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

Jobst Heitzig [*], **Kai Lessmann** [*], **and Yong Zou** [*]

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

Final authors' draft as accepted for publication in PNAS on July 25, 2011. Original will appear as DOI:10.1073/pnas.1106265108 the week of August 29, 2011.

As the Copenhagen Accord indicates, most of the international community agrees that global mean temperature should not be allowed to rise more than two degrees Celsius above pre-industrial levels to avoid unacceptable damages from climate change. The scientific evidence distilled in the IPCC's 4th Assessment Report and recent reports by the U.S. National Academies shows that this can only be achieved by vast reductions of greenhouse gas (GHG) emissions.
Still, international cooperation on GHG emissions reductions suffers from incentives to free-ride and to renegotiate agreements in case of non-compliance, and the same is true for other so-called 'public good games.' Using game theory, we show how one might overcome these problems with a simple dynamic strategy of Linear Compensation (LinC) when the parameters of the problem fulfill some general conditions and players can be considered to be sufficiently rational.
The proposed strategy redistributes liabilities according to past compliance levels in a proportionate and timely way. It can be used to implement any given allocation of target contributions, and we prove that it has several strong stability properties.

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

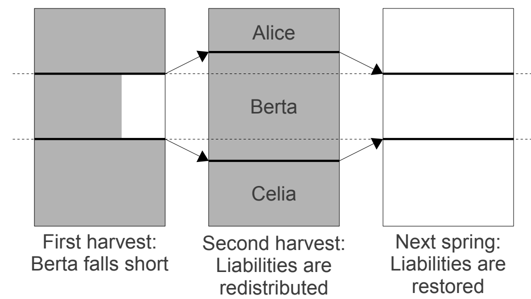

Fig. 1. Illustration of Linear Compensation in a simple public good game. Alice, Berta, and Celia farm their back-yard for carrots. Each has her individual farming liability (thick separators) but harvests are divided equally. In the first year, Berta falls short of her target by some amount (white area).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

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

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

We show that players can ensure compliance with a given initially negotiated target allocation of contributions by adopting a certain simple dynamic *strategy* to choose their actual contributions over time. In each period, the allocation of liabilities is redistributed in 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 a strategically important adjustment to keep total liabilities constant. This strategy will be called 'Linear Compensation' (LinC), and its basic idea is illustrated in Fig. 1 in a fictitious community gardening example. In the emissions game, these liabilities to reduce emissions then translate into emissions allowances via the formula *allowance = reference emissions – liability*. Our assumptions and the proposed strategy are summarized in Table 1.

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

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

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

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

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

The public good game:

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

The strategy of Linear Compensation (LinC):

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

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

Framework

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

In the emissions game, $q_i(t)$ would be the difference between i's hypothetical amount of GHG emissions in period t in some 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 Discussion and in SI: Validity of assumptions.

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

Unlike in many other models of public goods, we assume here that total costs are shared in a way that equalizes marginal costs. E.g., costs might be shared in proportion to individual contributions, giving $c_i(t) = q_i(t)C(t)/Q(t)$. Or, what is more realistic if there is a perfect competition market for contributions, costs might be shared according to a rule based on marginal cost pricing.[2] In both cases, one has the following convexity property on which our results will

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

rely: for each Q, there is some 'cost sensitivity' $\gamma(Q)$ so that (i) if q_i and Q are both lowered by an amount $x > 0$, then c_i gets lowered by at most $x\gamma(Q)$, (ii) if q_i is raised by an amount $x > 0$ but Q is kept constant by lowering the other values q_j, then c_i raises by at least $x\gamma(Q)$, and (iii) $c_i = 0$ for $Q \leqslant 0$. In other words, lowering your contributions by x saves you at most costs of $x\gamma(Q)$, but if x contributions are redistributed from others to you, your costs raise by at least that same amount. It is easy to see that in the proportional cost-sharing case, $\gamma(Q)$ equals average costs $g(Q)/Q$, while in the marginal cost pricing case, $\gamma(Q)$ equals marginal costs $g'(Q)$.

In the emissions game, the benefits of reducing emissions by 1 Gt CO_2-equivalents in period t correspond to all avoided welfare losses that would have been caused at times after t by that additional 1 Gt of emissions, properly discounted to reflect the corresponding time difference, and using any suitable welfare measure such as consumption, income, gross domestic product (GDP), etc. (20–22). The above form of the costs c_i seems justified when we assume an international emissions market between firms, similar to the European Union Emission Trading Scheme (EU ETS). A simple example cost-benefit structure is that of linear benefits and linear marginal costs (23): $f_i(Q) = \beta_i Q$ with $\beta_i > 0$, $g(Q) = Q^2$ for $Q > 0$, and $g(Q) = 0$ for $Q \leqslant 0$. 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 $Q(t) \to \pm\infty$. In the emissions game, large positive or negative values for some q_i can obtain if large amounts of emissions permits are traded. Although the strategy we will propose below prescribes such large values of q_i only in cases where there has already been an irrationally large earlier deviation, this might still lead to problems in practice (for an alternative model, see SI: Bounded liabilities).

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

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

$$f(Q^\star) - g(Q^\star) = \max, \quad g'(Q^\star) = f'(Q^\star) > f_i'(Q^\star). \qquad [1]$$

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

Finally, let us assume that players can enter no legally binding and enforceable agreements (since this is the worst case assumption when studying the possibility of cooperation) but have somehow chosen in advance (before period one) an allocation of the optimum target into individual targets $q_i^\star$, 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 $Q_G^\star = \sum_{i\in G} q_i^\star$.[3]

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

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
$f(Q), f_i(Q)$	benefit functions, total and for player i
$g(Q)$	total cost function
$\gamma(Q), \gamma^\star$	cost sensitivity function, and value at target $Q^\star$
$\ell_i(t)$	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

Table 2. Main symbols used in this article

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

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

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

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

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

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

a defection is void and cannot deter free-riding under assumptions of rationality and without commitment possibilities (38).[4]

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

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

Results

Avoiding renegotiations. Let us deal with the question of renegotiations first. The crucial idea to avoid those in our kind of game is to keep total contributions constant and only redistribute them as a reaction to past behaviour. Consider a strategy s which, in each period t, tells all players to choose their contributions $q_i(t)$ in a certain way which makes sure that the total target is met, $Q(t) = Q^\star$. 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 there is no possible situation in the game that would cause all players to agree to change the strategy. In game-theoretic terms, such a strategy is 'strongly perfect', i.e., Pareto-efficient in all subgames. It will thus be *strongly renegotiation-proof* (26; 27) if we manage to do the redistribution of contributions in $t+1$ in a way that makes free-riding in t unprofitable in the long run. This we will do next.[5]

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

$$x(\gamma^\star - \beta_G). \qquad [2]$$

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

$$x(\gamma^\star - \beta_G)/\delta. \qquad [3]$$

So, G's joint gains in t are overcompensated by these losses in $t + 1$. Although free-riding for one period might be profitable for some individual members of G, there is always at least one member of G for whom it is not. Fig. 1 illustrates the basic idea. We will show next how the same kind of redistribution can be used to deter also every conceivable sequence of deviations from the target path.

The strategy of Linear Compensation (LinC). A simple strategy that does this assigns each player i in each period t a certain *individual liability* $\ell_i(t)$ which that player should contribute in t. In period one, liabilities equal the negotiated targets, $\ell_i(1) = q^\star_i(1)$. 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

$$\ell_i(t + 1) = q^\star_i + [d_i(t) - \bar{d}(t)] \cdot \alpha. \qquad [4]$$

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 the previous section to apply in all possible situations, whatever the contributions have been before t. In the simple free-riding situation discussed in the previous section, the group's joint shortfall equals x and the mean shortfall is $\bar{d}(t) = x/n$. Hence G's joint additional liability in $t + 1$ is $[x - |G|x/n] \cdot \alpha$, where $|G| < n$ is the number of players in G. If this is at least x/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

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

where the maximum is taken over all possible groups of players G. If it is known that the benefit functions of all players are equal, then $\beta_G = C'(Q^\star)|G|/n \geqslant \gamma^\star|G|/n$ and Eqn. [5] simplifies to $\alpha > [n\gamma^\star - C'(Q^\star)]/\gamma^\star\delta(n - 1)$, 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. Now, a player i who complies with the liabilities defined by Eqns. [4] and [5] by putting $q_i(t) = \ell_i(t)$ is said to apply the strategy of 'Linear Compensation' (LinC).

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

[5] 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' (26; 27). See also SI: Renegotiations when targets are not optimal.

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

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

Discussion

We have presented a simple strategy by which players in a public good game can keep each other in check in the provision of agreed target contributions. Our approach can be interpreted as a combination of a proportionate version of the punishment approach that strategies like 'tit for tat' use in the Prisoners' Dilemma, and the repayment approach that is already included in the Kyoto mechanism. This combination has been formally shown here to have strong game-theoretic stability properties in situations where some simplifying assumptions hold, a feature that is not true of strategies that use only one of the two ingredients. In Axelrod's (36) terminology, our strategy, LinC, is 'nice' in that it cooperates unless provoked, 'retaliating' when provoked, 'forgiving' when deviators repay, and uses '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. E.g., future work might use an improved model of the emissions game in which the assumption of identical periods is replaced by certain path-dependencies: Real-world benefit functions f_i depend on GHG stocks and hence on time and emission history, and also the cost function g depends on time and emission history because of technological progress. Since past contributions will reduce future marginal costs, this will lead to a non-constant optimal abatement path $Q^\star(t)$. However, these effects will probably not weaken LinC's stability when $q_i^\star$ is replaced by a time-dependent target allocation $q_i^\star(t)$ of $Q^\star(t)$ that is computed according to some initially negotiated rule (e.g., in fixed proportions). This is because then the Pareto-efficiency argument for renegotiation-proofness still holds, while shortfalls would slow down technological progress and lead to even higher marginal costs in the punishment period.

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

In addition to such model refinements, future work should also (i) assess the possibility of players to "bind their hands" ahead of time by making long-term investment decisions that reduce their own ability to choose $q_i(t)$ at t, (ii) study the influence of incomplete information due to restricted monitoring capacities, finite planning horizons and of other forms of 'bounded rationality' (43), (iii) link emissions reductions with other issues (44), (iv) include possible altruism, reputation, and status effects, also using experimental approaches such as (45).

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

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

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

---

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

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

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

[10] The value of δ however does play a role when, in addition to our assumptions, liabilities shall be bounded. This is further explored in SI: Bounded liabilities.

Appendix: Why infinite sequences of deviations do not pay

Suppose all players apply LinC by putting $q_i(t) = \ell_i(t)$ except that from some period t_0 on, a group G of players play a *deviation strategy* s that leads to joint shortfalls $\sum_{i\in G} d_i(t) = x_t$ in each period $t \geqslant t_0$. 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 $w_{t,r}$ fulfill the conditions

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

E.g., players could use exponential discounting with $w_{t,r} = \varepsilon^r$, $\delta < \varepsilon < 1$, and $W_t = 1/(1-\varepsilon)$.[11] 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 apply 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 $x_t\gamma^\star$ according to Eqn. [2]. Hence $\Delta(s, \tilde{s}) \leqslant \sum_{t\geqslant t_1} w_{t_0,t-t_0} x_t \gamma^\star$. 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, so neither is s.

(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, $L_G(t) = \sum_{i\in G} \ell_i(t) \geqslant Q^\star_G$. Hence their joint costs $C_G(t)$ are either zero if $x_t \geqslant Q^\star$, since then total costs are zero, or they are by at most $Q^\star\gamma^\star$ smaller than in the case where $L_G(t) = Q^\star_G$. In other words, $C_G(t)$ is bounded from below by some value $C_G^{\min}$. Concerning benefits, let $f_G(Q) = \sum_{i\in G} f_i(Q)$ and let $\beta_G = f'_G(Q^\star)$ 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 $(Q^\star - Q^\star_G)\gamma^\star + f_G(Q^\star) - \beta_G x_t$, so that G's discounted long-term payoff $U_G(t_0)$ is then at most

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

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

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

References.

1. Dinar A, Albiac J, & Sánchez-Soriano J (2008) *Game Theory and Policy Making in Natural Resources and the Environment* (Routledge, London).
2. Aldy JE & Stavins RN, eds (2010) *Post-Kyoto International Climate Policy: Implementing Architectures for Agreement* (Cambridge University Press).
3. Karp L & Zhao J. (2010) in (2), pp 530–562
4. Finus M (2008) Game theoretic research on the design of international environmental agreements: insights, critical remarks, and future challenges. *Int Rev Environ Res Econ* 2:29–67.
5. Carraro C & Siniscalco D (1993) Strategies for the international protection of the environment. *J Public Econ* 52:309–328.
6. Barrett S (1994) Self-enforcing international environmental agreements. *Oxford Econ Pap* 46:878–894.
7. Barrett S (2001) International cooperation for sale. *Eur Econ Rev* 45:1835–50.
8. Barrett S (1999) A theory of full international co-operation. *J Theor Polit* 11:519–541.
9. Asheim GB, Froyn CB, Hovi J, & Menz FC (2006) Regional versus global cooperation for climate control. *J Environ Econ Manag* 51:93–109.
10. Froyn CB & Hovi J (2008) A climate agreement with full participation. *Econ Lett* 99:317–319.
11. Asheim GB & Holtsmark B (2009) Renegotiation-proof climate agreements with full participation: conditions for Pareto-efficiency. *Environ Resour Econ* 43:519–533.
12. Weikard HP, Dellink R, & van Ierland E (2010) Renegotiations in the greenhouse. *Environmental and Resource Economics* 45:573–596.
13. Dutta PK & Radner R (2004) Self-enforcing climate-change treaties. *P Natl Acad Sci USA* 101:5174–9
14. Rubio SJ & Casino B (2005) Self-enforcing international environmental agreements with a stock pollutant. *Spanish Economic Review* 7:89–109.
15. Dutta PK & Radner R (2006) A game-theoretic approach to global warming. *Adv Math Econ* 8:135–153.
16. Rubio SJ & Ulph A (2007) An infinite-horizon model of dynamic membership of international environmental agreements. *J Environ Econ Manag* 54:296–310.
17. de Zeeuw A (2008) Dynamic effects on the stability of international environmental agreements. *J Environ Econ Manag* 55:163–174.
18. Dutta PK & Radner R (2009) A strategic analysis of global warming: theory and some numbers. *J Econ Behav Organ* 71:187–209.
19. Mason CF, Polasky S, & Tarui N (2010) Cooperation on Climate-Change Mitigation. *University of Exeter Working Papers.*
20. Intergovernmental Panel on Climate Change (2007) *IPCC Fourth Assessment Report: Climate Change 2007 (AR4)* (Cambridge University Press, Cambridge, UK).
21. U.S. National Academies of Sciences (2010) *Advancing the Science of Climate Change* (National Academies Press, Washington, D.C.).
22. U.S. National Research Council (2010) *Limiting the Magnitude of Climate Change* (National Academy Press, Washington, D.C.).
23. Finus M, Ierland EV, & Dellink R (2006) Stability of climate coalitions in a cartel formation game. *Economics of Governance* 7:271–291.
24. Ellerman AD & Decaux A (1998) Analysis of post-Kyoto CO_2 emissions trading using marginal abatement curves. *MIT Joint Program on the Science and Policy of Global Change* 40:1–33.
25. Ellerman AD (2010) in (2), pp 88–150.
26. Farrell J & Maskin E (1989) Renegotiation in repeated games. *Game Econ Behav* 1:327–360.
27. Bergin J & MacLeod B (1993) Efficiency and renegotiation in repeated games. *J Econ Theory* 61:42–73.
28. Carraro C, Eyckmans J, & Finus M (2006) Optimal transfers and participation decisions in international environmental agreements. *The Review of International Organizations* 1:379–396.
29. Carraro C, Marchiori C, & Oreffice S (2009) Endogenous minimum participation in international environmental treaties. *Environ Resour Econ* 42:411–425.
30. Lessmann K & Edenhofer O (2010) Research cooperation and international standards in a model of coalition stability. *Resour Energy Econ* 33:36–54.
31. Weikard HP, Finus M, Altamirano-Cabrera JC (2006) The impact of surplus sharing on the stability of international climate agreements. *Oxford Econ Pap* 58:209–232.
32. Kverndokk S & Rose A (2008) Equity and justice in global warming policy. *Fondazione Eni Enrico Mattei Working Papers* 239:1–66.
33. 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).
34. Aldy JE & Stavins RN (2010) in (2), pp 899–929.
35. Aumann RJ (2006) War and peace. *P Natl Acad Sci USA* 103:17075–8.
36. Axelrod R (1981) The emergence of cooperation among egoists. *Am Polit Sci Rev* 75:306–318.
37. Axelrod R & Keohane RO (1985) Achieving cooperation under anarchy: strategies and institutions. *World Polit* 38:226–254.
38. van Damme E (1989) Renegotiation-proof equilibria in repeated prisoners' dilemma. *J Econ Theory* 47:206–217.
39. Dal Bó P & Fréchette GR (2011) The evolution of cooperation in infinitely repeated games: experimental evidence. *Amer Econ Rev* 101:411–429.
40. Davis DD & Holt CA (1999) Equilibrium cooperation in two-stage games: experimental evidence. *Intern J Game Theory* 28:89–109
41. Barrett S (2005) in *Handbook of Environmental Economics, Vol. 3,* eds Mäler KG & Vincent JR (Elsevier), pp 1457–1516.
42. Abreu D (1988) On the theory of infinitely repeated games with discounting. *Econometrica* 56:383–396.
43. Jones BD (2001) *Politics and the architecture of choice: Bounded rationality and governance* (University of Chicago Press, Chicago).
44. Carraro C & Marchiori C (2004) in *Endogenous strategic issue linkage in international negotiations*, eds Carraro C & Fragnelli V. (Edward Elgar), pp 65–84.
45. Milinski M, Semmann D, Krambeck HJ, & Marotzke J (2006) Stabilizing the Earth's climate is not a losing game: supporting evidence from public goods experiments. *P Natl Acad Sci USA* 103:3994–8.

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

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

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