^+$ . Since each  $U_i(t+1)$  is a linear function of the targets  $q_i^+$ , the latter can also be chosen so that for each individual  $i$ ,  $U_i(t+1)$  is larger for  $\tilde{s}$  than for  $s$ . Let  $q_i^0$  be those targets and consider the alternative targets  $q_i^+ = q_i^0 + (n - |G|)\lambda$  for  $i \in G$  and  $q_i^+ = q_i^0 - |G|\lambda$  for  $i \notin G$ , with some  $\lambda > 0$ . Then  $U_i(t+1)$  is still larger for  $\tilde{s}$  than for  $s$  for all  $i \notin G$ , and  $U_G(t+1)$  is linearly decreasing with increasing  $\lambda$ . Now let  $s_0$  be the strategy of applying LinC with the original targets  $q_i^*$  and consider these four cases:

- (i) all play  $s_0$  from  $t$  on,
- (ii)  $G$  free-rides in  $t$  and all continue with  $s_0$  from  $t+1$  on,
- (iii)  $G$  free-rides in  $t$  and all switch to  $s$  from  $t+1$  on,
- (iv)  $G$  free-rides in  $t$  and all switch to  $\tilde{s}$  from  $t+1$  on.

We already know that  $U_G(t)$  is larger in case (i) than in case (ii) and  $U_G(t+1)$  is larger in case (iii) than in cases (ii) and (iv). Hence  $\lambda$  can be chosen so that  $U_G(t)$  is smaller in case (iv) than in case (i), but  $U_G(t+1)$  is still larger in case (iv) than in case (ii). Since also  $U_i(t+1)$  is larger in case (iv) than in case (iii) for all  $i \notin G$ , this means that when  $G$  proposes switching to  $s$  after the free-riding, the rest has incentives to argue for switching to  $\tilde{s}$  instead which at  $t+1$  still all prefer to continuing with  $s_0$ , but which makes sure the free-riding by  $G$  did not pay in the long run.

### SI: Properties of the one-shot game

Here we consider the *one-shot* version of the game (also called the *stage game* of the repeated game) in which only one period is played and a strategy just consists of choosing the individual contributions  $q_i$  of that period.

**Pareto-efficient contributions.** Since the game has transferable utility and the total period payoff  $P$  has a unique maximum  $P^*$  for  $Q = Q^*$ , a vector of individual contributions  $q_i$  is Pareto-efficient if and only if  $\sum_i q_i = Q^*$ .

**Pure-strategy equilibria.** A pure-strategy equilibrium is a strong form of Nash equilibrium in which strategies do not use randomization. Let  $Q_{-i} = Q - q_i$  be the joint contributions of all players except  $i$ . A *best response*  $q_i$  of player  $i$  to a given value of  $Q_{-i}$  is a value of  $q_i$  that maximizes the individual period payoff  $P_i$ . A best response must make total contributions non-negative,  $Q \geq 0$ , since for  $Q < 0$  we have  $\partial P_i / \partial q_i \geq f'_i(0) > 0$ . Hence  $q_i \geq -Q_{-i}$ . Note that for  $q_i = -Q_{-i}$  we have  $P_i = 0$ , and for  $q_i \rightarrow +\infty$  we have  $P_i \rightarrow -\infty$ . Thus, for any best response  $q_i$  to  $Q_{-i}$ , either (i)  $q_i > -Q_{-i}$  and  $P_i$  as a function of  $q_i$  has a local maximum with  $P_i > 0$  at  $q_i$ , in particular  $\partial P_i / \partial q_i = 0$ , or (ii)  $q_i = -Q_{-i}$  and  $P_i$  as a function of  $q_i$  has a global maximum  $P_i = 0$  at that value. Let  $f = \sum_{i=1}^n f_i$ .

Now a *pure-strategy equilibrium (PSE)* is a vector of contributions  $q_i$  for all  $i$  such that  $q_i$  is a best response to  $Q_{-i}$  for all  $i$ . So for a PSE, either (i)  $Q > 0$  and  $\partial P_i / \partial q_i = 0$  for all  $i$ , or (ii)  $Q = 0$  and  $P_i$  as a function of  $q_i$  has a global maximum at that value for all  $i$ .

In the *proportional cost sharing case*, denote average unit costs by  $h(Q) = g(Q)/Q \geq 0$ , so that  $c_i = h(Q)q_i$  with  $h'(Q) = (g'(Q) - h(Q))/Q \geq 0$  for  $Q > 0$  and  $h'(Q) = 0$  for  $Q \leq 0$ . In case (i), we then have  $\partial P_i / \partial q_i = f'_i(Q) - h(Q) - q_i h'(Q) = 0$ , and taking the sum over all  $i$  gives  $0 = f'(Q) - nh(Q) - Qh'(Q)$  or  $f'(Q) = (n-1)h(Q) + g'(Q)$ . Since the left-hand side is non-increasing, the right-hand side non-decreasing, and since  $f'(Q^*) = g'(Q^*)$ , this condition has at least one (and often unique) solution  $Q^{\text{PSE}} > 0$  for which

$$f'(Q^{\text{PSE}}) = (n-1)h(Q^{\text{PSE}}) + g'(Q^{\text{PSE}}). \quad [8]$$

Given  $Q^{\text{PSE}}$ , the individual conditions  $\partial P_i / \partial q_i = 0$  have a unique solution

$$q_i^{\text{PSE}} = \frac{f'_i(Q^{\text{PSE}}) - h(Q^{\text{PSE}})}{h'(Q^{\text{PSE}})} \quad [9]$$

if  $h'(Q^{\text{PSE}}) > 0$ . This solution leads to individual period payoffs

$$P_i^{\text{PSE}} = f_i(Q^{\text{PSE}}) + h(Q^{\text{PSE}}) \frac{h(Q^{\text{PSE}}) - f'_i(Q^{\text{PSE}})}{h'(Q^{\text{PSE}})}. \quad [10]$$

If  $P_i^{\text{PSE}} > 0$  for all  $i$ , this solution is the unique one-shot PSE with  $Q = Q^{\text{PSE}}$ . If  $h'(Q^{\text{PSE}}) = 0$  or some of the  $P_i^{\text{PSE}}$  are non-positive, the analysis is more complicated.

In the *marginal cost pricing case*, we have  $c_i = g_i(a_i(Q)) + [q_i - a_i(Q)]g'(Q)$ . In case (i), we then have  $\partial P_i / \partial q_i = f'_i(Q) - g'(Q) - (q_i - a_i(Q))g''(Q) = 0$ , and taking the sum over all  $i$  gives the following condition for a PSE which again has at least one, typically unique solution  $Q^{\text{PSE}}$ :

$$f'(Q^{\text{PSE}}) = ng'(Q^{\text{PSE}}). \quad [11]$$

Given  $Q^{\text{PSE}}$ , the individual conditions  $\partial P_i / \partial q_i = 0$  have a unique solution

$$q_i^{\text{PSE}} = a_i(Q^{\text{PSE}}) + \frac{f'_i(Q^{\text{PSE}}) - g'(Q^{\text{PSE}})}{g''(Q^{\text{PSE}})} \quad [12]$$

if  $g''(Q^{\text{PSE}}) > 0$ . This solution leads to individual period payoffs

$$P_i^{\text{PSE}} = f_i(Q^{\text{PSE}}) - g_i(a_i(Q^{\text{PSE}})) - g'(Q^{\text{PSE}}) \frac{f'_i(Q^{\text{PSE}}) - g'(Q^{\text{PSE}})}{g''(Q^{\text{PSE}})}. \quad [13]$$

Because  $P^{\text{PSE}} < P^*$ , there are allocations  $q_i$  of the total optimal contributions  $Q^*$  that give each  $i$  a strictly higher payoff than  $P_i^{\text{PSE}}$ . Hence each such PSE is Pareto-dominated but may serve as a kind of benchmark in negotiations of the target allocation  $q_i^*$  in the sense that one could restrict attention to target allocations that Pareto-dominate the PSE. This idea to use non-cooperative solutions as a benchmark for cooperative ones is also used in the context of consensus voting rules (26). See also SI: Cooperative analysis.

### SI: Examples

**Linear benefits, monomial costs.** Many examples from the literature are of the following form:

- Individual benefits  $f_i(Q) = \beta_i Q$  with  $f'_i(Q) = \beta_i > 0$ .
- Total marginal benefits  $f'(Q) = \beta = \sum_i \beta_i > 0$ .
- Total costs  $g(Q) = \max\{Q, 0\}^\zeta$  with  $\zeta > 1$ .
- Marginal total costs  $g'(Q) = \zeta \max\{Q, 0\}^{\zeta-1}$ .
- Total period payoff  $P(Q) = \beta Q - \max\{Q, 0\}^\zeta$  with  $P'(Q) = \beta - \zeta \max\{Q, 0\}^{\zeta-1}$ .

The optimal total contributions  $Q^* > 0$  then fulfill

$$0 = P'(Q^*) = \beta - \zeta(Q^*)^{\zeta-1},$$

hence

$$Q^* = \gamma_0^{\frac{1}{\zeta-1}},$$

$$P^* = (\zeta - 1)\gamma_0^{\frac{\zeta}{\zeta-1}},$$

where  $\gamma_0 = \beta/\zeta$  are the target average unit costs.

In the *proportional cost sharing case*, we then have average unit costs  $h(Q) = \max\{Q, 0\}^{\zeta-1}$  with  $h'(Q) = (\zeta-1)Q^{\zeta-2}$  for  $Q > 0$  and  $h'(Q) = 0$  for  $Q < 0$ , leading to individual period payoffs

$P_i = \beta_i Q - q_i h(Q)$  with  $\partial P_i / \partial q_i = \beta_i - (\zeta - 1)(q_i + Q)Q^{\zeta-2}$  for  $Q > 0$  and  $\partial P_i / \partial q_i = \beta_i$  for  $Q < 0$ . A one-shot PSE has  $Q^{\text{PSE}} > 0$  and fulfills

$$\begin{aligned} 0 &= f'(Q^{\text{PSE}}) - (n-1)h(Q^{\text{PSE}}) - g'(Q^{\text{PSE}}) \\ &= \beta - (n-1 + \zeta)(Q^{\text{PSE}})^{\zeta-1}, \end{aligned}$$

hence the unique PSE is given by

$$\begin{aligned} Q^{\text{PSE}} &= \tilde{\beta}^{\frac{1}{\zeta-1}}, \\ q_i^{\text{PSE}} &= \frac{\beta_i - \tilde{\beta}}{\zeta-1} \tilde{\beta}^{\frac{2-\zeta}{\zeta-1}}, \\ P_i^{\text{PSE}} &= \frac{(\zeta-2)\beta_i + \tilde{\beta}}{\zeta-1} \tilde{\beta}^{\frac{1}{\zeta-1}}, \\ P^{\text{PSE}} &= (n-2 + \zeta)\tilde{\beta}^{\frac{\zeta}{\zeta-1}} = O\left(P^*/n^{\frac{1}{\zeta-1}}\right), \end{aligned}$$

where  $\tilde{\beta} = \beta/(n-1 + \zeta)$  is slightly smaller than the average individual marginal benefits. For  $\zeta = 2$ , total period payoff is then shared equally between players, and individual payoffs are  $P_i^{\text{PSE}} \propto 1/(n+1)^2$ , bearing a surprising similarity to Cournot-Nash payoffs in Cournot oligopolies (see also SI: Cooperative analysis). For  $\zeta > 2$ , part of it is shared in proportion to marginal benefits, while for  $\zeta < 2$ , those with larger marginal benefits get smaller payoffs.

In the *marginal cost pricing case*, the one-shot PSE fulfills

$$\beta = ng'(Q^{\text{PSE}}) = n\zeta(Q^{\text{PSE}})^{\zeta-1},$$

hence the unique PSE is given by

$$\begin{aligned} Q^{\text{PSE}} &= \tilde{\beta}^{\frac{1}{\zeta-1}}, \\ q_i^{\text{PSE}} &= a_i(\tilde{\beta}^{\frac{1}{\zeta-1}}) + \frac{\beta_i/\zeta - \tilde{\beta}}{\zeta-1} \tilde{\beta}^{\frac{2-\zeta}{\zeta-1}}, \\ P_i^{\text{PSE}} &= \frac{\zeta}{\zeta-1} \tilde{\beta}^{\frac{\zeta}{\zeta-1}} + \frac{\zeta-2}{\zeta-1} \beta_i \tilde{\beta}^{\frac{1}{\zeta-1}} - g_i(a_i(\tilde{\beta}^{\frac{1}{\zeta-1}})), \\ P^{\text{PSE}} &= (n\zeta - 1)\tilde{\beta}^{\frac{\zeta}{\zeta-1}} = O\left(P^*/n^{\frac{1}{\zeta-1}}\right), \end{aligned}$$

where  $\tilde{\beta} = \beta/n\zeta$  this time.

**Decreasing marginal benefits, quadratic costs.** A simple model with decreasing instead of constant marginal benefits that can still be solved analytically is this:

- Individual benefits  $f_i(Q) = \beta_i \ln(1+Q)$  for  $Q \geq 0$  and  $f_i(Q) = \beta_i(Q - Q^2/2 + Q^3/3)$  for  $Q \leq 0$ , with  $\beta_i > 0$ .
- Individual marginal benefits  $f'_i(Q) = \beta_i/(1+Q)$  for  $Q \geq 0$  and  $f'_i(Q) = \beta_i(1-Q+Q^2)$  for  $Q \leq 0$ .
- Total costs  $g(Q) = \max\{Q, 0\}^2$ .
- Marginal total costs  $g'(Q) = \max\{2Q, 0\}$ .
- Total period payoff for  $Q \geq 0$ :  
 $P(Q) = \beta \ln(1+Q) - Q^2$  with  $P'(Q) = \beta/(1+Q) - 2Q$ .

Optimal total contributions  $Q^* > 0$  must fulfill

$$0 = P'(Q^*) = \beta/(1+Q^*) - 2Q^*,$$

hence

$$\begin{aligned} Q^* &= \frac{\sqrt{1+2\beta} - 1}{2}, \\ P^* &= \beta \ln \frac{\sqrt{1+2\beta} + 1}{2} + \frac{\sqrt{1+2\beta} - 1 - \beta}{2}. \end{aligned}$$

In the *proportional cost sharing case*, we then have average unit costs  $h(Q) = \max\{Q, 0\}$  with  $h'(Q) = 1$  for  $Q > 0$  and

$h'(Q) = 0$  for  $Q < 0$ , leading to individual period payoff for  $Q \geq 0$ :  $P_i = \beta_i \ln(1+Q) - q_i Q$  with  $\partial P_i / \partial q_i = \beta_i/(1+Q) - q_i - Q$ . The one-shot PSE has  $Q^{\text{PSE}} > 0$  and thus fulfills

$$\begin{aligned} 0 &= f'(Q^{\text{PSE}}) - (n-1)h(Q^{\text{PSE}}) - g'(Q^{\text{PSE}}) \\ &= \beta/(1+Q^{\text{PSE}}) - (n+1)Q^{\text{PSE}}, \end{aligned}$$

hence the unique PSE is given by

$$\begin{aligned} Q^{\text{PSE}} &= \frac{\varrho - 1}{2}, \\ q_i^{\text{PSE}} &= 2 \frac{\beta_i - \tilde{\beta}}{\varrho + 1}, \\ P_i^{\text{PSE}} &= \beta_i \left( \ln \frac{\varrho + 1}{2} - \frac{\varrho - 1}{\varrho + 1} \right) + \tilde{\beta} \frac{\varrho - 1}{\varrho + 1}, \\ P^{\text{PSE}} &= \beta \ln \frac{\varrho + 1}{2} + \frac{\varrho - 1}{2} - \tilde{\beta}, \end{aligned}$$

where  $\tilde{\beta} = \beta/(n+1)$  is slightly smaller than the average individual marginal benefits, and  $\varrho = \sqrt{1+4\tilde{\beta}}$ . Again, part of the total period payoff is shared in proportion to marginal benefits, and that part grows with  $\beta$ .

For small  $\beta$  and large  $n$ ,  $P^* \approx \beta^2/4$  and  $P^{\text{PSE}} \approx \beta^2/n = O(P^*/n)$ , i.e., the cooperative payoff is of the order  $n$  larger than the PSE payoff. For large  $\beta \gg n$ ,  $P^* \approx (\beta \ln \beta)/2 \approx P^{\text{PSE}}$ , i.e., the cooperative and PSE payoffs are approximately equal.

In the *marginal cost pricing case*, the one-shot PSE fulfills

$$\beta/(1+Q^{\text{PSE}}) = ng'(Q^{\text{PSE}}) = 2nQ^{\text{PSE}},$$

hence the unique PSE is given by

$$\begin{aligned} Q^{\text{PSE}} &= \frac{\varrho - 1}{2}, \\ q_i^{\text{PSE}} &= a_i(Q^{\text{PSE}}) + \beta_i - 1 - \tilde{\beta} - \varrho, \\ P^{\text{PSE}} &= \beta \ln \frac{\varrho + 1}{2} + \frac{\varrho - 1 - \tilde{\beta}}{2}, \end{aligned}$$

where  $\tilde{\beta} = \beta/n$  and  $\varrho = \sqrt{1+2\tilde{\beta}}$  this time.

As above, for small  $\beta$  and large  $n$ ,  $P^* \approx \beta^2/4$  and  $P^{\text{PSE}} \approx \beta^2/2n = O(P^*/n)$ , and for large  $\beta \gg n$ ,  $P^* \approx (\beta \ln \beta)/2 \approx P^{\text{PSE}}$ .

**Diverging costs for some maximal contributions.** A simple model in which contributions are effectively bounded from *above* by diverging costs is this:

- Linear individual benefits  $f_i(Q) = \beta_i Q$  with  $\beta_i > 0$ .
- Individual marginal benefits  $f'_i(Q) = \beta_i$ .
- Total costs  $g(Q) = Q^2/(1-Q)$  for  $Q \in [0, 1)$  and  $g(Q) = 0$  for  $Q < 0$ .
- Marginal total costs  $g'(Q) = Q(2-Q)/(1-Q)^2$  for  $Q \in [0, 1)$ .
- Total period payoff  $P(Q) = \beta Q - Q^2/(1-Q)$  for  $Q \in [0, 1)$ , with  $P'(Q) = \beta - Q(2-Q)/(1-Q)^2$ .

Optimal total contributions  $Q^* \in (0, 1)$  fulfill

$$0 = P'(Q^*) = \beta - Q^*(2-Q^*)/(1-Q^*)^2,$$

hence

$$\begin{aligned} Q^* &= 1 - 1/\sqrt{\beta+1}, \\ P^* &= \beta + 2 - 2\sqrt{\beta+1}. \end{aligned}$$

In the *proportional cost sharing case*, we then have average unit costs  $h(Q) = Q/(1-Q)$  for  $Q \in [0, 1)$ , with  $h'(Q) = 1/(1-Q)^2$ , leading to individual period payoff  $P_i = \beta_i Q - q_i Q/(1-Q)$  for

$Q \in [0, 1)$ , with  $\partial P_i / \partial q_i = \beta_i - Q / (1 - Q) - q_i / (1 - Q)^2$ . The one-shot PSE has  $Q^{\text{PSE}} \in (0, 1)$  and thus fulfills

$$\begin{aligned} 0 &= f'(Q^{\text{PSE}}) - (n-1)h(Q^{\text{PSE}}) - g'(Q^{\text{PSE}}) \\ &= \beta - Q^*(n+1 - nQ^*) / (1 - Q^*)^2, \end{aligned}$$

hence the unique PSE is given by

$$Q^{\text{PSE}} = 1 - \frac{\frac{n-1}{2} + \sqrt{\beta + n + (\frac{n-1}{2})^2}}{\beta + n}.$$

For large  $n$ ,  $P^{\text{PSE}} \approx \beta^2/n = O(P^*/n)$ , i.e., the cooperative payoff is of the order  $n$  larger than the PSE payoff.

In the *marginal cost pricing case*, the unique PSE is given by

$$Q^{\text{PSE}} = 1 - 1/\sqrt{\tilde{\beta} + 1},$$

where  $\tilde{\beta} = \beta/n$ . For large  $n$ , again  $P^{\text{PSE}} \approx \beta^2/2n = O(P^*/n)$ .

### SI: Bounded liabilities

In some applications, it might be desirable or necessary to *restrict* the range of possible liabilities LinC might allocate in reaction to deviations. Let's assume liabilities must be bounded by some lower bounds  $\ell_i^{\min} < q_i^*$  for all players  $i$ , so that only liabilities with  $\ell_i(t) \geq \ell_i^{\min}$  are feasible allocations. E.g., if individual contributions  $q_i$  cannot be negative, one could choose  $\ell_i^{\min} = 0$ . Any strategy that still keeps total liabilities fixed to the optimal target  $Q^*$  in order to be strongly renegotiation-proof can then assign any group  $G$  of players at most the liability  $L_G^{\max} = Q^* - \sum_{i \notin G} \ell_i^{\min}$ .

We suggest to use the following modified strategy of *Bounded Linear Compensations (BLinC)* in that case: For those players  $i$  without shortfalls in  $t$ , liabilities in  $t+1$  are calculated as in LinC, but are capped at their lower bounds. For those with shortfalls, the liability adjustments are then scaled down to keep the total target:

$$\ell_i(t+1) = \begin{cases} \max\{q_i^* + [d_i(t) - \bar{d}(t)] \cdot \alpha, \ell_i^{\min}\} & \text{if } d_i(t) = 0 \\ q_i^* + [d_i(t) - \bar{d}(t)] \cdot \alpha / s(t) & \text{if } d_i(t) > 0, \end{cases} \quad [14]$$

where  $s(t) > 0$  is chosen so that  $\sum_i \ell_i(t+1) = Q^*$ . If shortfalls are moderate so that  $\bar{d}(t) \leq (q_i^* - \ell_i^{\min})/\alpha$  for all  $i$  with  $d_i(t) = 0$ , then  $s(t) = 1$  and the allocation is the same as in LinC (Eqn. [4]).

While LinC's subgame-perfectness follows from the ability to assign additional liabilities proportional to a large enough multiple of the shortfalls, BLinC can do so no longer in case of large shortfalls. Hence it depends on the choice of the bounds  $\ell_i^{\min}$  and on the discounting factor  $\delta$  whether BLinC is subgame-perfect or not.

Note that, in the *proportional cost-sharing case*, the gain that any group  $G$  of players would get from a shortfall of size  $x \geq 0$  in a situation in which its liability is already maximal,  $L_G = L_G^{\max}$ , is at most  $L_G^{\max}\gamma^* - (L_G^{\max} - x)\gamma_x - x\beta_G$ , where  $\gamma_x = g(Q^* - x)/(Q^* - x)$  are the average unit costs at  $Q^* - x$ , with  $\gamma_x \leq \gamma^*$  because average costs are non-decreasing. And the discounted loss that  $G$  would have in  $t+1$  from having assigned maximal liabilities again is at least  $\delta\gamma^*(L_G^{\max} - Q_G^*)$ . Hence a sufficient condition for such a shortfall to be unprofitable is that the former be smaller than the latter, which is equivalent to

$$x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x > Q_G^*\gamma^*\delta + L_G^{\max}\gamma^*(1 - \delta). \quad [15]$$

We can now show the following 'strong' form of a 'folk theorem': if the target allocation  $q^*$  is *profitable* for each player, so that  $Q^*\beta_G - Q_G^*\gamma^* > 0$  for all  $G$ , and if  $\delta$  is close enough to unity and the bounds  $\ell_i^{\min}$  are small enough, then the above condition is fulfilled for all  $G$  and all  $x \geq 0$ . Let  $\varepsilon_G = (Q^*\beta_G - Q_G^*\gamma^*)/2(1 + \beta_G) > 0$ ,

$\varepsilon = \min_G \varepsilon_G > 0$ , and  $x_0 = Q^* - \varepsilon$ . Choose the bounds  $\ell_i^{\min}$  small enough so that  $L_G^{\max} \geq Q^*$  and  $L_G^{\max} > (Q_G^*\gamma^* + \varepsilon - x_0(\beta_G - \gamma^*))/\gamma_{x_0}$  for all  $G$ . Then, for all  $G$  and  $x$ ,

$$x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x > Q_G^*\gamma^* + \varepsilon. \quad [16]$$

This is because (i) for  $x \in [x_0, L_G^{\max}]$ , we have  $x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x \geq x\beta_G \geq (Q^* - \varepsilon_G)\beta_G > Q_G^*\gamma^* + \varepsilon_G \geq Q_G^*\gamma^* + \varepsilon$ , (ii) for  $x \geq L_G^{\max} \geq Q^*$ , we have  $\gamma_x = 0$  and thus  $x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x = x\beta_G \geq Q^*\beta_G > Q_G^*\gamma^* + \varepsilon$ , and (iii) for  $x \leq x_0$ , we have  $0 \leq \gamma_{x_0} \leq \gamma_x \leq \gamma^*$  and thus  $x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x \geq x_0(\beta_G - \gamma^*) + L_G^{\max}\gamma_{x_0} > Q_G^*\gamma^* + \varepsilon$ . Now if  $\delta$  is close enough to unity,  $Q_G^*\gamma^* + \varepsilon > Q_G^*\gamma^*\delta + L_G^{\max}\gamma^*(1 - \delta)$  and the claim is proved.

This means that, for large enough compensation factor  $\alpha$ , no group of players has ever an incentive to deviate from BLinC for one period, and thus neither for a finite number of periods. In contrast to LinC, the bounds on liabilities in BLinC imply that also the possible payoffs are bounded. Hence a standard argument as in (27) shows that then also no infinite number of deviations can pay. In particular, this shows that with individually profitable targets  $q_i^*$ , large enough  $\alpha$  and  $\delta$ , and small enough  $\ell_i^{\min}$ , the modified strategy BLinC is still subgame-perfect.

In the example with linear benefits and marginal costs ( $\zeta = 2$ ), and for  $\ell_i^{\min} = 0$  (non-negative liabilities), Eqn. [15] is fulfilled when  $Q_G^* < \min\{\beta_G - \beta(1 - \delta)/2, \beta_G(1 - \beta_G/2\beta)/\delta - \beta(1 - \delta)/2\delta\}$  for all  $G$ . For large enough  $\delta$ , this can be fulfilled by a target allocation proportional to marginal benefits,  $q_i^* = \beta_i/2$ . That allocation leads to payoffs which are also proportional to marginal benefits,  $P_i = \beta_i\beta/4$ .

In the *marginal cost pricing case*, a condition similar to Eqn. [15] can be found, but no similar general existence result as above is obvious.

For the emissions game, we simulated whether BLinC can be used instead of LinC in a slightly modified version of the STACO cost-benefit-model which is frequently used in the literature (1; 2; 8).<sup>1</sup> and which calculates time-dependent individual benefit functions  $b_i(t)$  and a time-dependent global optimal emissions abatement path  $Q^*(t)$ . For the chosen model parameters, a moderate  $\alpha$  of 1.22 fulfills Eqn. [5]. We tested several possible allocations of the global target  $Q^*(t)$  into time-dependent individual targets  $q_i^*(t)$  under the assumption of proportional cost sharing. A promising compromise between an "egalitarian" and a "grandfathering" allocation distributes half of the long-term global payoff as compared to the business-as-usual scenario in a way so that each region's per capita payoff in purchasing power (PPP) increases by the same amount, and the other half in proportion to regional GDP (based on 1995 population, PPP, and GDP data).<sup>2</sup> That allocation gives four players negative contribution targets  $q_i^*(t)$ , i.e., more emissions permits than under business-as-usual, which is often termed "hot air" in the policy literature, so that those players can profit from selling unused permits on the market. When the liability bounds  $\ell_i^{\min}(t)$  were chosen so that those four players never have liabilities lower than twice this (negative) value, and all others never have negative liabilities at all, we could verify using Eqn. [15] that none of the 4095 possible groups of players ever had incentives to deviate from BLinC. An alternative allocation that completely achieved equal per capita payoffs in PPP did not allow to use the same kind of bounds  $\ell_i^{\min}(t)$  since then some groups of in-

<sup>1</sup>Model parameters: 12 players (economic world regions); 2-year periods; exponential discounting at 2% yearly; costs based on cubic regional abatement cost functions as estimated by (28); benefits = avoided emissions-related economic damages in linear approximation, properly discounted; damages estimated as 2.7% of regional GDP if atmospheric GHG concentrations double; GDP estimated with the DICE integrated assessment model (1994 version with "no controls", scenario B2 (29)); play simulated from 2010–2110.

Other authors use similar models, e.g. the WITCH model with 12 regions in (31) or the EPPA model with 16 regions in (32).

<sup>2</sup>This is basically the average of the sharing rules 2 and 3 from (33), for which those authors found that only very small long-term coalitions were stable in their model. For a more recent study of the distributional effects of various sharing rules see (32).

dustrialized regions could profit from free-riding. Still, with a different cost-benefit model and different liability bounds this completely egalitarian allocation rule might work. If marginal cost pricing is assumed instead of proportional cost sharing with the same model and bounds, both allocation rules gave several groups of industrialized countries incentives to free-ride. Using numerical optimization, we were able to find for sufficiently large values of  $\delta$  some alternative sets of liability bounds that would remove these incentives, but these results are very preliminary. Future work should assess this question more thoroughly, and using more elaborate cost-benefit models.

## SI: Cooperative analysis

We proved that for each conceivable target allocation, playing LinC constitutes a strong form of strategic equilibrium that realizes this allocation. Hence the problem of negotiating a target allocation can be seen as a problem of selecting a particular equilibrium of the game.

The game-theoretic literature does not answer clearly which equilibria rational players can, will, or should select in a game that has many equilibria, and there are quite different approaches to this.

**Coalition formation.** One approach is to envision that players might end up partitioned into some *coalition structure*  $\pi = \{S_1, \dots, S_m\}$ , i.e., a partition of all players into  $m$  disjoint coalitions of one or more players each, who will cooperate internally but not with each other. The coalition structure  $\{N\}$  in which all players cooperate is called the *grand coalition*. In the public good game, such a coalition structure can reach a large number of alternative equilibria as follows: Consider the  $m$ -player version of the game in which each coalition  $S_j$  is treated as one player with benefit function  $f_{S_j} = \sum_{i \in S_j} f_i$ , and let  $(Q_1^{\text{PSE}}, \dots, Q_m^{\text{PSE}})$  be the contributions in a PSE of this game. These can be determined by replacing  $n$ ,  $f_i$ , and  $q_i$  in Eqns. 8 and 9 by  $m$ ,  $f_{S_j}$ , and  $Q_j^{\text{PSE}}$ , respectively. Now assume each  $S_j$  has agreed internally on some individual target allocation  $q_i^*$  of  $Q_j^{\text{PSE}}$ , so that  $\sum_{i \in S_j} q_i^* = Q_j^{\text{PSE}}$ , and applies LinC to these targets *internally* (i.e., ignoring players outside  $S_j$  in the calculation of liabilities). Then it is easy to see that this constitutes an equilibrium of the whole game with similar stability properties as when LinC is applied by the grand coalition, but total payoffs are sub-optimal when the coalition structure is not the grand coalition.

Let  $v^{\text{PSE}}(S_j, \pi)$  be the joint payoff of  $S_j$  in such an equilibrium, given the coalition structure  $\pi$ . Then it might be considered plausible that  $v^{\text{PSE}}(S_j, \pi)$  is the joint payoff that the players in  $S_j$  can expect to get should initial negotiations lead to the coalition structure  $\pi$ .

Both classical “cooperative” game theory and the newer more sophisticated theory of coalition formation (34) now try to predict which coalition structures might arise and what allocations the coalitions will agree to, by only considering what each coalition can expect to get given each coalition structure, and assuming players can influence the coalition structure in various ways independent from those payoffs, by individually or jointly leaving, joining, or blocking coalitions. Such an analysis then only depends on the *partition function*  $v = v^{\text{PSE}}$ . Depending on the precise assumptions, that theory sometimes somewhat surprisingly predicts that not the grand coalition but a partition into more than one coalition will form, resulting in sub-optimal payoffs.<sup>3</sup>

Consider for example the public good game with linear benefits  $f_i(Q) = \beta_i Q$ , quadratic costs  $g(Q) = \max\{0, Q\}^2$ , and proportional cost sharing  $c_i = Cq_i/Q$ . Then it can be shown (see SI: Examples) that  $v^{\text{PSE}}$  has a particularly simple form that only depends on the number of coalitions and not at all on their size or their individual benefit functions:  $v(S_j, \pi) = A/(|\pi| + 1)^2$  for some constant  $A$ . (If marginal cost pricing is used and all individual cost curves are identical, one gets the very similar form  $v(S_j, \pi) = A(|\pi| - 3/2)/|\pi|^3$  instead). This extreme form of  $v$  has been analysed in the literature as a kind of quintessential example of cooperative games with ex-

*ternalities* since it also arises naturally from Cournot-Nash equilibria in *Cournot oligopolies*<sup>4</sup> For  $n = 5$  (and similarly for larger  $n$ ), one approach (35) predicts that a coalition structure with one individual player  $S_1$  and two coalitions  $S_2, S_3$  consisting of two players each will arise, each coalition getting a payoff of  $1/16$ . The argument for this is that any allocation of the grand coalition’s payoff of  $1/4$  must give at least one player at most a payoff of  $1/20 < 1/16$ , so that that player will leave the grand coalition and the remaining four players will then split in two pairs for similar reasons. Another approach by the same authors (36) assumes that the actual bargaining process follows a certain particular protocol and predicts that the result is one individual player and a coalition of the remaining four players, not splitting any further into two pairs. Other authors (37; 38) arrive at still different coalition structures for different values of  $n$  (e.g.,  $n = 6$ ).

In such analyses, however, it remains unclear why the predicted coalitions should not afterwards negotiate an additional agreement with each other in order to realize and share also the additional total payoff that is possible by forming the grand coalition. Following Coase (39), such behavior should always be expected so that only optimal allocations can result. We support this point of view with an analysis of the case  $n = 5$  of the above example, in the next subsection.

Other choices of the partition function  $v$  than  $v^{\text{PSE}}$  might also be plausible. Assume players can make each other believe that, should no global agreement be reached, they will contribute nothing. Then each coalition  $S_j$  can only expect to benefit from its own contributions, resulting in a maximal payoff  $v^0(S_j, \pi) = v^0(S_j)$  that only depends on  $S_j$  (actually only on the functions  $f_{S_j}$  and  $g$ ) and is *superadditive*:  $v^0(S_j \cup S_k) \geq v^0(S_j) + v^0(S_k)$ . For such superadditive *value functions*, a rich literature exists which holds that the grand coalition will indeed form. Its most prominent solution concept is the *Shapley value* (40) which suggests that player  $i$ ’s share of  $v(N)$  should be a certain linear combination of the differences  $v(S \cup \{i\}) - v(S)$  for all  $S$  with  $i \notin S$ . For situations with players of unequal “size”, there are weighted versions of this (41) that give players with larger weight  $w_i$  (e.g., a country’s population in the emissions game) larger payoffs. Depending on the chosen weights, this can lead to any payoff allocation in the so-called *core* of the game (42). Given  $v$  and weights  $w_i$  with  $\sum_i w_i = 1$ , the (weighted) Shapley values are  $\phi_i = w_i[P(N) - P(N \setminus \{i\})]$ , where the *potential function*  $P$  is defined recursively as  $P(\emptyset) = 0$  and

$$P(S) = \left[ v(S) + \sum_{i \in S} w_i P(S \setminus \{i\}) \right] / \sum_{i \in S} w_i. \quad [17]$$

A third choice of  $v$  relies on the assumption that players can make each other believe that, should no global agreement be reached, they will not enter any other agreement with a smaller coalition but still maximize their individual payoff by playing a best response of the one-shot game. In that case, we get the value function  $v(N) = P^*$  and  $v(S) = \sum_{i \in S} P_i^{\text{PSE}}$  for  $S \neq N$ , which is not only superadditive but even additive for all coalitions except the grand coalition. Such a situation is often called a *pure bargaining* or *unanimity game*, and its weighted Shapley values are simply  $\phi_i = P_i^{\text{PSE}} + w_i(P^* - P^{\text{PSE}})$ , that is, the surplus from cooperation is shared in proportion to the weights. In the example of linear benefits and marginal costs, the weighted Shapley values are then proportional to  $4 + w_i(n - 1)^2$ .

**Coalition formation when inter-coalitional agreements are possible.** Before turning to a more general case, we present this idea

<sup>3</sup>Note that in many relating papers the authors use public goods examples with a different cost structure than ours, assuming non-decreasing individual marginal costs that depend on *individual* contributions,  $C_i = g_i(q_i)$ , instead of our assumption of marginal costs that depend only on *total* contributions,  $C_i = q_i g(Q)/Q$ .

<sup>4</sup>Although the corresponding game has a different individual payoff structure that cannot be interpreted as a public good game, only  $v$  is considered relevant in this line of reasoning.

by first discussing the example of five players with linear benefits and marginal costs, for which the value function has the form  $v(S_j, \pi) = 1/(|\pi| + 1)^2$ . Suppose the grand coalition, denoted by (12345), meets to negotiate an allocation of the total payoff of  $1/4$ , and the current proposal is to split it equally into  $5 \cdot 1/20$ . In (35) it is argued that each player, say player 1, can then hope to get  $1/16$  if he leaves the room, since he can then expect that (i) another pair, say players 23, will leave, so that the coalition structure (1, 23, 45) of one singleton and two pairs will arise, and that (ii) the resulting coalitions will then behave like three individual players, so that their payoffs are those in the PSE,  $1/16$  for each coalition.

But if those three coalitions would agree on an additional inter-coalitional agreement, they could realize a surplus of  $1/4 - 3/16 = 1/16$  and share it to everyone's profit. Collective rationality requires that we assume this would indeed happen, leading to some individual payoffs  $a_i$  with  $\sum_i a_i = 1/4$ ,  $a_1 \geq 1/16$ ,  $a_2 + a_3 \geq 1/16$ , and  $a_4 + a_5 \geq 1/16$ . A similar assumption must be made for any possible refinements of that structure that might arise should one of the players 2345 leave her coalition. If 2 leaves, the resulting structure is either (1, 2, 3, 4, 5) or (1, 2, 3, 45), depending on whether the latter can stabilize itself. Whether it can do so depends on what an additional leaving player, say 5, can expect to get.

If 5 leaves (1, 2, 3, 45), we get the all-singletons structure (1, 2, 3, 4, 5), and collective rationality implies that all five would then come back to the table and start a new round of negotiations, probably starting with the allocation that was discussed last for the grand coalition. As this allocation is  $5 \cdot 1/20$ , player 5 can hence expect to get  $1/20$  when leaving (1, 2, 3, 45). Collective rationality now requires that (1, 2, 3, 45) would not agree on an allocation  $b$  that destabilizes their structure, so we can assume that structure will stabilize itself like this: First, coalition 45 has an intra-coalitional agreement on how to share their PSE payoff of  $1/25$ , and then the four coalitions have an inter-coalitional agreement on how to share the additionally possible payoff of  $1/4 - 4/25$  that gives neither 4 nor 5 an incentive to leave. Hence all players can expect that, should the structure (1, 2, 3, 45) arise, their payoffs would be some  $b_i$  with  $\sum_i b_i = 1/4$ ,  $b_1 \geq 1/25$ ,  $b_2 \geq 1/25$ ,  $b_3 \geq 1/25$ ,  $b_4 \geq 1/20$ , and  $b_5 \geq 1/20$ .

Let us now assume that each player announces in advance to accept no less than  $1/20$  should the structure (1, 2, 3, 45) arise. This is certainly a credible announcement since it corresponds to the currently discussed allocation, can be realized by putting  $a_i = 1/20$ , leads to a stable agreement, and gives no incentive to deviate from it and accept less than  $1/20$  when the structure (1, 2, 3, 45) indeed arises. We will argue below that these announcement will finally stabilize the grand coalition. In other words, all players can expect the payoffs to be  $b_i = 1/20$  if structure (1, 2, 3, 45) arises, and similarly for all other structures with three singletons and a pair.

Now for the stability of (1, 23, 45): If  $a_2 < b_2 = 1/20$ , player 2 has an incentive to leave (1, 23, 45). A similar condition holds for players 345, so (1, 23, 45) is unstable if not  $a_2, a_3, a_4, a_5 \geq 1/20$ . But then  $a_1 \leq 1/4 - 4/20 = 1/20 < 1/16$ , so 1 would not agree on that allocation since he can realize  $1/16$  in the PSE. Hence (1, 23, 45) can *not* stabilize itself, in contrast to the expectation (ii) above, and will instead fall apart to give one of the stable coalitions (1, 2, 3, 45) and (1, 23, 4, 5).

Similarly, also a two-singletons-and-a-triple structure, say (1, 2, 345), cannot stably agree on a payoff allocation  $a'$ . It would require  $\sum_i a'_i = 1/4$ ,  $a'_1 \geq 1/16$ ,  $a'_2 \geq 1/16$ , and  $a'_3 + a'_4 + a'_5 \geq 1/16$ . But since  $1/4 - 2/16 < 3/20$ , one of  $a'_3, a'_4, a'_5$  must be smaller than  $1/20$ , so that that player would leave to get  $1/20$  in a four-singletons-and-a-pair structure.

Now we check expectation (i) by checking the stability of (1, 2345): They would agree on a payoff allocation  $c$  with  $\sum_i c_i = 1/4$ ,  $c_1 \geq 1/9$ , and  $c_2 + c_3 + c_4 + c_5 \geq 1/9$ . If at least two of the latter four summands are  $< 1/20$ , the corresponding players, say 45, have an incentive to leave since the unstable intermedi-

ate structure (1, 23, 45) would split further into either (1, 2, 3, 45) or (1, 23, 4, 5), and both players get  $1/20$  in each of them. Hence stability of (1, 2345) requires that three of the values  $c_2, c_3, c_4, c_5$  are  $\geq 1/20$ , so that  $c_1 \leq 1/4 - 3/20 = 1/10 < 1/9$  in contradiction to  $c_1 \geq 1/9$ . Thus (1, 2345) cannot stabilize itself either, and neither can any other structure with a four-player coalition.

Finally, we can now check whether the grand coalition can expect anyone to leave should they propose the allocation  $5 \cdot 1/20$ : If a player  $i$  leaves the room, he can expect that the other four players will split into two singletons and a pair that will first reach an intra-coalitional agreement and then meet again with the rest to negotiate an allocation of the additional surplus they can get from an inter-coalitional agreement. Because each other player announced to accept no less than  $1/20$  in that case,  $i$  cannot expect to get more than  $1/4 - 4/20 = 1/20$  when he leaves. Hence there is no incentive for individuals to leave the grand coalition in the first place when  $5 \cdot 1/20$  is proposed. With similar arguments, one can show that neither any coalition has an incentive to leave the grand coalition, and that the same also holds for larger values of  $n$  with the assumed cost-benefit functions. In other words, it seems likely that there will be an agreement in the grand coalition when inter-coalitional agreements are possible.

Now for a more general but *symmetric* case, where a similar analysis can be performed for most other cost-benefit structures. Assume benefits are symmetric,  $f_i = f_0$  for all  $i$ , and that for each  $m \in \{1, \dots, n\}$ , the equation

$$f'(Q) = (m-1)h(Q) + g'(Q) \quad [18]$$

has a unique solution  $Q_m$  with  $h'(Q_m) > 0$ . Then for each coalition structure  $\pi$  with  $|\pi| = m$  and each coalition  $S \in \pi$  with  $|S| = k$ , we have

$$v^{\text{PSE}}(S, \pi) = k f_0(Q_m) + h(Q_m) \frac{h(Q_m) - k f'_0(Q_m)}{h'(Q_m)}. \quad [19]$$

Now assume all players announce they will not accept a payoff less than  $v^{\text{PSE}}(N, \{N\})/n = P^*/n$ , no matter what structure arises.

Then each structure  $\pi$  can either stabilize itself by giving each player exactly  $P^*/n$ , or cannot stabilize itself at all. To see this, call this symmetric allocation  $a$ , and proceed inductively from finer to coarser structures: The all-singletons structure  $\pi$  is stable with  $a$  since it gives each coalition at least the same as in the PSE,  $P^*/n > v^{\text{PSE}}(\{i\}, \pi)$  for all  $i \in N$ , and no-one can leave any coalition since they are all singletons already. Given that the claim is true for all refinements of a structure  $\pi$ , we distinguish two cases to show that it is also true for  $\pi$ :

(i) If  $a$  gives each coalition  $S \in \pi$  at least  $v^{\text{PSE}}(S, \pi)$ , it is a possible outcome of an inter-coalitional agreement, and no player or subcoalition has an incentive to leave. The latter is because for every finer structure  $\pi'$  that might arise from leaving, they must expect that, because of the announcements,  $\pi'$  will stabilize itself by agreeing on the same allocation  $a$  if it can stabilize at all.

(ii) On the other hand, assume  $a$  gives some coalition  $S \in \pi$  less than  $v^{\text{PSE}}(S, \pi)$ , but some other allocation  $b$  stabilizes  $\pi$ . Then  $v^{\text{PSE}}(S, \pi) > k P^*/n$  where  $k = |S|$ , and  $b$  gives each coalition  $T \in \pi$  at least  $v^{\text{PSE}}(T, \pi)$ . Because  $S$  gets more under  $b$  than under  $a$ , some other coalition  $T \in \pi$  must get less under  $b$  than under  $a$ . The crucial point of the proof now is that this  $T$  cannot be a singleton; otherwise it would get under  $b$  at least

$$\begin{aligned} v^{\text{PSE}}(T, \pi) &= f_0(Q_m) + h(Q_m) \frac{h(Q_m) - f'_0(Q_m)}{h'(Q_m)} \\ &\geq \left( k f_0(Q_m) + h(Q_m) \frac{h(Q_m) - k f'_0(Q_m)}{h'(Q_m)} \right) / k \\ &= v^{\text{PSE}}(S, \pi) / k > P^*/n, \end{aligned} \quad [20]$$

but the latter is what a singleton gets under  $a$ . So  $T$  contains at least two players and gets less under  $b$  than under  $a$ . Hence at least one

player in  $T$  gets less under  $b$  than under  $a$ . That player has an incentive to leave  $T$  since she gets  $a$  in any stable structure that might arise from her leaving  $T$ . This proves that when  $a$  does not stabilize  $\pi$ , no other allocation  $b$  will. Finally, taking  $\pi = \{N\}$ , this proves that the grand coalition can stabilize by agreeing on the symmetric allocation  $a_i = P^*/n$ .<sup>5</sup>

So, in contrast to (35), the possibility of players or coalitions leaving negotiations need not destabilize the grand coalition if later inter-coalitional agreements are possible. We will further explore this line of thought in a forthcoming paper.

**The tracing procedure.** A quite different approach is that of Harsanyi and Selten (43) based on *payoff-dominance* and a so-called *tracing procedure*. It suggests that the grand coalition will indeed form to realize an optimal (i.e., payoff-undominated) equilibrium which is selected in a procedure in which all players gradually adapt their beliefs about the others' choices in a Bayesian fashion, depending not on a value function  $v$  but on the actual strategies that constitute the available equilibria. Unfortunately, that theory is mainly developed for games with bounded payoffs and only finitely many strategies, and therefore does not apply easily to our situation. We may however at least pick up the main idea of the tracing procedure (44) and interpret it in our context, making a number of assumptions on the beliefs of players during negotiations:

All players assess the progress of negotiations by the same parameter  $\tau \in [0, 1]$  that increases monotonically from zero at the beginning to one at the time agreement is reached. All players start at  $\tau = 0$  with the assumption that the remaining players will use their PSE strategies  $q_i^{\text{PSE}}$  as given by Eqns. 8 and 9. At each point  $\tau$  during negotiations, all players expect some allocation  $\vec{q}^\tau$  to be focal at this point and that all other players will apply the strategy LinC with targets  $\vec{q}^\tau$  if agreement will be reached, but expect that all other players will use their PSE strategies if no agreement will be reached. In particular,  $\vec{q}^0 = \vec{q}^{\text{PSE}}$ . At each point  $\tau$ , each player  $i$  considers the probability that agreement will be reached to be  $\tau$ . We now require that the focal allocation  $\vec{q}^\tau$  is rational for each player  $i$  if she maintains these beliefs. For this, playing LinC with targets  $\vec{q}^\tau$  must be a best response for  $i$  to the strategy mixture of the other players that her beliefs imply. For  $\tau = 1$ , all players will assume the rest will apply LinC with the agreed allocation with certainty, and our paper proves that for  $i$  it is a best response to that if she applies LinC with the same allocation. So, for  $\tau = 1$ , the rationality requirement does not restrict the set of possible allocations  $\vec{q}^\tau$ . But for  $\tau < 1$ , player  $i$  expects that there is a positive probability  $1 - \tau$  that the other players play their PSE strategies instead of LinC, in which case the best response would be to play  $q_i^{\text{PSE}}$  as well. The long-term payoff player  $i$  expects if she contributes  $q_i$  in each period is

$$(1 - \tau)W_1 P_i(q_i, q_i + Q_{-i}^{\text{PSE}}) + \tau V_i(q_i, \vec{q}^\tau) \quad [21]$$

where  $Q_{-i}^{\text{PSE}} = Q^{\text{PSE}} - q_i^{\text{PSE}}$ ,  $P_i(q_i, Q) = f_i(q_i) - q_i h(Q)$  is the period payoff of  $i$  if she contributes  $q_i$  and total contributions are  $Q$ , and  $V_i(q_i, \vec{q}^\tau)$  is the long-term payoff for player  $i$  if she contributes  $q_i$  in each period while all other players apply LinC with targets  $\vec{q}^\tau$ . Unfortunately, the Appendix on infinite sequences of deviations (case ii) shows that  $V_i(q_i, \vec{q}^\tau) = -\infty$  if  $q_i \neq q_i^\tau$ . This means that the best response would always consist in accepting  $q_i^\tau$  whatever it is. But then, also for  $\tau \in (0, 1)$ , the rationality requirement does not restrict the set of possible allocations  $\vec{q}^\tau$ , and the tracing procedure could not predict how the beliefs develop and whether the  $\vec{q}^\tau$  would converge.

Let us now assume that the cost-benefit structure is so that we could restrict liabilities to non-negative values and use BLinC instead of LinC to stabilize an agreement, as discussed in the SI: Bounded liabilities. If we replace LinC by BLinC in the above discussion, the value  $V_i(q_i, \vec{q}^\tau)$  will not be  $-\infty$  if  $q_i \neq q_i^\tau$ . Instead, the contributions by the other players will quickly converge to zero so the per-period payoff of  $i$  will converge to  $P_i(q_i, q_i)$ . If  $i$  is sufficiently patient,  $V_i(q_i, \vec{q}^\tau)$  will then approximately equal  $W_1(f_i(q_i) - g_i(q_i))$ .

The best response  $q_i$  to the current beliefs of  $i$  at  $\tau$  is then approximately that  $q_i$  which maximizes the function  $\pi_i^\tau(q_i, \vec{q}^\tau)$  that is given by  $\pi_i^\tau(q_i, \vec{q}^\tau) = a_i^\tau(\vec{q}^\tau)$  if  $q_i = q_i^\tau$  and by  $\pi_i^\tau(q_i, \vec{q}^\tau) = p_i^\tau(q_i)$  if  $q_i \neq q_i^\tau$ , where

$$a_i^\tau(\vec{q}^\tau) = (1 - \tau)P_i(q_i^\tau, q_i^\tau + Q_{-i}^{\text{PSE}}) + \tau P_i(q_i^\tau, Q^\tau), \quad [22]$$

$$p_i^\tau(q_i) = (1 - \tau)P_i(q_i, q_i + Q_{-i}^{\text{PSE}}) + \tau P_i(q_i, q_i). \quad [23]$$

The function  $p_i^\tau(q_i)$  has its maximum at that  $q_i$  for which

$$0 = (1 - \tau)[f_i'(q_i + Q_{-i}^{\text{PSE}}) - h(q_i + Q_{-i}^{\text{PSE}}) - q_i h'(q_i + Q_{-i}^{\text{PSE}})] + \tau[f_i'(q_i) - g'(q_i)]. \quad [24]$$

Denote this  $q_i$  by  $\tilde{q}_i^\tau$  and note that it depends on  $\tau$  but not on the focal allocation  $\vec{q}^\tau$ . As our rationality requirement maintains that the best response is  $q_i = q_i^\tau$ , the value  $a_i^\tau(\vec{q}^\tau)$  must be larger than  $p_i^\tau(\tilde{q}_i^\tau)$  for all  $i$ . Still, this does not much restrict the choice of  $\vec{q}^\tau$ .

So far, we only required *individual* rationality during negotiations. But it is natural also to assume a form of *collective* rationality and require that the focal allocation  $\vec{q}^\tau$  must be so that there is no alternative allocation that *payoff-dominates* it in the sense that the expected payoff  $a_i^\tau(\vec{q}^\tau)$  is larger for each  $i$ . This condition is equivalent to requiring that  $A^\tau(\vec{q}^\tau) = \sum_i a_i^\tau(\vec{q}^\tau)$  is maximal at  $\vec{q}^\tau$ , which requires that for all  $i$ ,

$$0 = \partial A^\tau(\vec{q}^\tau) / \partial q_i^\tau = (1 - \tau)[f_i'(q_i^\tau + Q_{-i}^{\text{PSE}}) - h(q_i^\tau + Q_{-i}^{\text{PSE}}) - q_i^\tau h'(q_i^\tau + Q_{-i}^{\text{PSE}})] + \tau[f'(Q^\tau) - g'(Q^\tau)]. \quad [25]$$

For  $\tau = 1$ , these equations are all equivalent to the optimality condition  $f'(Q^\tau) = g'(Q^\tau)$ . Although this implies that the final agreement realizes the optimum total contributions  $Q^\tau = Q^*$ , it does not pose any further restriction on  $\vec{q}^\tau$ . But for  $\tau < 1$ , these  $n$  equations might be independent and thus have a unique solution  $\vec{q}^\tau$ . If this is so for all  $\tau \in (\tau_0, 1)$  for any  $\tau_0 < 1$ , the tracing procedure maintains that in the last phase of the negotiations, the focal allocations will "trace" the path of those unique solutions  $\vec{q}^\tau$ , converging to some limit  $\vec{q}^1$  for  $\tau \rightarrow 1$ . This limit could now be considered a likely final outcome of the negotiations if suitable liability bounds can be found that allow the application of BLinC to actually realize it.

Let us look at the simple example of linear benefits  $f_i(Q) = \beta_i Q$  and quadratic costs  $g(Q) = \max\{0, Q\}^2$  again (SI: Examples). In that case, Eqn. [25] is

$$0 = (1 - \tau)[\beta_i - 2q_i^\tau - Q_{-i}^{\text{PSE}}] + \tau[\beta - 2Q^\tau]. \quad [26]$$

We can first determine  $Q^\tau$  from their sum, giving

$$Q^\tau = \frac{[\beta - (n - 1)Q^{\text{PSE}}](1 - \tau) + n\beta\tau}{2(1 - \tau) + 2n\tau} \quad [27]$$

which converges for  $\tau \rightarrow 1$  to  $Q^* = \beta/2$  as required. Then we get

$$q_i^\tau = \frac{\beta_i - Q_{-i}^{\text{PSE}}}{2} + \tau \frac{\beta - 2Q^\tau}{2(1 - \tau)} \quad [28]$$

which converges to

$$q_i^1 = \frac{\beta_i - Q_{-i}^{\text{PSE}}}{2} + \frac{n - 1}{2n} Q^{\text{PSE}} = \beta_i - \beta/2n. \quad [29]$$

The resulting payoffs are then all equal,  $P_i = \beta^2/4n$ , but this is a consequence of this particularly simple payoff structure.

If  $g(Q) = \max\{0, Q\}^\zeta$  with  $\zeta \neq 2$ , the resulting payoffs are larger for those with larger  $\beta_i$  if  $\zeta < 2$ , and they are larger for those with smaller  $\beta_i$  if  $\zeta > 2$ , opposite to how the PSE payoffs behave. An example of this is the emissions game with the STACO cost-benefit-model (1; 2; 8), using the same parameters as in SI: Bounded liabilities. It has approximately cubic costs ( $\zeta = 3$ ), and when we solve

<sup>5</sup>Note that this is also relevant for Cournot oligopolies of any size, since the Cournot-Nash equilibrium leads to the same  $v$  as the public good game with symmetric linear benefits and quadratic costs.

Eqns. 25 numerically, the resulting allocation of the optimal global payoff gives the US, Japan, and the EU (having large  $\beta_i$ ) a share of about 4%, 6%, and 4% of the payoff, respectively, and the remaining nine world regions (having small  $\beta_i$ ) a share of about 10% each. However, such an allocation could not be stabilized using BLinC with similar liability bounds as we discussed in SI: Bounded liabilities, so it does not seem a likely outcome of the emissions game.

## References.

1. Finus M, Ierland EV, & Dellink R (2006) Stability of climate coalitions in a cartel formation game. *Economics of Governance* 7:271–291.
2. Weikard HP, Dellink R, & van Ierland E (2010) Renegotiations in the greenhouse. *Environmental and Resource Economics* 45:573–596.
3. Intergovernmental Panel on Climate Change (2007) *IPCC Fourth Assessment Report: Climate Change 2007 (AR4)* (Cambridge University Press, Cambridge, UK).
4. U.S. National Academies of Sciences (2010) *Advancing the Science of Climate Change* (National Academies Press, Washington, D.C.).
5. U.S. National Research Council (2010) *Limiting the Magnitude of Climate Change* (National Academy Press, Washington, D.C.).
6. Fankhauser S (1995) *Valuing climate change: the economics of the greenhouse* (Earthscan Publications).
7. Tol RSJ (1997) *A decision-analytic treatise of the enhanced greenhouse effect* (Vrije Universiteit Amsterdam, Ph.D. thesis).
8. Rubio SJ & Ulph A (2006) Self-enforcing international environmental agreements revisited. *Oxford Econ Pap* 58:233–263.
9. Lenton TM, Held H, Kriegler E, Hall JW, Lucht W, Rahmstorf S, & Schellnhuber HJ (2008) Tipping elements in the Earth's climate system. *P Natl Acad Sci USA* 105:1786–93.
10. Schellnhuber HJ (2009) Tipping elements in the Earth System. *P Natl Acad Sci USA* 106:20561–3.
11. Enkvist PA, Naucler T, & Rosander J (2007) A cost curve for greenhouse gas reduction. *McKinsey Quarterly*, Feb 2007.
12. Carbone JC, Helm C, & Rutherford TF (2009) The case for international emission trade in the absence of cooperative climate policy. *J Environ Econ Manag* 58:266–280.
13. Kolstad CD (2007) Systematic uncertainty in self-enforcing international environmental agreements. *J Environ Econ Manag* 53:68–79.
14. Held H, Kriegler E, Lessmann K, & Edenhofer O (2009) Efficient climate policies under technology and climate uncertainty. *Energ Econ* 31:S50–S61.
15. Barrett S (1994) Self-enforcing international environmental agreements. *Oxford Econ Pap* 46:878–894.
16. Dasgupta P (2008) Discounting climate change. *J Risk Uncertainty* 37:141–169.
17. Traeger CP (2009) Recent developments in the intertemporal modeling of uncertainty. *Annu Rev Resour Econ* 1:261–285.
18. Keohane RO & Raustiala K (2008) Toward a post-Kyoto climate change architecture: a political analysis. *The Harvard Project on International Climate Agreements Discussion Papers* 08-01:1–32.
19. Bovensmann H *et al.* (2010) A remote sensing technique for global monitoring of power plant CO<sub>2</sub> emissions from space and related applications. *Atmos Meas Tech* 3:781–811.
20. Jones BD (2001) *Politics and the architecture of choice: Bounded rationality and governance* (University of Chicago Press, Chicago).
21. Carraro C & Marchiori C (2004) in *Endogenous strategic issue linkage in international negotiations*, eds Carraro C & Fragnelli V. (Edward Elgar), pp 65–84.
22. Axelrod R (2005) in *Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*, eds Judd KL & Tesfatsion L (North-Holland), pp 1567–1584.
23. Crandall JW & Goodrich MA (2005) Learning to compete, compromise, and cooperate in repeated general-sum games. *ACM International Conference Proceeding Series* 119:161–168.
24. Fowler JH & Christakis NA (2010) Cooperative behavior cascades in human social networks. *P Natl Acad Sci USA* 107:5334–8.
25. Milinski M, Semmann D, Krambeck HJ, & Marotzke J (2006) Stabilizing the Earth's climate is not a losing game: supporting evidence from public goods experiments. *P Natl Acad Sci USA* 103:3994–8.
26. Heitzig J & Simmons FW (2010) Some chance for consensus: voting methods for which consensus is an equilibrium. *Soc Choice Welf* (in press), DOI:10.1007/s00355-010-0517-y.
27. Abreu D (1988) On the theory of infinitely repeated games with discounting. *Econometrica* 56:383–396.
28. Ellerman AD & Decaux A (1998) Analysis of post-Kyoto CO<sub>2</sub> emissions trading using marginal abatement curves. *MIT Joint Program on the Science and Policy of Global Change* 40:1–33.
29. Nordhaus WD (1994) *Managing the Global Commons: The Economics of Climate Change*. (MIT Press, Cambridge, MA).
30. Aldy JE & Stavins RN, eds (2010) *Post-Kyoto International Climate Policy: Implementing Architectures for Agreement* (Cambridge University Press).
31. Frankel J (2010) in (30), pp 31–87.
32. Jacoby HD *et al.* (2010) in (30), pp 753–785.
33. Weikard HP, Finus M, Altamirano-Cabrera JC (2006) The impact of surplus sharing on the stability of international climate agreements. *Oxford Econ Pap* 58:209–232.
34. Ray D (2007) *A Game-Theoretic Perspective on Coalition Formation (The Lipsy Lectures)*. (Oxford University Press, USA, Oxford).
35. Ray D & Vohra R (1997) Equilibrium binding agreements. *J Econ Theory* 73:30–78.
36. Ray D & Vohra R (1999) A theory of endogenous coalition structures. *Game Econ Behav* 26:286–336.
37. Yi SS (1997) Stable coalition structures with externalities. *Game Econ Behav* 20:201–237.
38. Morasch K (2000) Strategic alliances as Stackelberg cartels-concept and equilibrium alliance structure. *Int J Ind Organ* 18:257–282.
39. Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44.
40. Shapley LS (1951) Notes on the n-person game II: the value of an n-person game. *Rand Corporation Research Memoranda* RM-670:1–19.
41. Kalai E & Samet D (1987) On weighted Shapley values. *Int J Game Theory* 16:205–222.
42. Monderer D, Samet D, & Shapley LS (1992) Weighted values and the core. *Int J Game Theory* 21:27–39.
43. Harsanyi JC & Selten R (1988) *A General Theory of Equilibrium Selection in Games*. (MIT Press Books, Cambridge, MA).
44. Harsanyi JC (1975) The tracing procedure: a Bayesian approach to defining a solution for n-person noncooperative games. *Int J Game Theory* 4:61–94.