Keeping climate in check: a self-enforcing strategy for cooperation in 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

Submitted to Proceedings of the National Academy of Sciences of the United States of America

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 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, whether optimal or sub-optimal, and we prove that it has several strong stability properties.

greenhouse gas emissions | free-riding | 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 avoid as compared to a reference scenario (e.g., a "business as usual" emissions path). Whereas the existing literature on this *emissions game* is mainly pessimistic about the likelihood of cooperation (2–11), our version of the game allows for much more positive results.

The general situation is modeled here as a *repeated game* played in a sequence of periods, with *continuous* control variables (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 a market for contributions that has perfect competition (12).

We propose that players adopt a simple dynamic *strategy* to choose their contributions. In each period, an initially negotiated *target* allocation of *liabilities* will be redistributed in reaction 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 an 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*.

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 (13; 14) and a Pareto-efficient and strong Nash-equilibrium in each subgame if all players use LinC. Our assumptions and the proposed strategy are summarized in Fig. 2.

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 concentrate on first negotiating an allocation of the *optimal* total payoff achievable and then implementing that allocation by using LinC. Regarding the emissions game, we will finally discuss why our results are in contrast to the frequent claim in the literature that self-enforcing agreements can achieve only sub-optimal targets, and will hint at a possible modification of the Kyoto mechanism that might enhance compliance levels.

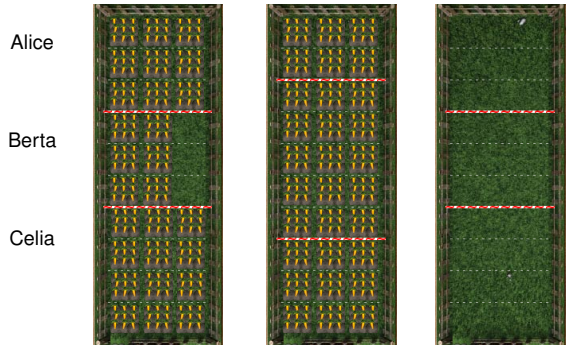

Fig. 1. Illustration of Linear Compensations in a simple public good game. Alice, Berta, and Celia farm their back-yard for carrots. Each has her individual farming liability (red-white separators) but harvests are divided equally. In the first year, Berta falls short of her target by three square meters.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

Reserved for Publication Footnotes

Framework

The public good game. Assume that there are infinitely many periods, numbered $1, 2, \ldots$, and finitely many players, numbered $1, \ldots, n$. 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 $Q(t) = \sum_i q_i(t)$. Throughout this article, individual quantities are denoted by small letters, and totals by large letters.

In the emissions game, $q_i(t)$ would be the difference between i's hypothetical amount of GHG emissions in period t in some predetermined 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 Appendix.

The 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.[1] Unlike in some other models of public goods, we assume here that total costs are shared in proportion to individual contributions, e.g., because there is a market for contributions that has perfect competition or because marginal costs are constant. Hence

$$c_i(t) = C(t) \cdot \frac{q_i(t)}{Q(t)} = g(Q(t)) \cdot \frac{q_i(t)}{Q(t)} \qquad [1]$$

if $Q(t) > 0$, and $c_i(t) = 0$ otherwise.

In the emissions game, the benefits of reducing emissions by 1 Gton CO_2-equivalents in period t correspond to all avoided welfare losses that would have been caused at times after t by that additional 1 Gton 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. 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 (EUETS), to which all players have equal access. A simple example of such a cost-benefit structure is that of linear benefits and marginal costs (4): $f_i(Q) = \beta_i Q$ with $\beta_i > 0$, $g(Q) = Q^2$ for $Q > 0$, and $g(Q) = 0$ for $Q \leqslant 0$. 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 $P(t) \to -\infty$ for $Q \to \pm\infty$. In the emissions game, large positive or negative values for some $q_i(t)$ can obtain if large amounts of emissions permits are traded. Although the strategy we will propose below prescribes such large values of $q_i(t)$ only in cases where there has already been an irrationally large earlier deviation, this might still lead to problems in practice. Therefore we also analyse in the Supporting Information the alternative case in which contributions and liabilities are bounded.

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 δ payoff units in t.

For some known (or estimated) *optimal total contributions* $Q^\star$, total payoff is maximal, and marginal total costs equal marginal total benefits but exceed marginal individual benefits:

$$F(Q^\star) - g(Q^\star) = \max, \quad g'(Q^\star) = F'(Q^\star) > f_i'(Q^\star). \qquad [2]$$

Optimal total payoffs are usually much larger than the total payoffs the players would end up if they do not cooperate. In the simple example with linear benefits and marginal costs, e.g., optimal total payoffs are larger than the non-cooperative equilibrium payoffs by a factor of $(n + 1)^2/4$, showing that the potential gains of cooperation can be 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* $q_i^\star$, with $\sum_i q_i^\star = Q^\star$. 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 $Q_G^\star = \sum_{i \in G} q_i^\star$.[2] However, the allocation need not be profitable for each player as compared to the reference scenario, i.e., some target payoffs may be negative.

In the emissions game, the targets could for example be negotiated using equity criteria such as per capita emissions permits, per capita payoffs, historical responsibility, etc. (6; 15; 16). In game-theoretic terms, this initial negotiation poses a problem of *equilibrium selection* that might be solved by *coalition formation* and pre-

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 in proportion to individual contributions
- 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

Fig. 2. Main assumptions and solution for the public good game

[1] Formally, f_i and g are twice differentiable, $b_i(t) = f_i(Q(t))$, $C(t) = g(Q(t)) \geqslant 0$, $f_i(0) = g(0) = 0$, $f_i'(Q) > 0$, $g'(Q) \geqslant 0$, $f_i''(Q) \leqslant 0$, and $g''(Q) \geqslant 0$.

[2] Formally: $\sum_{i \in G} f_i'(Q^\star) < h'(0)$ where $h(x) = (Q_G^\star + x) g(Q^\star + x)/(Q^\star + x)$.

cedes the problem of *cooperation* which we are concerned with in this article (see also SI: Target allocation).

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. 2). 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. When reacting suitably to free-riding, its immediate gains might be compensated by later losses. 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 (17)).

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 (13).

A famous example of such a non-credible strategy, though in a different game, is the strategy "tit for tat" that can be observed in various versions of the repeated Prisoners' Dilemma in which players *can* commit themselves beforehand (18; 19). In that game, each of two players decides to either "cooperate" or "defect" in each period, and the strategy is to start with "cooperate" and then do whatever the other player did in the previous period, thereby *punishing* non-cooperation by non-cooperation. 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 a defection is void and cannot deter free-riding under assumptions of rationality and without commitment possibilities (20).

Another problematic strategy is to simply treat free-riding as some form of debt to be *repayed* with interest, as it is done, e.g., in the "Procedures and mechanisms relating to compliance under the Kyoto protocol" that were adopted in 2001 in the so-called "Marrakech Accord". According to its Article XV 5 (a), a country free-riding 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 contributions in $t+1$ that exceed the optimal value $Q^\star$. Hence if renegotiations are possible after a shortfall, all players would agree to rather jointly contribute the smaller but optimal value $Q^\star$ in $t+1$, allocating individual contributions in a way so that the additional payoff is positive for each country. Even worse, if a player never fulfils 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 own shortfalls is both efficient and extremely stable, even when players make small errors in implementing it. Fig. 2 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, $Q(t) = Q^\star$. 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 at no possible situation in the game, all players would agree to change the strategy. In game-theoretic terms, such a strategy is *Pareto-efficient in all subgames.* It will thus be *strongly renegotiation-proof* (13; 14) 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.[3]

Deterring simple free-riding by groups of players. In this section, we will need many of the denotations we introduced earlier and which are summarized in Fig. 3.

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: $Q_G(t) = Q^\star_G - x$. Note that G's benefits are given by $f_G(Q) = \sum_{i\in G} f_i(Q)$, so that $\beta_G = f'_G(Q^\star)$ is G's target marginal benefit. Let $\gamma = g(Q^\star)/Q^\star$ be the average unit costs at the target contributions. Then the free-riding reduces G's 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

$$x(\gamma - \beta_G). \qquad [3]$$

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 δ are no smaller than the right-hand side of Eqn. 3. 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 let G's joint contributions $Q_G(t+1)$ be at least $Q^\star_G + x(\gamma-\beta_G)/\gamma\delta$ and reduce the other players' contributions

α	compensation factor
$B(t), b_i(t)$	benefits in period t, total and for player i
β_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
δ	lower bound for discounting factors
$F(Q), f_i(Q)$	benefits of total contributions Q, total and for player i
$g(Q)$	total costs of total contributions Q
γ	average unit costs at target
$\ell_i(t)$	liability of player i in period t
$Q(t), q_i(t)$	contributions in period t, total and by player i
$Q^\star, q_i^\star$	target contributions, total and for player i
x	size of potential shortfall by a group of players G

Fig. 3. Main symbols used in this article

[3] If we drop the assumption that the global target $Q^\star$ 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* (13; 14). See also the Appendix.

[4] If *all* players are in G, optimality of $Q^\star$ implies that shortfalls give no gains for G in period t.

accordingly.[4] This leads to additional costs for G of size

$$[Q^\star_G + x(\gamma - \beta_G)/\gamma\delta]\gamma - Q^\star_G\gamma = x(\gamma - \beta_G)/\delta. \qquad [4]$$

So, G's joint gains in t are overcompensated by these losses in $t+1$. Although the free-riding for one period might be profitable for some individual members of G, there is always at least one member of G for which 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, $\ell_i(1) = q^\star_i(1)$. 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:

new liability* = *target* + [*own shortfall* − *mean shortfall*] · *factor

$$\ell_i(t+1) = q^\star_i + [d_i(t) - \bar{d}(t)] \cdot \alpha. \qquad [5]$$

In this, $\bar{d}(t) = \sum_i d_i(t)/n$ is the mean shortfall and α 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 α has to be large enough for the argument of Eqns. 3 and 4 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/δ, 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

$$\alpha > \frac{n}{\gamma\delta} \cdot \max_G \frac{\gamma - \beta_G}{n - |G|}, \qquad [6]$$

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 $\beta_G = C'(Q^\star)|G|/n \geqslant \gamma|G|/n$ and Eqn. 6 simplifies to $\alpha > [n\gamma - C'(Q^\star)]/\gamma\delta(n-1)$, so that in particular $\alpha > 1/\delta$ suffices. Note that liabilities do not depend on costs and benefits explicitly, only via the negotiated targets $q^\star_i$ and the factor α, 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. Let us call the strategy defined by Eqns. 5 and 6 *Linear Compensation (LinC)*.

In game-theoretic terms, we have shown above that when all players comply with 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 argument (21): Assume $m > 0$ is the smallest integer for which a sequence of m successive deviations exists that are profitable for some group G. Let $t, \ldots, t+m-1$ be the periods in which they deviate, and $t+m$ the period in which they return to compliance with LinC. But we already proved above that in period $t+m-1$, it is not profitable to deviate for one period and then return to LinC. Hence it must be even more profitable to only do the first $m-1$ deviations and then return to LinC already in period $t+m-1$. So there is a sequence of $m-1$ successive deviations that are profitable, in contradiction to our choice of m. 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 $q^\star_i$, it builds a *strong Nash equilibrium in each subgame* if all players comply with LinC given these targets. Roughly speaking, the reason is that if G continually shortfalls, liabilities and contributions of the other players will decrease fast enough so that G's gains from saved costs are overcompensated by its losses from decreased total contributions in the long-term. Note that the others do not need to use a threat of contributing nothing forever, which would be incredible, but only the threat of answering a period of shortfalls by 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 shortfalling again, no matter how many shortfalls have happened already.

Discussion

We have presented here a simple strategy by which players in a public good game can try to keep each other in check in the provision of agreed target contributions. Our approach can be interpreted as a combination of, on the one hand, a proportionate version of the punishment approach that strategies like "tit for tat" use in the Prisoners' Dilemma, and, on the other hand, the repayment approach that is already included in the Kyoto mechanism. Unlike each one of these ingredients alone, this combination has then been formally shown here to have strong game-theoretic stability properties if the situation fulfils some simplifying assumptions. In Axelrod's (18) terminology, our strategy, LinC, is *nice* in that it cooperates unless provoked, *retaliating* when provoked, *forgiving* when deviators repay, and uses *contrition* 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 (see the discussion in the Appendix).

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. If their errors can be modelled as being random with a given variance, it is easy to see from Eqn. 5 that errors do not add up or lead away from the target, nor do one-time deviations initiate an infinite "echoing" sequence of reactions as strategies like "tit for tat" would.[5] 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 "contrite tit-for-tat" for the repeated Prisoners' Dilemma (22).

All the above stability properties of LinC hold independently of the value of the discounting factor δ if only the compensation factor α is chosen accordingly.[6] This is in contrast to n-player versions of the repeated Prisoners' Dilemma, in which such stable strategies only exist when players are sufficiently patient, i.e., when δ is close to unity (20). The reason is that when the only possible two actions are to "cooperate" or to "defect", no redistribution that keeps the optimal total target is possible. Strategies must then use sophisticated reciprocation to construct credible threats not subject to renegotiation.

It may come as a surprise that while in many other games there are no strong Nash equilibria at all, the public good game studied here even allows players to sustain *any* conceivable allocation of the optimal total payoff with such an equilibrium. This means that although

[5] With implementation errors of variance σ^2, the mean squared deviation of $\ell_i(t+1)$ from the target $q^\star_i$ will be at most $\sigma^2\alpha^2(n-1)/n$, hence the mean squared deviation between actual and target contributions is of magnitude $\sigma^2(1+\alpha^2(n-1)/n)$.

[6] The value of δ however does play a role when, in addition to our assumptions, liabilities shall be bounded. This is further explored in the Appendix.

the *cooperation* problem can be solved in this game, the *equilibrium selection* problem might be even harder than in games with no or only a small number of strong equilibria. In SI: Target allocation, we give some hints on possible game-theoretic approaches. In the emissions game, our scheme can be interpreted as a form of "cap and trade" with a dynamic allocation of permits that does not seem to require negative initial targets ("hot air") for some countries to secure compliance (23).

Comparison with the literature on the emissions game. Some of the existing literature on the emissions game (8; 9) treat it as a form of Prisoners' Dilemma with a choice to either "cooperate" or "defect", and the above discussion explains why those studies are much more pessimistic about cooperation. Although there are also positive contributions like (24), they do not consider deviations by groups of players and cannot be applied to a model with more than two choices (25). We believe that a model in which countries choose an actual level of emissions is more appropriate than one in which they can only choose between two pre-determined emission levels.

In (3; 26), several dynamic physico-economic models of the emissions game are described which are much more sophisticated than our simplistic model and differ in features that are important for a strategic analysis and give it a different strategy set than in our game. A slightly similar approach as ours (27) uses harsh punishment strategies that last several periods to arrive at a subgame-perfect equilibrium. As those authors do not discuss renegotiations or deviations by groups of players, it is not straightforward to compare their partly negative and partly positive results to our optimistic result, which will be an important task for future research.

Other authors neglect the time dimension (2; 4–6) and assume countries choose only once, whether to join a long-term coalition or not. Usually only one coalition is assumed to form which then tries to maximize their joint payoff, while each non-member tries to maximize its individual payoffs, leading to sub-optimal total payoffs. As these studies have to assume that players can sign legally binding and enforceable long-term agreements, it remains unclear why the predicted coalition would not later on find another such agreement with all non-members to realize the optimal total payoff and share the resulting surplus in some way. The present paper shows that such an agreement would indeed be stable if players agree to use LinC for its implementation, which we further discuss in SI: Coalition formation.

In one paper (7), a complicated iterated game is used as a model that has, however, only finitely many periods, so that the game can be "solved backwards". Unfortunately, with a finite time horizon, one cannot react to free-riding in the last period, and this has also negative effects in earlier periods.

As mentioned, the Kyoto mechanism already includes a form of compensation rule, and our results indicate that it should be analysed whether compliance can be expected to improve if the compensation is modified to keep total liabilities constant as in Eqn. 5 and if the current compensation factor of 1.3 is adjusted according to Eqn. 6.

Appendix: Validity of assumptions on the emissions game

Typical models of the emissions game (4; 7) fulfil our assumptions on costs and benefits.

Concerning *benefits,* the economic literature on climate change as distilled in the IPCC's 4th Assessment Report indicates that the global society as a whole would benefit from reduced GHG concentrations. The regional distribution of the consequences of climate change is much more uncertain, but some studies (28; 29) 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 (4; 5; 7). Otherwise, i has to be excluded from our analysis and its contributions (if any) could be treated as an exogenous variable for our solution to be applicable.

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 (30; 31), our analysis would still be valid if we let β_G denote the value $\inf_{Q \leqslant f_G'(Q^\star)} f_G'(Q)$ instead of $f_G'(Q^\star)$ and raise the compensation factor α 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 (32) 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 proportion of contributions, and whether this is justified depends on whether the market has perfect competition or prices can be influenced strategically by countries as it is assumed in some models (33). With a suitable choice of α, LinC should work also for other cost sharing schemes as long as individual costs g_i grow in a convex way when contributions are redistributed. Future research should investigate this issue in detail.

To be able to apply the powerful analysis tools developed for *repeated games,* we had to assume time-independent cost and benefit functions, although in a more accurate model the functions f_i and g, and hence the quantity $Q^\star$ would display some time-dependency, e.g., because of technological progress or the stock pollutant nature of GHG, making this an *iterated* or even *dynamic* instead of a repeated game. However, most of our analysis depends only on a comparison of two successive periods, and if the period length is not too large, f_i, g, and $Q^\star$ will change only little from period t to period $t+1$. Future work should check whether a suitable increase in the compensation factor α might suffice to account for this variation.

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 (34; 35), e.g., due to the unknown value of future GDP and the fact that for a stock pullutant, 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 (2; 4; 5; 7). On the question of discounting in the context of the emissions game, see (36), 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 (23), 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. 5. 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 (37), 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 (38), and approaches based on agent-based model-

ing (39), learning theory (40), or complex networks (41) 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 (42).

Appendix: Why infinite sequences of deviations do not pay

Suppose all players comply with 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 $t \geqslant t_0$. Since excess contributions never pay, we can assume that $x_t \geqslant 0$.

Assume further that in each period t and for each integer $r \geqslant 0$, 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* $w_{t,r}$ fulfil the conditions

$$w_{t,0} = 1,\ \ w_{t,1} > \delta,\ \ w_{t,r} \geqslant 0,\ \ \sum_{r=0}^{\infty} w_{t,r} = W_t < \infty. \quad \textbf{[7]}$$

E.g., players could use exponential discounting with $w_{t,r} = \varepsilon^r$, $\delta < \varepsilon < 1$, and $W_t = 1/(1-\varepsilon)$.[7] G's *discounted long-term payoff* from t_0 on is then $U_G(t_0) = \sum_{t\geqslant t_0} w_{t_0,t-t_0} P_G(t)$ 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 comply with LinC instead. Assume $\Delta(s, \mathrm{LinC}) > 0$ 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 $\sum_{t\geqslant t_0} w_{t_0,t-t_0} x_t$ 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 $t_0 \leqslant t < t_1$ and playing LinC for $t \geqslant t_1$. 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 $t \geqslant t_1$ when playing s instead of LinC, which is at most $x_t\gamma$ according to Eqn. 3. Hence $\Delta(s, \tilde{s}) \leqslant \sum_{t\geqslant t_1} w_{t_0,t-t_0} x_t\gamma$. Because of the assumed series convergence, this goes to zero for $t_1 \to \infty$, so it is smaller than $\Delta(s, \mathrm{LinC})$ if t_1 is large enough. Then $\Delta(\tilde{s}, \mathrm{LinC}) = \Delta(s, \mathrm{LinC}) - \Delta(s, \tilde{s}) > 0$ which means that already the truncated deviation strategy $\tilde{s}$ is profitable. But we already proved that no finite sequence of deviations is profitable, a contradiction.

(ii) Suppose the discounted total long-term shortfalls are infinite, $\sum_{t\geqslant t_0} w_{t_0,t-t_0} x_t = \infty$. Because $x_{t-1} \geqslant 0$, the joint liability of G in period t is no smaller than the target, $L_G(t) = \sum_{i\in G} \ell_i(t) \geqslant Q^\star_G$. Their joint costs are

$$C_G(t) = (L_G(t) - x_t)\frac{g(Q^\star - x_t)}{Q^\star - x_t}. \quad \textbf{[8]}$$

If $x_t \geqslant Q^\star$, these are zero because total costs are. For $x_t < Q^\star$, we have $C_G(t) \geqslant (L_G(t) - Q^\star)g(Q^\star - x_t)/(Q^\star - x_t)$. The latter is non-negative for $L_G(t) \geqslant Q^\star$ and is otherwise at least $(Q^\star_G - Q^\star)\gamma \leqslant 0$ since average costs are non-decreasing. So in all cases, $C_G(t) \geqslant (Q^\star_G - Q^\star)\gamma$. Concerning benefits, let $f_G(Q) = \sum_{i\in G} f_i(Q)$ and let $\beta_G = f'_G(Q^\star)$ be the target marginal benefit of G. Then G's joint benefits are $f_G(Q^\star - x_t)$, which is at most $f_G(Q^\star) - \beta_G x_t$ because marginal benefits are non-increasing. Thus G's joint payoffs are at most $(Q^\star - Q^\star_G)\gamma + f_G(Q^\star) - \beta_G x_t$, so that G's discounted long-term payoff $U_G(t_0)$ is then at most

$$W_{t_0}[(Q^\star - Q^\star_G)\gamma + f_G(Q^\star)] - \beta_G \textstyle\sum_{t\geqslant t_0} w_{t_0,t-t_0} x_t.$$

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.

Cases (i) and (ii) exhaust the possibilities, hence no strategy of deviations whether of finite or infinite length can increase G's discounted long-term joint payoff. Hence LinC builds a strong Nash-equilibrium in each subgame.[8]

ACKNOWLEDGMENTS. We thank Ottmar Edenhofer, Gunnar Luderer, Robert Marschinski, John Schellnhuber, Arghya Mondal, and Bo Hu for fruitful discussions. 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).

References.

1. Dinar, A, Albiac, J, & Sánchez-Soriano, J. (2008) *Game Theory and Policy Making in Natural Resources and the Environment (Routledge Explorations in Environmental Economics)*. (Routledge, London), pp. 1–368.
2. Barrett, S. (1994) *Oxford Economic Papers*.
3. Dutta, P. K & Radner, R. (2006) *Advances in Mathematical Economics* **8**, 135–153.
4. Finus, M, Ierland, E. V, & Dellink, R. (2006) *Economics of Governance* **7**, 271–291.
5. Rubio, S. J & Ulph, A. (2006) *Oxford Economic Papers* **58**, 233–263.
6. Weikard, H.-P. (2006) *Oxford Economic Papers* **58**, 209–232.
7. Weikard, H.-P, Dellink, R, & van Ierland, E. (2010) *Environmental and Resource Economics* **45**, 573–596.
8. Barrett, S. (1998) *SSRN Electronic Journal*.
9. Asheim, G. B, Betteville Froyn, C, Hovi, J, & Menz, F. C. (2006) *Journal of Environmental Economics and Management* **51**, 93–109.
10. Finus, M. (2008) *International Review of Environmental and Resource Economics* **2**, 29–67.
11. Nagashima, M, Dellink, R, van Ierland, E, & Weikard, H.-P. (2009) *Ecological Economics* **68**, 1476–1487.
12. Ellerman, A. D & Decaux, A. (1998) *MIT Joint Program on the Science and Policy of Global Change* **40**, 1–33.
13. Farrell, J & Maskin, E. (1989) *Games and Economic Behavior* **1**, 327–360.
14. Bergin, J & MacLeod, B. (1993) *Journal of Economic Theory* **61**, 42–73.
15. Kverndokk, S & Rose, A. (2008) *Fondazione Eni Enrico Mattei Working Papers* **239**, 1–66.
16. 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).
17. Aumann, R. J. (2006) *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17075–8.
18. Axelrod, R. (1981) *The American Political Science Review* **75**, 306 – 318.
19. Axelrod, R & Keohane, R. O. (1985) *World Politics* **38**, 226–254.
20. van Damme, E. (1989) *Journal of Economic Theory* **47**, 206–217.
21. Abreu, D. (1988) *Econometrica* **56**, 383–396.
22. Sugden, R. (1986) *The evolution of rights, co-operation and welfare*. (Blackwell, Oxford).
23. Keohane, R. O & Raustiala, K. (2008) *The Harvard Project on International Climate Agreements Discussion Papers* **08-01**, 1–32.
24. Bretteville Froyn, C & Hovi, J. (2008) *Economics Letters* **99**, 317–319.
25. Asheim, G. B & Holtsmark, B. (2009) *Environmental and Resource Economics* **43**, 519–533.
26. Dutta, P. K & Radner, R. (2009) *Journal of Economic Behavior & Organization* **71**, 187–209.
27. Mason, C. F, Polasky, S, & Tarui, N. (2009) *Working paper*.
28. Fankhauser, S. (1995) *Valuing climate change: the economics of the greenhouse*. (Earthscan Publications).
29. Tol, R. S. J. (1997) Ph.D. thesis (Vrije Universiteit Amsterdam).
30. Lenton, T. M, Held, H, Kriegler, E, Hall, J. W, Lucht, W, Rahmstorf, S, & Schellnhuber, H. J. (2008) *Proceedings of the National Academy of Sciences of the United States of America*.
31. Schellnhuber, H. J. (2009) *Proceedings of the National Academy of Sciences of the United States of America* **106**, 20561.
32. Enkvist, P.-A, Naucler, T, & Rosander, J. (2007) *McKinsey Quarterly* pp. 1–17.
33. Carbone, J. C, Helm, C, & Rutherford, T. F. (2009) *Journal of Environmental Economics and Management* **58**, 266–280.
34. Kolstad, C. D. (2007) *Journal of Environmental Economics and Management* **53**, 68–79.
35. Held, H, Kriegler, E, Lessmann, K, & Edenhofer, O. (2009) *Energy Economics* **31**, S50–S61.
36. Dasgupta, P. (2008) *Journal of Risk and Uncertainty*.
37. Jones, B. D. (2001) *Politics and the architecture of choice: Bounded rationality and governance*. (University of Chicago Press, Chicago).
38. Carraro, C & Marchiori, C. (2004) *Endogenous strategic issue linkage in international negotiations*, eds. Carraro, C & Fragnelli, V. (Edward Elgar), pp. 65–84.
39. Axelrod, R. (2005) Agent-based modeling as a bridge between disciplines.
40. Crandall, J. W & Goodrich, M. A. (2005) *ACM International Conference Proceeding Series; Vol. 119* pp. 161–168.
41. Fowler, J. H & Christakis, N. A. (2010) *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5334–8.
42. Milinski, M, Semmann, D, Krambeck, H.-J, & Marotzke, J. (2006) *Proceedings of the National Academy of Sciences of the United States of America* **103**, 3994–8.
43. Jamison, D. T & Jamison, J. (2003) *Congress of the European Economic Association, Stockholm*.
44. Lehrer, E & Pauzner, A. (1999) *Econometrica* **67**, 393–412.

[7] Another example is *hyperbolic discounting* with $w_{t,r} = (1 + \alpha_t r)^{-1-\zeta_t}$ and certain parameters $\alpha_t, \zeta_t > 0$ (43). For the emissions game, see the discussion in (36). If individual players discount differently, one says they have different *time-preferences*, the analysis gets more complicated because intertemporal trade can be profitable (44), and our results concerning renegotiations might no longer hold. If efficient financial markets exist, they can be expected to equalize discount rates (44) so that our assumption would be valid in the emissions game.

[8] If 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.

Supporting Information for Keeping climate in check: a self-enforcing strategy for cooperation in 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

Supporting Information: 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 $P^\star$ for $Q = Q^\star$, a vector of individual contributions q_i is Pareto-efficient if and only if $\sum_i q_i = Q^\star$.

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, $Q \geqslant 0$, since for $Q < 0$ we have $\partial P_i/\partial q_i \geqslant f_i'(0) > 0$. Hence $q_i \geqslant -Q_{-i}$. Denote average unit costs by $h(Q) = g(Q)/Q \geqslant 0$, so that $h'(Q) = (g'(Q) - h(Q))/Q \geqslant 0$ for $Q > 0$ and $h'(Q) = 0$ for $Q \leqslant 0$. Note that for $q_i = -Q_{-i}$ we have $P_i = 0$, and for $q_i \to +\infty$ we have $P_i \to -\infty$. 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 $\partial P_i/\partial q_i = f_i'(Q) - h(Q) - q_i h'(Q) = 0$, 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.

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 case (i), 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 $F'(Q^\star) = g'(Q^\star)$, this condition has at least one (and often unique) solution $Q^{\mathrm{PSE}} > 0$:

$$F'(Q^{\mathrm{PSE}}) = (n-1)h(Q^{\mathrm{PSE}}) + g'(Q^{\mathrm{PSE}}). \qquad [9]$$

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

$$q_i^{\mathrm{PSE}} = \frac{f_i'(Q^{\mathrm{PSE}}) - h(Q^{\mathrm{PSE}})}{h'(Q^{\mathrm{PSE}})} \qquad [10]$$

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

$$P_i^{\mathrm{PSE}} = f_i(Q^{\mathrm{PSE}}) + h(Q^{\mathrm{PSE}})\frac{h(Q^{\mathrm{PSE}}) - f_i'(Q^{\mathrm{PSE}})}{h'(Q^{\mathrm{PSE}})}. \qquad [11]$$

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

Because $P^{\mathrm{PSE}} < P^\star$, there are allocations q_i of the total optimal contributions $Q^\star$ that give each i a strictly higher payoff than P_i^{PSE}. Hence each such PSE is Pareto-dominated but may serve as a kind of benchmark in negotiations of the target allocation $q_i^\star$ in the sense that one could only allow for target allocations that Pareto-dominate the PSE. See also SI: Target allocation.

Supporting Information: Renegotiations when targets are not optimal

Let us drop the assumption that the global target $Q^\star$ 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 α 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 $q_i^\star$ 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 $\sum_i U_i(t+1) = \sum_{r\geqslant 0} w_{t+1,r} P(t+1+r)$. Thus the supremum of the new total period payoffs, $P^+ = \sup_{r\geqslant 0} P(t+1+r)$, 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 $q_i^\star$, 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 λ. Now let s_0 be the strategy of applying LinC with the original targets $q_i^\star$ and consider these four cases:

(i) all play s_0 from t on,
(ii) G free-rides in t and all continue 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 λ 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 can 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.

Supporting Information: 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 $g'(Q) = \zeta \max\{Q, 0\}^{\zeta-1}$.
- 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 $Q < 0$.
- Total period payoff $P(Q) = \beta Q - \max\{Q, 0\}^\zeta$ with $P'(Q) = \beta - \zeta \max\{Q, 0\}^{\zeta-1}$.
- Individual period payoff $P_i = \beta_i Q - q_i h(Q)$ 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 $Q < 0$.

The optimal total contributions $Q^\star > 0$ then fulfil

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

hence

$$\begin{aligned} Q^\star &= \gamma^{\frac{1}{\zeta-1}}, \\ P^\star &= (\zeta - 1)\gamma^{\frac{\zeta}{\zeta-1}}, \end{aligned}$$

where $\gamma = \beta/\zeta$ are the target average unit costs. Similarly, a one-shot PSE has $Q^{\mathrm{PSE}} > 0$ and thus fulfils

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

hence the uniquely PSE is given by

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

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 $P_i^{\mathrm{PSE}} \propto 1/(n+1)^2$, bearing a surprising similarity to Cournot-Nash payoffs in Cournot oligopolies (see also SI: Target allocation). 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.

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 $Q \geqslant 0$ and $f_i(Q) = \beta_i(Q - Q^2/2 + Q^3/3)$ for $Q \leqslant 0$, with $\beta_i > 0$.
- Individual marginal benefits $f_i'(Q) = \beta_i/(1 + Q)$ for $Q \geqslant 0$ and $f_i'(Q) = \beta_i(1 - Q + Q^2)$ for $Q \leqslant 0$.
- Total costs $g(Q) = \max\{Q, 0\}^2$.
- Marginal total costs $g'(Q) = \max\{2Q, 0\}$.
- Average unit costs $h(Q) = \max\{Q, 0\}$ with $h'(Q) = 1$ for $Q > 0$ and $h'(Q) = 0$ for $Q < 0$.
- Total period payoff for $Q \geqslant 0$: $P(Q) = \beta \ln(1 + Q) - Q^2$ with $P'(Q) = \beta/(1 + Q) - 2Q$.
- Individual period payoff for $Q \geqslant 0$: $P_i = \beta_i \ln(1+Q) - q_i Q$ with $\partial P_i/\partial q_i = \beta_i/(1+Q) - q_i - Q$.

Optimal total contributions $Q^\star > 0$ must fulfil

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

hence

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

Similarly, a one-shot PSE has $Q^{\mathrm{PSE}} > 0$ and thus fulfils

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

hence the unique PSE is given by

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

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 β.

For small β and large n, $P^\star \approx \beta^2/4$ and $P^{\mathrm{PSE}} \approx \beta^2/n = O(P^\star/n)$, i.e., the cooperative payoff is of the order n larger than the PSE payoff. For large β, $P^\star \approx \beta(\ln \beta - \ln 2)/2$ and $P^{\mathrm{PSE}} \approx \beta(\ln \beta - \ln(n+1))/2 = O(P^\star)$, i.e., the ratio between cooperative and PSE payoffs does not diverge for large n.

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 $Q \in [0, 1)$.
- Average unit costs $h(Q) = Q/(1 - Q)$ for $Q \in [0, 1)$, with $h'(Q) = 1/(1 - Q)^2$
- Total period payoff $P(Q) = \beta Q - Q^2/(1 - Q)$ for $Q \in [0, 1)$, with $P'(Q) = \beta - Q(2 - Q)/(1 - Q)^2$,
- Undividual period payoff $P_i = \beta_i Q - q_i Q/(1 - Q)$ for $Q \in [0, 1)$, with $\partial P_i/\partial q_i = \beta_i - Q/(1-Q) - q_i/(1-Q)^2$.

Optimal total contributions $Q^\star \in (0, 1)$ fulfil

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

hence

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

Similarly, a one-shot PSE has $Q^{\mathrm{PSE}} \in (0, 1)$ and thus fulfils

$$\begin{aligned} 0 &= F'(Q^{\mathrm{PSE}}) - (n-1)h(Q^{\mathrm{PSE}}) - g'(Q^{\mathrm{PSE}}) \\ &= \beta - Q^\star(n + 1 - NQ^\star)/(1 - Q^\star)^2, \end{aligned}$$

hence the unique PSE is given by

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

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

Supporting Information: 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 $\ell_i^{\min} < q_i^\star$ for all players i, so that only liabilities with $\ell_i(t) \geqslant \ell_i^{\min}$ 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 $Q^\star$ in order to be strongly renegotiation-proof can then assign any group G of players at most the liability $L_G^{\max} = Q^\star - \sum_{i \notin G} \ell_i^{\min}$.

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:

$$\ell_i(t+1) = \begin{cases} \max\{\, q_i^\star + [d_i(t) - \bar{d}(t)] \cdot \alpha \,, \ell_i^{\min}\} & \text{if } d_i(t) = 0 \\ q_i^\star + [d_i(t) - \bar{d}(t)] \cdot \alpha \,/ s(t) & \text{if } d_i(t) > 0, \end{cases} \quad [12]$$

where $s(t) > 0$ is chosen so that $\sum_i \ell_i(t+1) = Q^\star$. If shortfalls are moderate so that $\bar{d}(t) \leqslant (q_i^\star - \ell_i^{\min})/\alpha$ for all i with $d_i(t) = 0$, then $s(t) = 1$ and the allocation is the same as in LinC (Eqn. 5).

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 δ whether BLinC is subgame-perfect or not. Note that the gain that any group G of players would get from a shortfall of size $x \geqslant 0$ 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 $\gamma_x = g(Q^\star - x)/(Q^\star - x)$ are the average unit costs at $Q^\star - x$, with $\gamma_x \leqslant \gamma$ 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 $\delta\gamma(L_G^{\max} - Q_G^\star)$. Hence a sufficient condition for such a shortfall to be unprofitable is that the former be smaller than the latter, which is equivalent to

$$x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x > Q_G^\star\gamma\delta + L_G^{\max}\gamma(1-\delta). \quad [13]$$

We will now show that if the target allocation $q^\star$ is *profitable* for each player, so that $Q^\star\beta_G - Q_G^\star\gamma > 0$ for all G, and if δ 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 $x \geqslant 0$. Let $\varepsilon_G = (Q^\star\beta_G - Q_G^\star\gamma)/2(1+\beta_G) > 0$, $\varepsilon = \min_G \varepsilon_G > 0$, and $x_0 = Q^\star - \varepsilon$. Choose the bounds $\ell_i^{\min}$ small enough so that $L_G^{\max} \geqslant Q^\star$ and $L_G^{\max} > (Q_G^\star\gamma + \varepsilon - x_0(\beta_G - \gamma))/\gamma_{x_0}$ for all G. Then, for all G and x,

$$x(\beta_G - \gamma_x) + L_G^{\max}\gamma_x > Q_G^\star\gamma + \varepsilon. \quad [14]$$

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

This means that, for large enough compensation factor α, 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 (1) shows that then also no infinite number of deviations can pay. In particular, this shows that with individually profitable targets $q_i^\star$, large enough α and δ, 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. 13 is fulfilled when $Q_G^\star < \min\{\beta_G - \beta(1-\delta)/2, \beta_G(1 - \beta_G/2\beta)/\delta - \beta(1-\delta)/2\delta\}$ for all G. For large enough δ, this can be fulfilled by a target allocation proportional to marginal benefits, $q_i^\star = \beta_i/2$. That allocation leads to payoffs which are also proportional to marginal benefits, $P_i = \beta_i\beta/4$.

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 (2–4)[1] if the global optimal emissions abatement path $Q^\star(t)$ is allocated in a certain way and the explicit time-dependency of the benefit functions $b_i(t)$ is taken into account properly. For the chosen model parameters, a moderate α of 1.22 fulfils Eqn. 6. We tested an allocation under which half of the long-term global payoff as compared to the business-as-usual scenario was distributed so that each region's per capita payoff in purchasing power (PPP) increases by the same amount, and the other half was distributed in proportion to regional GDP (based on 1995 population, PPP, and GDP data).[2] That allocation gives four players negative contribution targets $q_i^\star(t)$, i.e., more emissions permits than under business-as-usual. 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, we could verify 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 industrialized regions could profit from free-riding. Still, such targets can be stabilized by using the unbounded strategy LinC.

Supporting Information: Target allocation

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* $\pi = \{S_1, \ldots, S_m\}$, 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 $(Q_1^{\mathrm{PSE}}, \ldots, Q_m^{\mathrm{PSE}})$ be the contributions in a PSE of this game. These can be determined by replacing n, f_i, and q_i in Eqns. 9 and 10 by m, f_{S_j}, and Q_j^{PSE}, respectively. Now assume each S_j has agreed internally on some individual target allocation $q_i^\star$ of Q_j^{PSE}, so that $\sum_{i \in S_j} q_i^\star = Q_j^{\mathrm{PSE}}$, 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.

[1] 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 (5); 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 (6)); play simulated from 2010–2110.

[2] This is basically the average of the sharing rules 2 and 3 from (7), for which those authors found that only very small long-term coalitions were stable in their model.

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

Both classical "cooperative" game theory and the newer more sophisticated theory of zt coalition formation (8) 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* $v = v^{\mathrm{PSE}}$. 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.[3]

Consider for example the public good game with symmetric linear benefits $f_i(Q) = Q$ and quadratic costs $g(Q) = \max\{0, Q\}^2$. Then it can be shown (see SI: Examples) that v^{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. This extreme form of v has been analysed in the literature as a kind of quintessential example of cooperative games with *externalities* since it also arises naturally from Cournot-Nash equilibria in *Cournot oligopolies*[4] For $n = 5$ (and similarly for larger n), one approach (9) 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 (10) 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 (11; 12) 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 (13), such behaviour 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^{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:* $v^0(S_j \cup S_k) \geqslant v^0(S_j) + v^0(S_k)$. 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* (14) 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 (15) 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 (16). 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

$$P(S) = \left[v(S) + \sum_{i \in S} w_i P(S \setminus \{i\}) \right] / \sum_{i \in S} w_i. \qquad \textbf{[15]}$$

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 $v(N) = P^\star$ and $v(S) = \sum_{i \in S} P_i^{\mathrm{PSE}}$ 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 $\phi_i = P_i^{\mathrm{PSE}} + w_i(P^\star - P^{\mathrm{PSE}})$, 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 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 (9) 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 $\sum_i a_i = 1/4$, $a_1 \geqslant 1/16$, $a_2 + a_3 \geqslant 1/16$, and $a_4 + a_5 \geqslant 1/16$. 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 $\sum_i b_i = 1/4$, $b_1 \geqslant 1/25$, $b_2 \geqslant 1/25$, $b_3 \geqslant 1/25$, $b_4 \geqslant 1/20$, and $b_5 \geqslant 1/20$.

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

[3] 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, $C_i = q_i g(Q)/Q$.

[4] 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.

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 $a_2, a_3, a_4, a_5 \geqslant 1/20$. But then $a_1 \leqslant 1/4 - 4/20 = 1/20 < 1/16$, 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 $\sum_i a'_i = 1/4$, $a'_1 \geqslant 1/16$, $a'_2 \geqslant 1/16$, and $a'_3 + a'_4 + a'_5 \geqslant 1/16$. 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 $\sum_i c_i = 1/4$, $c_1 \geqslant 1/9$, and $c_2 + c_3 + c_4 + c_5 \geqslant 1/9$. 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 intermediate 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 $\geqslant 1/20$, so that $c_1 \leqslant 1/4 - 3/20 = 1/10 < 1/9$ in contradiction to $c_1 \geqslant 1/9$. 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 $m \in \{1, \ldots, n\}$, the equation

$$F'(Q) = (m-1)h(Q) + g'(Q) \qquad \textbf{[16]}$$

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

$$v^{\mathrm{PSE}}(S, \pi) = k f_0(Q_m) + h(Q_m) \frac{h(Q_m) - k f'_0(Q_m)}{h'(Q_m)}. \qquad \textbf{[17]}$$

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

Then each structure π can either stabilize itself by giving each player exactly $P^\star/n$, 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 π is stable with a since it gives each coalition at least the same as in the PSE, $P^\star/n > v^{\mathrm{PSE}}(\{i\}, \pi)$ 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 π, we distinguish two cases to show that it is also true for π:

(i) If a gives each coalition $S \in \pi$ at least $v^{\mathrm{PSE}}(S, \pi)$, 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 π' that might arise from leaving, they must expect that, because of the announcements, π' 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 $v^{\mathrm{PSE}}(S, \pi)$, but some other allocation b stabilizes π. Then $v^{\mathrm{PSE}}(S, \pi) > kP^\star/n$ where $k = |S|$, and b gives each coalition $T \in \pi$ at least $v^{\mathrm{PSE}}(T, \pi)$. 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

$$\begin{aligned} v^{\mathrm{PSE}}(T, \pi) &= f_0(Q_m) + h(Q_m) \frac{h(Q_m) - f'_0(Q_m)}{h'(Q_m)} \\ &\geqslant \left(k f_0(Q_m) + h(Q_m) \frac{h(Q_m) - k f'_0(Q_m)}{h'(Q_m)} \right) / k \\ &= v^{\mathrm{PSE}}(S, \pi)/k > P^\star/n, \end{aligned} \qquad \textbf{[18]}$$

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 π, no other allocation b will. Finally, taking $\pi = \{N\}$, this proves that the grand coalition can stabilize by agreeing on the symmetric allocation $a_i = P^\star/n$.[5]

So, in contrast to (9), 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 (17) 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 (18) 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^{PSE} as given by Eqns. 9 and 10. At each point τ 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, $\vec{q}^{\,0} = \vec{q}^{\,\mathrm{PSE}}$. At each point τ, each player i considers

[5] 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.

the probability that agreement will be reached to be τ. 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 $q_i^{\rm PSE}$ as well. The long-term payoff player i expects if she contributes q_i in each period is

$$(1-\tau)W_1 P_i(q_i, q_i + Q_{-i}^{\rm PSE}) + \tau V_i(q_i, \vec{q}^{\,\tau}) \qquad [19]$$

where $Q_{-i}^{\rm PSE} = Q^{\rm PSE} - q_i^{\rm PSE}$, $P_i(q_i, Q) = f_i(q_i) - q_i h(Q)$ 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^τ 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 Appendix on 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 $W_1(f_i(q_i) - g(q_i))$. The best response q_i to the current beliefs of i at τ 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

$$a_i^\tau(\vec{q}^{\,\tau}) = (1-\tau)P_i(q_i^\tau, q_i^\tau + Q_{-i}^{\rm PSE}) + \tau P_i(q_i^\tau, Q^\tau), \qquad [20]$$

$$p_i^\tau(q_i) = (1-\tau)P_i(q_i, q_i + Q_{-i}^{\rm PSE}) + \tau P_i(q_i, q_i). \qquad [21]$$

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

$$\begin{aligned} 0 =& (1-\tau)[f_i'(q_i + Q_{-i}^{\rm PSE}) - h(q_i + Q_{-i}^{\rm PSE}) - q_i h'(q_i + Q_{-i}^{\rm PSE})] \\ &+ \tau[f_i'(q_i) - g'(q_i)]. \end{aligned} \qquad [22]$$

Denote this q_i by $\tilde{q}_i^\tau$ and note that it depends on τ 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 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,

$$\begin{aligned} 0 &= \partial A^\tau(\vec{q}^{\,\tau})/\partial q_i^\tau \\ &= (1-\tau)[f_i'(q_i^\tau + Q_{-i}^{\rm PSE}) - h(q_i^\tau + Q_{-i}^{\rm PSE}) - q_i^\tau h'(q_i^\tau + Q_{-i}^{\rm PSE})] \\ &\quad + \tau[F'(Q^\tau) - g'(Q^\tau)]. \end{aligned} \qquad [23]$$

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 $Q^\tau = Q^\star$, 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 $\tau \to 1$. 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. 23 is

$$0 = (1-\tau)[\beta_i - 2q_i^\tau - Q_{-i}^{\rm PSE}] + \tau[\beta - 2Q^\tau]. \qquad [24]$$

We can first determine Q^τ from their sum, giving

$$Q^\tau = \frac{[\beta - (n-1)Q^{\rm PSE}](1-\tau) + n\beta\tau}{2(1-\tau) + 2n\tau} \qquad [25]$$

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

$$q_i^\tau = \frac{\beta_i - Q_{-i}^{\rm PSE}}{2} + \tau\frac{\beta - 2Q^\tau}{2(1-\tau)} \qquad [26]$$

which converges to

$$q_i^1 = \frac{\beta_i - Q_{-i}^{\rm PSE}}{2} + \frac{n-1}{2n}Q^{\rm PSE} = \beta_i - \beta/2n. \qquad [27]$$

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 β_i if $\zeta < 2$, and they are larger for those with smaller β_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 (2–4), using the same parameters as in the Appendix: It has approximately cubic costs ($\zeta = 3$), and when we solve Eqns. 23 numerically, the resulting allocation of the optimal global payoff gives the US, Japan, and the EU (having large β_i) a share of about 4%, 6%, and 4% of the payoff, respectively, and the remaining nine world regions (having small β_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 the Appendix, so it does not seem a likely outcome of the emissions game.

References

1. Abreu, D. (1988) *Econometrica* **56**, 383–396.
2. Finus, M, Ierland, E. V, & Dellink, R. (2006) *Economics of Governance* **7**, 271–291.
3. Rubio, S. J & Ulph, A. (2006) *Oxford Economic Papers* **58**, 233–263.
4. Weikard, H.-P, Dellink, R, & van Ierland, E. (2010) *Environmental and Resource Economics* **45**, 573–596.
5. Ellerman, A. D & Decaux, A. (1998) *MIT Joint Program on the Science and Policy of Global Change* **Report 40**, 1–33.
6. Nordhaus, W. D. (1994) *Managing the global commons: the economics of climate change*. (MIT Press, Cambridge, MA).
7. Weikard, H.-P. (2006) *Oxford Economic Papers* **58**, 209–232.
8. Ray, D. (2007) *A Game-Theoretic Perspective on Coalition Formation (The Lipsey Lectures)*. (Oxford University Press, USA, Oxford).
9. Ray, D & Vohra, R. (1997) *Journal of Economic Theory* **73**, 30–78.
10. Ray, D & Vohra, R. (1999) *Games and Economic Behavior* **26**, 286–336.
11. Yi, S.-S. (1997) *Games and Economic Behavior* **20**, 201–237.
12. Morasch, K. (2000) *International Journal of Industrial Organization* **18**, 257–282.
13. Coase, R. H. (1960) *Journal of Law and Economics* **3**, 1–44.
14. Shapley, L. S. (1951) *Rand Corporation Research Memoranda* **RM-670**, 1–19.
15. Kalai, E & Samet, D. (1987) *International Journal of Game Theory* **16**, 205–222.
16. Monderer, D, Samet, D, & Shapley, L. S. (1992) *International Journal of Game Theory* **21**, 27–39.
17. Harsanyi, J. C & Selten, R. (1988) *MIT Press Books*.
18. Harsanyi, J. C. (1975) *International Journal of Game Theory* **4**, 61–94.