Dynamic cooperation with tipping points in the climate system

ROBERT C. SCHMIDT*

October 26, 2015

Abstract

Tipping points in the climate system can stabilize climate treaties; the stabilizing effect, however, vanishes when the location of the threshold is sufficiently uncertain (Barrett, 2013). We demonstrate that in a dynamic setting, additional welfare gains can improve the prospects of cooperation. In our model, intertemporal efficiency gains result from abatement costs that are convex in each period. While non-cooperative countries tend to postpone their abatement efforts “until the last minute” as a result of the free-rider incentive, cooperation allows countries to allocate their abatement efforts efficiently over time. We show that cooperation often improves the outcome substantially, and arises endogenously in the model. Our main theoretical results are confirmed by experimental evidence.

*School of Business and Economics, Humboldt University, Spandauer Str. 1, 10178 Berlin, Germany; E-mail: robert.schmidt.1@wiwi.hu-berlin.de

Keywords: climate treaty, abatement, long-term commitment, cooperation, free-riding

JEL classification: D62, F53, H23, Q55
1 Introduction

People often think of climate change as a gradual phenomenon: the more greenhouse gases are emitted, the higher is the temperature increase to be expected in the long run. Climate change is, however, not a gradual phenomenon. Due to feedback effects and nonlinearities in the climate system, thresholds (so-called ‘tipping-points’) for irreversible and potentially catastrophic changes can emerge. From a game-theoretic point of view, the existence of such dangerous thresholds can improve the prospects of cooperation (Barrett 2013). However, this stabilizing device for collective action against climate change fails if the location of the threshold becomes (sufficiently) uncertain.

While Barrett (2013) uses a static framework, this paper extends the model by splitting up the time horizon into two periods. This simple extension is sufficient to grasp some important dynamic aspects that a static model cannot capture. In particular, efficiency in abatement requires countries not only to internalize environmental externalities within a period, but also to allocate these efforts efficiently over time, assuming that abatement costs are convex within each period. Non-cooperative countries, by contrast, tend to “wait until the last minute” when preventing a catastrophe. This is because each country realizes that it can shift part of its abatement costs towards the other countries by lowering its effort in the first period, as long as the other countries respond to this change by raising their efforts in the second period.

Let us illustrate this with the help of a simple numerical example. Suppose, there are 10 countries that can reduce their emissions in two periods (we assume there is no discounting). In each period, a country can either abate 0, 1, or 2 units of emissions. The abatement of one unit costs 1 unit of money, whereas to abate two units, 4 units of money are needed. Suppose, the threshold is avoided if the aggregate abatement of all countries over both periods amounts to at least 20 units of emissions. Otherwise, a damage cost of 4 units of money is incurred by each country. The first-best is reached if each country abates one unit in each period, so that the total abatement equals 20 over both periods and the threshold is avoided. Can this outcome be sustained without cooperation? To answer this question, start from the first-best, and suppose some country unilaterally deviates by reducing its abatement effort in the first period to zero. Then in the second period avoidance of the threshold is still an equilibrium. This merely requires that countries can coordinate, because one country now has to abate two units in order to avoid the threshold. Since countries are symmetric, let us assume that the identity of this country is randomly determined.¹ The deviating country then saves 1 unit of money in period 1, and (in expectation) incurs additional costs in period 2 of only 0.3. Hence, the deviation in period 1 is profitable. There is a subgame perfect Nash equilibrium in

¹In the model that is presented in the main part of this paper, countries’ abatement decisions are continuous.
which the catastrophe is avoided. It entails zero abatement in period 1 by all countries, and two units of abatement per country in period 2. Welfare per country is, then, as low as when countries do not abate at all.

The reason for this pessimistic result is an *intertemporal free-rider incentive* that induces non-cooperative countries to “wait until the last minute” when avoiding the threshold, whereas efficiency requires them to start with their abatement efforts already in the first period when costs are convex in each period. Assuming that coalition members can commit to their current and future abatement efforts once a climate contract is signed, whereas non-signatories cannot commit in this way, cooperation in a setting with a known threshold is, however, in general stable. Signatories then choose their abatement efforts sufficiently low so that non-signatories still find it optimal to avoid the catastrophe (they “break even”). As a result, a grand coalition emerges.

The situation is different when the location of the threshold is uncertain. As Barrett (2013) highlights, under this type of uncertainty the stabilizing effect upon cooperation often vanishes because the expected marginal benefits of abatement are lower. In a dynamic context, however, there are additional gains of cooperation because signatories can allocate their abatement efforts more efficiently across periods. This raises the scope for cooperation.

We distinguish between two types of welfare losses in our model that can result from the free-rider incentive. On the one hand, a lack of cooperation can induce countries to allocate their abatement efforts inefficiently over time. These welfare losses can be severe, as our simple numerical example illustrated. This problem typically arises if the location of the threshold is (almost) certain. On the other hand, a lack of cooperation can lead to an inefficiently high probability of reaching the threshold because the overall efforts are inefficiently low. This problem often occurs if the uncertainty about the location of the threshold is high. Interestingly, we find that the degree of inefficiency is non-monotonic in the parameters of the model: there exist intermediate parameter values where the outcome under no cooperation coincides with the fully cooperative outcome, while for more extreme parameter values (in either direction), the outcome is inefficient.

We demonstrate that in many cases a coalition forms, and that cooperation often leads to significant welfare gains. In some cases, a boundary solution is obtained where the coalition is just large enough to become active. While signatories then just break even, welfare gains for the non-signatories can be significant. In other cases, a grand coalition emerges because signatories are better off than non-signatories. Our general conclusion is that cooperation can play an important role in a dynamic setting with an uncertain threshold. Our main theoretical predictions are confirmed by an experiment.

---

2An alternative reason are learning-by-doing externalities. Efficiency then also requires “smooth” abatement profiles so that knowledge is accumulated in the first period. Also in this case, cooperation is often a stable outcome. Details of this analysis can be obtained from the author upon request.
Related Literature

Recently, a number of experimental studies have been conducted that shed light on the impact of thresholds for dangerous climate damages on cooperation. Miliniski et al. (2008) conduct an experiment that is similar to the simple numerical example outlined above: a number of players can contribute to the public good of ‘climate stabilization’ in several periods, and only the total contribution of all players over all periods matters for the occurrence of a ‘catastrophe’. The main difference is that in our setup, the costs of contributing to the public good are convex in each period, so that the social optimum is uniquely determined. In the setup of Miliniski et al. (2008), by contrast, players can shift their contributions across periods and players without affecting the overall efficiency of the final outcome. This may help to explain why participants in the experiment often failed to coordinate, apart from their lack of a possibility to communicate.³

The impact of uncertainty about the location of a threshold upon players’ ability to coordinate on an efficient outcome is investigated experimentally by Barrett and Dannenberg (2014). Consistent with the theory (Barrett 2013), players often manage to avoid the threshold if its location is (almost) certain, whereas coordination breaks down if the location becomes more uncertain. Our paper demonstrates that the prospects of cooperation may not be so gloomy even when the location of the threshold is uncertain.

In an experimental setup with stochastic damages (but no threshold), Köke, Lange, and Nicklisch (2015) show that individuals tend to switch to more cooperative behavior following a damage event, in particular if voluntary contributions reduce the probability of such events (rather than the size of the damages). The authors argue that such behavior is consistent with ex post regret.

The possibility of countries (players) to cooperate by forming a coalition is investigated experimentally by Dannenberg, Lange, and Sturm (2014). The contributions of coalition members are either imposed exogenously so as to maximize their joint welfare, or determined by the players. While the former case mirrors the standard approach in the theory on climate cooperation (e.g., Barrett 1994), in the latter case coalition members may contribute less (or more) than what is optimal from a theoretical perspective.⁴

The results support the pessimistic findings of Barrett (1994) and other authors from the literature on climate cooperation. Lange and Vogt (2003) analyze the formation of a climate coalition when countries have a preference for equity.

³See also Tavoni et al. (2011) and Dannenberg et al. (2011).
⁴Finus and Maus (2008) show that suboptimal contributions by signatories of a climate treaty can raise participation in the coalition and, as a result, lead to higher welfare gains than under the optimal contributions (optimal for a given coalition size). See also Barrett (2002).
2 Model

There are \( N \) ex-ante symmetric countries, that can reduce their emissions (relative to some baseline) in two periods. Country \( i \)'s abatement in period \( t \) \((t = 1, 2)\) is denoted \( q_{t,i} \), and causes abatement costs of \( C(q_{t,i}) = cq_{t,i}^2/2 \). Hence, \( c > 0 \) is the slope of the linear marginal abatement cost function. Total abatement in period \( t \) is denoted by \( Q_t = \sum_{i=1}^{N} q_{t,i} \).

Emissions accumulate in the atmosphere over both periods, and the damages depend only on the aggregated emissions. For simplicity, we assume away natural decay of emissions, and do not discount damages or abatement costs between the two periods.

We assume that there exists a 'tipping point' in the climate system, such that when cumulated emissions over both periods exceed a threshold, fixed damages of \( X \) are incurred by each country, while damages are zero if the threshold is not reached. We assume away gradual damages (unrelated to the tipping point) to keep the model tractable.\(^5\) The threshold is avoided if cumulated abatement efforts \( Q \equiv Q_1 + Q_2 \) are at least as large as the critical level \( Q^{crit} \), while the threshold is reached and damages incurred if \( Q < Q^{crit} \).

The location of the threshold \( Q^{crit} \) is drawn from a uniform distribution with the support \([\bar{Q} - \Delta, \bar{Q}]\) at the end of the second period, when the cumulated abatement \( Q \) is already fixed.\(^6\) Note, that in case of a known threshold \((\Delta = 0)\), the critical abatement level is simply \( \bar{Q} \).

In this setting, countries play the following 4-stage game. In stage 1, countries simultaneously decide whether to join a coalition or not. We denote the number of signatories by \( k \). In stage 2, the signatories commit to cumulated abatement targets of \( Q^*_1 \) and \( Q^*_2 \) for period 1 resp. period 2. In stage 3, the non-signatories non-cooperatively choose their abatement efforts \( q_{1,i} \) in period 1, and in stage 4, they choose their abatement efforts \( q_{2,i} \) in period 2. Hence, signatories move first, and they commit to abatement efforts over both periods, while non-signatories choose their abatement efforts in the two periods sequentially.

We first analyze the fully cooperative (i.e., \( k = N \)) as well as the non-cooperative (i.e., \( k = 0 \)) benchmark case, before we move on to endogenize the degree of cooperation in the full model.

\(^5\)Neglecting gradual damages is a useful simplification in situations where the tipping point is the main concern. In a political context, the 2°C target may also serve as a 'threshold'.

\(^6\)The assumption of a finite time horizon may reflect the arrival of a radically new technology in the future, which allows countries to avoid the threshold in the future.
2.1 Full cooperation

Due to the convexity of the abatement cost function, under full cooperation each country abates the same amount in each period: \( q_{t,i} = Q/(2N) \). Welfare per country is, thus

\[
\pi_i = -c(Q/(2N))^2 - (\bar{Q} - Q)X/\Delta,
\]

if \( Q \in (\bar{Q} - \Delta, \bar{Q}) \) (interior range). If \( Q \leq \bar{Q} - \Delta \), the damages \( X \) are incurred with certainty, so in this range \( \pi_i = -c(Q/(2N))^2 - X \) is strictly decreasing in \( Q \). If \( Q \geq \bar{Q} \), the threshold is avoided with probability 1, and in this range \( \pi_i = -c(Q/(2N))^2 \) is again strictly decreasing in \( Q \). Hence, there are three possible equilibrium outcomes: an interior solution \( Q \in (\bar{Q} - \Delta, \bar{Q}) \), \( Q = \bar{Q} \) as a corner solution in which the threshold is avoided with certainty, or \( Q = 0 \) (no abatement).

For an interior solution, the necessary first-order condition yields: \(^7\)

\[
Q = \frac{2N^2X}{c\Delta}.
\]

Inserting this result back into the welfare function yields:

\[
\pi_i = \frac{N^2X^2}{c\Delta^2} - \frac{\bar{Q}X}{\Delta}.
\]

An interior solution is attained if the resulting welfare is at least as large as \(-X\), the payoff if countries do not abate any emissions, and if in addition, \( Q \) indeed lies in the interior range. Total abatement under full cooperation is thus given by: \(^8\)

\[
Q^{\text{full coop}} = \begin{cases} 
\bar{Q} & \text{if } \frac{2N^2X}{c\Delta} < \bar{Q} \leq \Delta + \frac{N^2X}{c\Delta}, \\
\frac{2N^2X}{c\Delta} & \text{if } \bar{Q} \leq \Delta + \frac{N^2X}{c\Delta}, \\
0 & \text{otherwise.} 
\end{cases}
\]

Hence, if the threshold \( \bar{Q} \) is sufficiently small (given the damages \( X \)), the grand coalition abates enough to avoid the threshold with certainty. In this case, welfare per country is equal to \( \pi_i = -c\bar{Q}^2/(4N^2) \) (inserting \( Q = \bar{Q} \) into the welfare function). If \( \bar{Q} \) lies in an intermediate range, an interior solution is reached where the coalition avoids the catastrophe with a positive probability smaller than 1. Finally, if \( \bar{Q} \) is sufficiently large, avoiding the catastrophe is too costly, and as we ruled out gradual damages unrelated to the tipping point, the optimal abatement is, then, zero.

\(^7\)It is easy to verify that the second-order condition for a maximum is fulfilled.

\(^8\)Applying the tie-braking rule that the coalition abates when this leads to the same welfare as no abatement: \( \pi_i = -X \).
The intermediate range only exists if \( \frac{2N^2X}{c\Delta} < \Delta + \frac{N^2X}{c\Delta} \). This can be rewritten as:

\[
\Delta > N\sqrt{X/c}.
\]

If the uncertainty parameter \( \Delta \) is smaller than this critical value, the optimal abatement under full cooperation becomes:\(^9\)

\[
Q_{\text{full coop}} = \begin{cases} 
\bar{Q} & \text{if } \bar{Q} \leq 2N\sqrt{X/c}, \\
0 & \text{otherwise}.
\end{cases}
\] (2)

Note, that this solution is independent of \( \Delta \) as long as \( \Delta \) is sufficiently small, and covers also the case without uncertainty (\( \Delta = 0 \)).

### 2.2 No cooperation

The outcome under no cooperation is derived by backwards induction. We first analyze the outcome in period 2, by following similar steps as in the previous subsection. If \( Q = Q_1 + Q_2 \in (\bar{Q} - \Delta, \bar{Q}) \), country \( i \)'s welfare in period 2 (neglecting costs incurred in the previous period, that are now sunk) is:

\[
\pi_{2,i} = -cq_{2,i}^2/2 - (\bar{Q} - Q_1 - Q_2)X/\Delta,
\]

where the total abatement in period 1, \( Q_1 \), is fixed and effectively reduces the value of the threshold parameter \( \bar{Q} \). As under full cooperation, if \( Q \leq \bar{Q} - \Delta \), the damages \( X \) are incurred with certainty, and if \( Q \geq \bar{Q} \), the threshold is avoided with probability 1. Hence, there are three possible outcomes in period 2: an interior solution \( Q_2 \in (\bar{Q} - Q_1 - \Delta, \bar{Q} - Q_1) \), \( Q_2 = \bar{Q} - Q_1 \) as a corner solution, or \( Q_2 = 0 \) (no abatement).\(^{10}\)

The maximization of \( \pi_{2,i} \) over \( q_{2,i} \) yields the following first-order condition: \( cq_{2,i} = X/\Delta \). Hence, total abatement in period 2 is

\[
Q_2 = NX/(c\Delta)
\] (3)

in case of an interior solution. This yields a welfare per country of

\[
\pi_{2,i} = (N - 1/2)X^2/(c\Delta^2) - (\bar{Q} - Q_1)X/\Delta.
\]

Comparing this with the reference level of \(-X\), we can characterize the outcome in the

---

\(^9\)The cutoff-value \( 2N\sqrt{X/c} \) is derived by comparing welfare in the corner solution (\( Q = \bar{Q} \)) with \(-X\).

\(^{10}\)In the corner solution, the individual abatement levels \( q_{2,i} \) are not fully determined. However, in what follows, we focus exclusively on the symmetric equilibrium where \( q_{2,i} = (\bar{Q} - Q_1)/N \). This outcome maximizes total welfare (given \( Q_2 \)), and can be interpreted as the focal point in a coordination game of equilibrium selection.
second period as follows:

\[
Q_{2\text{ no coop}}^\text{no coop} = \begin{cases} 
\bar{Q} - Q_1 & \text{if } \bar{Q} - Q_1 \leq \frac{NX}{c\Delta}, \\
\frac{NX}{c\Delta} & \text{if } \frac{NX}{c\Delta} < \bar{Q} - Q_1 \leq \Delta + \frac{N - 1/2}{c\Delta}, \\
0 & \text{otherwise.}
\end{cases}
\]  

(4)

An interior solution can only exist if \( \frac{NX}{c\Delta} < \Delta + (N - 1/2) \frac{X}{c\Delta} \), which can be rewritten as:

\[ \Delta > \sqrt{X/(2c)}. \]

If \( \Delta \) is smaller than this critical value, the total abatement in period 2 under no cooperation becomes:\(^{11}\)

\[
Q_{2\text{ no coop}}^\text{no coop} = \begin{cases} 
\bar{Q} - Q_1 & \text{if } \bar{Q} - Q_1 \leq N \frac{\sqrt{2X/c}}, \\
0 & \text{otherwise.}
\end{cases}
\]  

(5)

Let us now analyze the outcome in period 1, taking into consideration the resulting outcome in period 2 (for each possible value of \( Q_1 \)). First consider the case where the uncertainty parameter is small (\( \Delta \leq \sqrt{X/(2c)} \)). In this case, an interior solution in period 2 does not exist, so that (similarly as under full cooperation when \( \Delta \leq N \sqrt{X/c} \)) either the threshold is avoided with certainty, or countries do not abate at all (\( Q = 0 \)). Country \( i \)'s welfare function in stage 1 (anticipating costs and benefits in period 2) reads:\(^{12}\)

\[
\pi_i = \begin{cases} 
-\frac{c}{2} q_{1,i}^2 - \frac{c}{2} \left( \frac{Q + Q_1}{N} \right)^2 & \text{if } \bar{Q} \geq Q_1 \geq \bar{Q} - N \sqrt{2X/c}, \\
-\frac{c}{2} q_{1,i}^2 - X & \text{if } Q_1 < \bar{Q} - N \sqrt{2X/c}.
\end{cases}
\]  

(6)

Suppose, there exists an interior solution in period 1 that fulfills \( Q_1 \geq \bar{Q} - N \sqrt{2X/c} \). If it exists, it satisfies the following first-order condition:\(^{13}\)

\[ q_{1,i} = (\bar{Q} - Q_1)/N^2. \]  

(7)

This yields under symmetry:

\[ Q_1 = \bar{Q}/(N + 1), \]

\(^{11}\)As under full cooperation, the cutoff-value is derived by comparing welfare in the corner solution (\( Q = \bar{Q} \)) with \(-X\).

\(^{12}\)The case \( Q_1 > \bar{Q} \) never arises in equilibrium and can be neglected.

\(^{13}\)The second-order condition is easily checked.
and implies a welfare per country of

\[ \pi_i = -\frac{c}{2}\overline{Q}^2 \frac{N^2 + 1}{N^2(N+1)^2}. \]  

(8)

It remains to be checked when the above interior solution in period 1 exists. The welfare derived above is at least as large as \(-X\) if

\[ \overline{Q} \leq \frac{N(N+1)}{\sqrt{N^2+1}} \sqrt{\frac{2X}{c}}. \]

(9)

Furthermore, if this holds, the condition \(Q_1 \geq \overline{Q} - N\sqrt{2X/c}\), which is required for the corner solution in period 2 with full avoidance, is then automatically satisfied.\(^{14}\)

Recall that this outcome entails an interior solution in the first period, and a corner solution in period 2, hence, we will refer to this type of outcome as “C\(_2\)” from now on. Note, that under this outcome, the threshold is always avoided with certainty.

The outcome that is obtained in the absence of any uncertainty coincides with the \(C\(_2\)\) - outcome, because the case \(\Delta = 0\) is included as a special case (note, that \(\Delta\) does not affect the \(C\(_2\)\) - outcome). Since \(Q_1 = \overline{Q}/(N+1)\) and \(Q_2 = \overline{Q} - Q_1\), we observe that for large \(N\), almost the entire abatement efforts are carried out in period 2, while countries abate little in the first period. This is a consequence of the free-rider effect, and implies inefficiently high intertemporal abatement costs (i.e., total abatement costs could be reduced by shifting efforts from one period to the other). Recall that under full cooperation (in the corresponding case where the threshold is avoided with certainty), we found \(Q_1 = Q_2 = \overline{Q}/2\). Cost efficiency, thus, requires a ‘smooth abatement profile’, while in the \(C\(_2\)\) - outcome, this does not arise due to a lack of cooperation.

Now consider the case where the uncertainty parameter is sufficiently large (\(\Delta > \sqrt{X/(2c)}\)), so that an interior solution can emerge in period 2. Any outcome that entails \(0 < Q_1 < \overline{Q} - \Delta - (N - 1/2)\frac{X}{\Delta}\) can immediately be ruled out using our above results for period 2, because \(Q_1\) is, then, so small that the threshold will not be avoided with any positive probability, so (assuming symmetric contributions) all countries are strictly better off with \(Q = 0\) than under any \(Q_1\) in that range. Hence, assuming for the moment that an outcome exists in which the threshold is avoided with positive probability (we identify later under which condition this holds), country \(i\)’s welfare function in stage 1 (anticipating costs and benefits in period 2) reads:

\[ \pi_i = \begin{cases} 
-\frac{c}{2}g_{1,i} - \frac{c}{2}(\frac{Q - Q_1}{N})^2 & \text{if } \overline{Q} \geq Q_1 \geq \overline{Q} - \frac{N\Delta X}{\overline{Q} - \Delta}, \\
-\frac{c}{2}g_{1,i} - \frac{c}{2}(\frac{X}{\Delta})^2 - (\overline{Q} - Q_1 - \frac{N\Delta X}{\overline{Q} - \Delta})\frac{X}{\Delta} & \text{if } Q_1 < \overline{Q} - \frac{N\Delta X}{\overline{Q} - \Delta}.
\end{cases} \]

(10)

\(^{14}\)Using \(Q_1 = \overline{Q}/(N+1)\), this condition yields \(Q \leq (N + 1)\sqrt{2X/c}\), which is a weaker condition than the above as \(N/\sqrt{N^2+1} < 1\) holds for all \(N > 1\).
If an interior solution in the upper range exists (that fulfills \( Q_1 \geq \bar{Q} - \frac{NX}{c\Delta} \)), it coincides with the solution that we characterized above for the case where \( \Delta \leq \sqrt{X/(2c)} \). Hence, it is the \( C_2 \) - outcome. If there exists an interior solution in the lower range (that fulfills \( Q_1 < \bar{Q} - \frac{NX}{c\Delta} \)), it satisfies the following first-order condition:\(^{15}\)

\[
\begin{align*}
q_{1,i} &= \frac{X}{\Delta},
\end{align*}
\]

which implies under symmetry

\[
Q_1 = \frac{NX}{c\Delta}.
\]

Note, that this is the same total abatement that we found also for the second period in case of an interior solution. For the resulting welfare per country, we obtain in this case:

\[
\pi_i = (2N - 1)\frac{X^2}{(c\Delta^2)} - \bar{Q}X/\Delta.
\]

Hence, this outcome entails an *interior solution in both periods*, and we will refer to this as “\( I \) - outcome”. Note, that under this outcome, the threshold is reached with a positive probability (smaller than 1).

Below, we provide a full characterization of the ranges of parameter values for which the different types of outcomes are obtained. It turns out that there exists another type of outcome, that is intermediate between an outcome of type \( C_2 \) and type \( I \). It entails a *corner solution in period 1* and an interior solution in period 2, so we will refer to this outcome as “\( C_1 \)”. Recall that the equilibrium quantity in an interior solution in period 2 is given by \( Q_2 = \frac{NX}{c\Delta} \) (see (3)). The equilibrium quantity under a corner solution in period 1 is then

\[
Q_1 = \bar{Q} - \frac{NX}{c\Delta}.
\]

This corner solution in period 1 implies that the threshold is avoided with certainty (as in the \( C_2 \) - outcome), but in contrast to the \( C_2 \) - outcome the equilibrium quantity in period 2 corresponds to an *interior* solution in that period.

To provide an intuition why this can indeed be an equilibrium outcome, consider a unilateral deviation by some country in period 1. Any deviation to a higher quantity (greater than \( \bar{Q}/N - \frac{X}{c\Delta} \)) is unprofitable, because it induces a corner solution in period 2. But under a corner solution in period 2, any country would have an incentive to reduce its abatement effort in period 1 unilaterally, because the threshold is, then, still avoided with probability 1 but the additional costs are incurred by *all countries* in period 2, not just by the deviating country. Hence, no country has an incentive to deviate unilaterally towards a higher quantity in period 1 in a \( C_1 \) - outcome. Similarly, also a deviation to a lower quantity is not profitable in period 1, because the threshold would

\(^{15}\)The second-order condition is easily checked.
then be reached with a positive probability. But since the quantity per country in period 1 falls short of the quantity under an interior solution in that period, each country would be willing to raise its abatement quantity unilaterally in period 1 in order to reduce the probability of reaching the threshold. For a more formal discussion, see the Appendix (proof of Proposition 1).

Our results for the non-cooperative case are summarized by the following Proposition:

**Proposition 1.** We identify the following types of outcomes under no cooperation:

1. **Outcome C₂ (threshold avoided with certainty):**
   \[ Q_1 = \bar{Q}/(N + 1), \quad Q_2 = \bar{Q} - Q_1; \]
   obtained if \( \bar{Q} \leq \min\left\{ \frac{N(N+1)}{N^2+1} \sqrt{\frac{2X}{c}}, (N + 1)X/(c\Delta) \right\} \)

2. **Outcome C₁ (threshold avoided with certainty):**
   \[ Q_1 = \bar{Q} - NX/(c\Delta), \quad Q_2 = NX/(c\Delta); \]
   obtained if \( \Delta > \frac{\sqrt{X+1}}{N} \sqrt{\frac{X}{c}} \) and \( (N + 1)X/(c\Delta) < \bar{Q} \leq \min\{2NX/(c\Delta), N(\sqrt{2X/c - X^2/(c\Delta)^2} + X/(c\Delta))\} \)

3. **Outcome I (threshold reached with positive probability):**
   \[ Q_1 = Q_2 = NX/(c\Delta); \]
   obtained if \( \Delta > \sqrt{X/c} \) and \( 2NX/(c\Delta) < \bar{Q} \leq \Delta + (2N - 1)X/(c\Delta) \)

4. **In all other cases, the threshold is reached with certainty:** \( Q_1 = Q_2 = 0 \).

These results are most easily understood by visualizing the different types of outcomes, depending on the threshold parameter \( \bar{Q} \). Figure 1 illustrates the different types of equilibrium outcomes for a case where \( \Delta \) is sufficiently large \( (\Delta > \sqrt{X/c}) \) so that all possible types of outcomes occur for a range of \( \bar{Q} \)-values.

On the left-hand side of the figure, the \( C_2 \) - outcome is obtained where countries avoid the threshold with certainty, but postpone most of their abatement efforts until the second period. If \( \bar{Q} \) is raised (in a comparative statics sense), the second range is reached (\( C_1 \) - outcome), where abatement per country in period 2 is \( X/(c\Delta) \), the maximum abatement that an individual country is willing to undertake (per period) in the absence of cooperation when \( \Delta > \sqrt{X/(2c)} \). Although the abatement efforts in period 1 are lower than in period 2, the threshold is still avoided with probability 1. In the third range (\( I \) - outcome), abatement per country is \( X/(c\Delta) \) in each period, and the threshold is reached with a positive probability. Finally, if \( \bar{Q} \) is raised further, then under an \( I \) - outcome, the probability of avoiding the threshold would be so low that countries prefer not to abate at all. Hence, at the critical value of \( \bar{Q} \), the abatement discontinuously drops to zero \( (Q = 0) \).

Let us compare the \( C_2 \) - outcome with the \( C_1 \) - outcome (see Figure 1). In the latter, the ratio \( Q_2/Q_1 \) is smaller than in the former, which implies that the intertemporal inefficiency in abatement is lower. Intuitively, in the \( C_2 \) - outcome, countries (for large
postpone almost their entire abatement efforts until the second period, because each individual country realizes that a unilateral raise in $q_{1,i}$ will amount in a reduction of the other countries’ total abatement efforts in the following period of almost equal size. This is because the outcome in period 2 is a corner solution in which countries ‘fill the gap’ between $\bar{Q}$ and $Q_1$, so that each country abates $q_{2,i} = (\bar{Q} - Q_1)/N$. In other words, the advantage of a raise in $q_{1,i}$ by some individual country $i$ is enjoyed by all countries, so that a pronounced intertemporal free-rider incentive characterizes the $C_2$ - outcome.

The situation is different when $\bar{Q}$ gets larger so that a $C_1$ - outcome is obtained. In this case, the strategic incentive to delay most of the abatement activities until the second period is reduced, because in the second period, an interior solution is obtained. This implies that a unilateral reduction in $q_{1,i}$ will not result in additional abatement efforts by the other countries. Instead, the probability of reaching the threshold then becomes positive. But if $\bar{Q}$ is not too large then even non-cooperative countries have a sufficiently strong incentive to avoid the catastrophe with certainty so that they would raise their abatement efforts in the first period. Hence, the deviation in period 1 is not profitable. This change in the dynamic interaction implies that the $C_1$ - outcome is intertemporally more efficient than the $C_2$ - outcome. In fact, at the boundary towards the $I$ - outcome, located at $\bar{Q} = 2NX/(c\Delta)$ (assuming $\Delta > \sqrt{X/c}$), the $C_1$ - outcome coincides with the fully cooperative outcome, because $Q_1 = Q_2 = \bar{Q}/2$, then, holds in both cases.

Figure 2 illustrates the ranges of parameter values (in the $\bar{Q} - \Delta$ - space) for which the different types of equilibrium outcomes are obtained (for $X = c = 1$ and $N = 3$). The largest value of $\bar{Q}$ for which the threshold is avoided with positive probability under no cooperation (given $\Delta$) is indicated by green color.
2.3 Welfare losses under no cooperation

In the following, we evaluate the welfare losses that result under no cooperation, as compared to the benchmark case of full cooperation. As indicated earlier, under convex costs an intertemporal inefficiency in abatement arises under no cooperation whenever the threshold is (in equilibrium) avoided with certainty (except when $Q = 2NX/(c\Delta)$ holds exactly and $\Delta \geq \sqrt{X/c}$). This inefficiency (reflected by the observation that $Q_1 < Q_2$ while efficiency would require $Q_1 = Q_2$) is especially pronounced when $\bar{Q}$ is small, namely when $\bar{Q} \leq \min \left\{ \frac{N(N+1)}{\sqrt{N+1}} \sqrt{\frac{2X}{c}}, (N+1)X/(c\Delta) \right\}$ ($C_2$ - outcome; see Proposition 1 and Figure 2).

When $\bar{Q}$ is larger, so that (for $\Delta \geq \sqrt{X/c}$) an $I$ - outcome is obtained, the intertemporal inefficiency vanishes, as each country abates the same amount in each period, so that the total abatement $Q = 2NX/(c\Delta)$ is reached at minimal costs. In this range, however, another inefficiency arises that is often even more severe. Namely, while under no cooperation the threshold is reached with a positive probability or even with certainty, under full cooperation it is avoided with certainty for a larger range of parameter values.

Figure 3 compares the ranges of parameter values for which the different types of outcomes are obtained under full and under no cooperation (for $X = c = 1$ and $N = 3$). Thick curves indicate results under full, and thin curves under no cooperation. The largest values of $\bar{Q}$ for which the threshold is avoided with positive probability under full (no) cooperation are indicated by red (green) color.

The figure illustrates that for all values of $\Delta$ except $\Delta = \sqrt{X/c}$, the maximum value
of $\bar{Q}$ where the threshold is avoided with positive probability is strictly larger under full than under no cooperation. The problem is particularly pronounced when

$$\Delta + \frac{(2N - 1)X}{(c\Delta)} < \bar{Q} \leq \min\{2N\sqrt{X/c}, \Delta + \frac{N^2X}{(c\Delta)}\}.$$ 

In this range of parameter values, the threshold is reached with certainty under no cooperation, while it is avoided with certainty under full cooperation. This result crucially depends on the assumption of uncertainty about the location of the threshold. While (in this range of parameter values) a grand coalition behaves exactly like under certainty about the threshold, a strong free-rider incentive makes the stabilizing effect of the threshold collapse under no cooperation, because each country only considers its marginal impact upon the probability of avoiding the threshold, while a grand coalition internalizes the positive externality upon all other countries when some country $i$ raises

---

$^{16}$Interestingly, while this cutoff-level (red curve in Figure 3) is weakly increasing in $\Delta$ under full cooperation, under no cooperation (green curve) it is non-monotonic in $\Delta$. To understand the intuition, consider the no cooperation outcome, and suppose $\Delta$ is raised, starting from an initial value of $\sqrt{X/c}$. Then the cutoff-value of $\bar{Q}$ is first declining because a larger $\Delta$ implies that the marginal incentive to abate (in an interior outcome) is lower, so countries achieve less in an $I$ - outcome, and no avoidance becomes relatively more attractive. However, when $\Delta$ is raised further, the cutoff-value increases again. This is because for larger $\Delta$, the lower boundary of the support $[\bar{Q} - \Delta, \bar{Q}]$ is reduced, so it is cheaper for countries to avoid the threshold with some positive probability. Under full cooperation, only this latter effect exists, so that the cutoff-value of $\bar{Q}$ that separates the partial or full avoidance outcome from the no avoidance outcome does not decline in $\Delta$. 

$^{17}$Note, that $\Delta > \sqrt{X/c}$ has to hold for this range to be non-empty. See Figure 3.
its abatement effort. However, also when \( \Delta \) is smaller (\( \Delta < \sqrt{X/c} \)), full avoidance occurs for larger values of \( \bar{Q} \) under full than under no cooperation. The ratio of cutoff-values of \( \bar{Q} \) where the switch from full to no avoidance occurs is about \( \sqrt{2} \) when \( N \) is large.

Let us summarize the above findings:

**Proposition 2.** There are two sources of inefficiency under no cooperation:

1. **An intertemporal inefficiency in abatement as countries delay most of their abatement efforts until period 2;** this inefficiency arises if \( \bar{Q} \) is sufficiently small, and is particularly pronounced when \( \bar{Q} \leq \min \left\{ \frac{N(N+1)}{\sqrt{N^2+1}} \sqrt{\frac{2X}{c}}, \frac{(N+1)X}{(c\Delta)} \right\} \), which yields a \( C_2 \)-solution under no cooperation;

2. **An inefficiently low probability of avoiding the threshold;** it arises if \( \bar{Q} \) is sufficiently large, and is particularly pronounced when \( \Delta + (2N-1)X/(c\Delta) < \bar{Q} \leq \min \left\{ 2N\sqrt{X/c}, \Delta + N^2X/(c\Delta) \right\} \), which implies full avoidance under full cooperation, while the threshold is reached with certainty under no cooperation.

What is surprising about these results is that the degree of inefficiency of the final outcome under no cooperation is not monotonic in the parameters. Consider again changes in the parameter \( \bar{Q} \). If \( \bar{Q} \) is small, the outcome under no cooperation is intertemporally inefficient because most abatement efforts are carried out in period 2, and too little in period 1. If \( \bar{Q} \) is large, then the outcome under no cooperation is inefficient because it entails too little overall abatement efforts, so that the catastrophe occurs with an inefficiently high probability (compared to the outcome under full cooperation). There is an intermediate value of \( \bar{Q} \) where neither of the two inefficiencies arises so that the outcome under no cooperation coincides with the fully cooperative outcome. This happens at \( \bar{Q} = 2NX/(c\Delta) \).

Figure 4 compares the abatement efforts under full and under no cooperation as a function of the parameter \( \bar{Q} \), for a case with high uncertainty (\( \Delta > \sqrt{X/c} \)). Red curves indicate the results under full, and green curves under no cooperation. The figure illustrates that under full cooperation, the catastrophe is still avoided for parameter values of \( \bar{Q} \) where countries resort to zero abatement efforts under no cooperation. For intermediate values of \( \bar{Q} \), however, the outcome under no cooperation resembles the outcome under full cooperation, and for \( \bar{Q} = 2NX/(c\Delta) \) the outcomes coincide. It is worth noting that the two types of inefficiency that were identified above (and illustrated in Figure 4) are a result of the same underlying problem: countries’ incentives to free-ride on the abatement efforts of other countries. It is interesting to observe that the free-rider incentive can manifest itself in two different ways (intertemporal inefficiency vs. inefficiently low overall abatement efforts). Both inefficiencies can lead to significant welfare losses as compared to the fully cooperative outcome, but they do not occur simultaneously.
2.4 International Environmental Agreements

Before non-signatories choose their abatement efforts in period 1 and 2 (stages 3 and 4 of the IEA-game), signatories in stage 2 commit to their abatement targets for the two periods. From the perspective of the $N-k$ non-signatories, the total commitment $Q^s = Q_1^s + Q_2^s$ of the signatories effectively leads to a reduction of the parameter $\bar{Q}$, and has no other effect because for these countries it is irrelevant how the coalition allocates its total abatement effort over the two periods. Therefore, the strategic situation that non-signatories face in stages 3 and 4 is exactly as in the non-cooperative case characterized in the previous subsection, when $\bar{Q}$ is replaced by $\bar{Q} - Q^s$, and $N$ is replaced by $N-k$ in all formulas. With these two adjustments, we can directly use the results of the previous subsection to fully characterize the outcomes of stages 3 and 4 of the IEA-game. For example, the first statement in Proposition 1, then, yields for a $C_2$ - outcome in stages 3 and 4: $Q_{1}^{n} = (\bar{Q} - Q^s)/(N - k + 1)$, $Q_{2}^{n} = \bar{Q} - Q^s - Q_{1}^{n}$ if $\bar{Q} - Q^s \leq \min\{\frac{(N-k)(N-k+1)}{2} \sqrt{\frac{2X}{c}}, (N - k + 1)X/(c\Delta)\}$. What remains to be done is to characterize the outcome of the first two stages. This is done in the following where we distinguish a number of cases, depending on the parameters $\bar{Q}$ and $\Delta$ (see Figure 2).

Suppose first, that $\bar{Q}$ and $\Delta$ are such that under no cooperation, a $C_2$ or a $C_1$ - outcome is obtained (first two cases in Proposition 1). Hence, the threshold is avoided with probability 1 even under no cooperation, and note that a grand coalition then
also avoids the threshold with certainty. However, unless $\bar{Q} - Q^s = 2(N - k)X/(c\Delta)$ holds exactly and $\Delta \geq \sqrt{X/c}$, there is an intertemporal inefficiency in non-signatories’ abatement efforts (see Proposition 2). Collectively, countries would benefit from shifting part of their abatement efforts into the first period so that a smooth abatement profile is reached. The freerider incentive, however, prevents this under no cooperation. Now suppose, there are $1 \leq k < N$ signatories. Clearly, they allocate their abatement efforts efficiently over the two periods: $Q^1_s = Q^2_s = Q^s/2$. Furthermore, they will choose their total abatement target $Q^s$ at the lowest level that still assures full avoidance of the threshold, anticipating the behavior of the non-signatories. But since the threshold is avoided with certainty even without cooperation, this implies that each signatory, then, abates less than each of the non-signatories. Furthermore, a signatory achieves any given abatement target more cost-effectively than a non-signatory because signatories allocate their efforts efficiently over the two periods. This implies that for this range of parameter values, the payoff of a signatory is higher than that of a non-signatory (when $k < N$), and it follows that in the first stage of the IEA-game, a grand coalition ($k = N$) forms. Note, that although the threshold is avoided also in the absence of cooperation, the formation of the coalition improves welfare because it allows countries to overcome the intertemporal inefficiency in abatement.

Now suppose, $\bar{Q}$ and $\Delta$ are such that under no cooperation, an $I$-outcome is obtained (third case in Proposition 1). This implies that under full cooperation, either the threshold is avoided with certainty, or with a strictly positive probability (see Figure 3). In the latter case, however, each signatory abates more than a country in the non-cooperative case. This implies that for any $k > 1$, the formation of the coalition raises the probability that the threshold is avoided, and the additional abatement costs are incurred only by the signatories, while the non-signatories also enjoy the benefits of the higher total abatement $Q$. Hence, in this range of parameter values, welfare per non-signatory is higher than that of a signatory whenever $k > 1$, while for $k = 1$ the no cooperation outcome is obtained. In this range of parameter values, the model is very similar to one with continuous damages but no threshold. In particular, as long as $Q$ lies in the interior range, the expected marginal benefits of abatement are linear and the abatement costs

---

18See Figure 3.

19And strictly less if $\Delta < \sqrt{X/c}$ or $\bar{Q} - Q^s < 2(N - k)X/(c\Delta)$ and $\Delta \geq \sqrt{X/c}$.

20The abatement per country and period under full cooperation is $NX/(c\Delta)$ in an interior solution, while a non-cooperative country abates $X/(c\Delta)$.

21If the signatories choose $Q^s$ sufficiently large so that non-signatories ‘fill the gap’ by choosing $Q^n = \bar{Q} - Q^s$ and the threshold is avoided with certainty, then the abatement per non-signatory is even lower than $X/(c\Delta)$, whereas each signatory abates more than this amount.

22This is because in an $I$-outcome, non-signatories do not respond to changes in the abatement target by the signatory, so this single signatory has no incentive to raise or lower its abatement target. Furthermore, since non-signatories also achieve intertemporal efficiency in their abatement efforts, there is clearly no advantage in being a first-mover when $k = 1$. 

17
are quadratic. Hence, as shown by Barrett (1994), the maximum coalition size is $k^* = 3$. However, if $\bar{Q}$ is relatively small (close to the range where a $C_1$ - outcome is reached in the non-cooperative case), then the equilibrium coalition size can also be smaller. E.g., suppose that with $k = 2$ the signatories already find it optimal to abate so much that the non-signatories “fill the gap” between $Q^s$ and $\bar{Q}$ so that the catastrophe is avoided with certainty. Then a third country would not enter into this coalition because the coalition would not respond to this by choosing a higher abatement target (as this would only lead to reduced abatement efforts by the remaining non-signatories). Therefore, there is no benefit for the third country that enters the coalition, there are only additional costs (as the coalition members split the abatement efforts equally among them). The equilibrium coalition size is, then, $k^* = 2$. In sum, in the range of parameter values where an $I$ - outcome is reached under no cooperation, the stable coalition size does not exceed three countries.

Next suppose that $\bar{Q}$ and $\Delta$ are such that even under full cooperation, the threshold is reached with certainty (no avoidance – see the red curve in Figure 3). In this case it is clear that there exists no $k < N$ for which a smaller coalition would choose $Q^s$ sufficiently large to induce non-signatories to avoid the threshold with positive probability. In this case, $Q = 0$ holds irrespective of $k$, and equilibrium participation is indeterminate.

Finally consider the ranges of parameter values for which the threshold is avoided with strictly positive probability under full, but with zero probability under no cooperation (area between the green and the red curve in Figure 3). Since avoidance is optimal under full cooperation, there clearly exists a minimum value of $k \leq N$ so that signatories in stage 2 are willing to choose $Q^s > 0$, which is sufficiently large to assure that also non-signatories become active and full or partial avoidance is reached as equilibrium outcome. Because the no cooperation outcome entails zero abatement, and non-signatories are not willing to contribute more than $X/(c\Delta)$ (per period and country) to total abatement, it is clear that each signatory has to contribute more than a non-signatory in equilibrium. Therefore, a corner solution is reached at the participation stage, where $k$ is just large enough to induce signatories to choose a positive $Q^s$. Hence, the equilibrium value of $k$ is the smallest integer that yields a welfare of at least $-X$ to each signatory. Formally, by using $Q^n = 2(N - k)X/(c\Delta)$ and $Q^s = 2k^2 X/(c\Delta)$ in the welfare function of a signatory, and equalizing this payoff with $-X$, we obtain the following critical value of $k$:\(^{23}\)

\[
k^{\text{crit}} = 1 + \sqrt{(\bar{Q} - \Delta)c\Delta/X - 2N + 1}.
\]

\(^{23}\)This formula is valid if $Q^n + Q^s \leq \bar{Q}$ holds for the given value of $k$. However, the lowest value of $k$ that induces the signatories to choose a positive abatement effort may be sufficiently large so that a corner solution with full avoidance is reached. In this case $Q^n = 2(N - k)X/(c\Delta)$ continues to hold because the signatories then choose the smallest $Q^s$ that still yields full avoidance, but $Q^s$ is smaller than $2k^2 X/(c\Delta)$. A modified expression for $k^{\text{crit}}$ is then obtained (not shown). Otherwise, this does not change the results.
Equilibrium participation is the smallest integer at least as large as $k^{crit}$.\footnote{This is due to the integer constraint on $k$.}

Although signatories just break even (welfare per signatory is approximately $-X$), nevertheless we cannot conclude that cooperation does not achieve much in this type of equilibrium. This is because non-signatories are strictly better off than the signatories, so that the \textit{global} welfare gains (as compared to the non-cooperative outcome that entails $Q = 0$) may still be substantial.

Consider the following numerical example. Let $X = c = 1$, $N = 10$, $\bar{Q} = 18$, and $\Delta = 13$. It is straight-forward to verify that for these parameter values, under no cooperation no avoidance is obtained ($Q = 0$), whereas under full cooperation, partial avoidance ($0 < Q^{full\ coop} < \bar{Q}$) is reached. Using the above formula we find $k^{crit} = 7.78$, so that equilibrium participation is given by $k = 8$. We, then, find the following results: $Q^n = 0.31$, $Q^s = 9.85$, $\pi^n_i = -0.63$, and $\pi^s_i = -0.98$. Hence, signatories break even and achieve a welfare only slightly larger than $-X$, whereas non-signatories are significantly better off than under no avoidance (which yields $\pi_i = -1$). In equilibrium, the threshold is avoided with a probability of about $2/3$.

The following proposition summarizes the above findings:

\textbf{Proposition 3.} If $\bar{Q}$ and $\Delta$ are in the following ranges:

1. (full avoidance under no cooperation)
   \begin{enumerate}
   \item $(a)$ $\Delta > \sqrt{X/c}$ and $\bar{Q} \leq 2NX/(c\Delta)$, or
   \item $(b)$ $\frac{\sqrt{N^2+1}}{N} \sqrt{\frac{X}{c\Delta}} \leq \Delta \leq \sqrt{X/c}$ and $\bar{Q} \leq N(\sqrt{2X/c - X^2/(c\Delta)^2} + X/(c\Delta))$, or
   \item $(c)$ $\Delta < \frac{\sqrt{N^2+1}}{N} \sqrt{\frac{X}{c\Delta}}$ and $\bar{Q} \leq \frac{N(N+1)}{N+1} \sqrt{\frac{X}{c\Delta}}$,
   \end{enumerate}
   then a grand coalition forms ($k = N$) in the IEA-game and the threshold is avoided with certainty;

2. $\Delta > \sqrt{X/c}$ and $2NX/(c\Delta) < \bar{Q} \leq \Delta + (2N - 1)X/(c\Delta)$ (I - solution under no cooperation), then the stable coalition size is not larger than $k = 3$. The threshold is reached with positive probability;

3. (no avoidance under full cooperation)
   \begin{enumerate}
   \item $(a)$ $\Delta \leq N\sqrt{X/c}$ and $\bar{Q} > 2N\sqrt{X/c}$, or
   \item $(b)$ $\Delta > N\sqrt{X/c}$ and $\bar{Q} > \Delta + N^2X/(c\Delta)$,
   \end{enumerate}
   then $k$ is indeterminate and the threshold is reached with certainty ($Q = 0$);
4. for all other parameter values (no avoidance under no cooperation, and partial or full avoidance under full cooperation) the (maximum) stable coalition size $k^*$ is just large enough so that signatories become ‘active’ and choose a positive abatement effort; the threshold is avoided with a positive probability.

3 Experiment

In order to check the robustness of the main predictions of the theory that was presented in the previous section of this paper, we conducted a small experiment. Most importantly, we wanted to investigate whether the intertemporal free-rider incentive can be observed, that is, whether non-cooperating players postpone their abatement efforts “until the last minute”. Furthermore, we wanted to investigate whether cooperation plays a significant role in a dynamic experimental setup with a threshold for discontinuous climate damages.

3.1 Experimental setup

We conducted a paper and pencil experiment with 18 Bachelor students from various fields of study. Four different variants of a game were played during the experiment. Each variant of the game was played ten times, with an additional mock game in the beginning to familiarize students with each new variant. In each game, the 18 participants were divided into three different groups of six players (A, B, and C). Hence, each of the 40 games was played simultaneously by three separate groups of players. The assignment of a participant to group A, B, or C changed in each game, in a pre-determined way that made it impossible for any participant to infer who the other five players in his or her group were.

Our experimental setting was inspired by Milinski et al. (2008), who also consider a game with a potential catastrophe, where each individual game is played by six participants. However, we introduce a number of significant modifications. First of all, in our experiment a game lasts only for two periods, whereas in Milinski et al. (2008) each game has ten periods. Apart from the fact that this makes the experimental results more easy to compare with our theoretical findings, this approach has the advantage that players cannot build up a reputation (e.g., for being non-cooperative). This allows us to isolate the intertemporal free-rider incentive, as it is not confounded with other incentives.

Another significant modification is that we gave players the possibility to form a climate coalition in two variants of our game. This possibility was embedded as follows in our experiment. In each game with the possibility to cooperate, the six participants of a group could initially decide (individually) whether to join a coalition or not. The

\[25\] The four mock games were not payoff-relevant, and the results in these games will be neglected in the following.
total number of signatories was then announced to the group, without revealing the identity of other coalition members or outsiders. In the next step, each coalition member could propose contributions for all signatories in his or her group for both periods, under the restriction that all signatories (including the proposer) have to implement identical contributions. Hence, a coalition member could not shift efforts towards other coalition members by proposing a lower contribution for herself than for the others. In each group, only the proposed contributions of one of the signatories became binding for all signatories; the identity of this signatory was randomly determined (in each group). In theory, each of the signatories’ proposals in a group should then coincide with the decision of a social planner who determines the contributions of all coalition members on their behalf, with the goal to maximize their joint welfare. Hence, also the incentives of players to join the coalition in the first place should coincide with those in a setup where coalition members decide collectively about their contributions. Our approach is thus suitable to analyze coalition formation in a simple experimental setup, without any need for coalition members to negotiate or to discuss about their collective decision making. An alternative approach is suggested by Dannenberg, Lange, and Sturm (2014). Under this approach, the proposal with the lowest contribution becomes binding for all coalition members.

In each game, each player had a (virtual) endowment of 12 Euros. The (virtual) payoff in a game is the endowment minus the sum of abatement costs in both periods, in case there is no catastrophe. In each period, each player could contribute 0, 1, or 2 units to the overall abatement in his or her group. The corresponding abatement cost in a period is 0, 2, and 5 Euros, respectively. Whenever a catastrophe occurs, the (virtual) payoff of all members of the group is zero. The actual payoff of a player was determined at the end of the experiment, by selecting at random three out of the 40 payoff-relevant games in which each player participated (at most one game from each variant). For each of these three games, the payoffs computed in this way were payed out in Euros.

The first variant of the game was characterized by a known threshold: if the sum of all contributions of the six players in a group over both periods was 10 or more, then no catastrophe occurred. Otherwise, a catastrophe occurred, and the (virtual) payoffs of all group members were zero. Players first determined their contributions for the first period. Then the total contribution was announced in each group, and contributions for the second period were determined. Finally the total contributions of all players in a group were announced, and whether a catastrophe occurred or not in this group.

The second variant of the game was also characterized by a known threshold. It differed from the first variant only in that players were allowed to form a climate coalition. The coalition members could determine their contributions for the two periods first (before the non-signatories), in the manner described earlier. Then the total contribution of all signatories over both periods was announced to the remaining non-signatories in each
group, and for these players the game proceeded in the same way as in the first variant.

The third variant of the game differed from the first variant only in that the location of the threshold was uncertain (unlike in the second variant, players had no possibility to cooperate). If the sum of all contributions in a group over both periods was 12 or more, then no catastrophe occurred. If this number was 11, then the probability of a catastrophe was 1/4. If the contributions summed up to 10 then this probability was 1/2, and if they summed up to 9 the probability was 3/4. If the total contributions were even lower then a catastrophe occurred for sure.\(^{26}\)

The fourth variant of the game differed from the third only in that players again had a possibility to cooperate, as in the second variant.

Before we summarize the main results of the experiment, let us formulate hypotheses about the likely behavior of the participants, that we would expect according to the theory presented earlier in this paper:

**H1:** The average contribution of non-cooperating players is smaller in the first than in the second period.

**H2:** This effect is weaker when the location of the threshold is uncertain.

**H3:** When players can form coalitions, then cooperation plays a significant role.

The first hypothesis reflects the presence of the intertemporal free-rider incentive (see Propositions 1 and 2). This effect should be particularly pronounced when the location of the threshold is (almost) certain, and less pronounced (or absent) when there is uncertainty about this location (see Figures 1 and 2, in particular the transition from the \(C_2\) to the \(C_1\) and then the \(I\) - outcome when \(\Delta\) increases); this is reflected by Hypothesis 2. According to Proposition 3, cooperation should play a significant role for various parameter constellations. In particular, if the uncertainty about the location of the threshold is sufficiently small, then according to the theory a grand coalition should form. This is because the coalition sets its overall abatement target just large enough so that the threshold is still avoided with certainty, anticipating that any remaining non-signatories will fill the gap towards the level of abatement that achieves this goal (\(C_1\) or \(C_2\) - outcome). In the experiment, the degree of uncertainty in the variants of the game with uncertainty (variants 3 and 4) was sufficiently low so that non-signatories still had this incentive (although they were close to being indifferent). Hence, in both variants with a possibility to cooperate (variants 2 and 4), we would expect that large coalitions form; this is the contents of the third hypothesis.

\(^{26}\)All probabilities in the experiment were implemented with the help of dices.
3.2 Results

Figure 5 shows the average contributions per player in each period for all 40 games. The averages are taken over the contributions of the players in the three groups (groups A, B, C) in each game.

Let us first discuss the two variants of the game without a possibility to cooperate (variants 1 and 3). The results illustrated in Figure 5 confirm Hypothesis 1: we observe that in variants 1 and 3, in all games the average contribution of players is (weakly) larger in the second than in the first period (there is only one game where the average contributions in both periods are identical: game 2 in variant 3).\footnote{We conducted a Wilcoxon signed-rank test to verify that in each of these two variants, the differences between the contributions in period 1 and 2 are significant. For variant 1, we found a z-value of 30.6, and for variant 2 this value was 14.9. In both cases, the zero hypothesis (that contributions are on average equal in both periods) can be rejected on any desired level of significance (the p-value is below $10^{-10}$ in both cases).} Hence, our experimental results support the theoretical prediction of an intertemporal free-rider incentive: non-cooperating countries tend to postpone their abatement efforts “until the last minute”, because each of them realizes that by lowering its contribution early on, it can shift part of its abatement costs towards the other countries.\footnote{The effect occurred also for the non-signatories in the variants with a possibility to cooperate. In variant 2, non-signatories contributed on average 0.54 in period 1 and 1.03 in period 2. In variant 4, their average contributions were 0.66 resp. 0.68 in period 1 resp. 2.}

Also Hypothesis 2 is confirmed by the experimental evidence. In particular, the
difference between the contributions in period 1 and 2 is on average larger in variant 1 (without uncertainty about the location of the threshold), as compared with variant 3 where the location of the threshold is uncertain. Intuitively, if the location of the threshold is uncertain, then the incentives of any individual player to raise her contribution in the second period to two are lower, because this lowers the probability of a catastrophe by at most 1/4. Anticipating that the other players will be less inclined to “fill the gap” in the second period, an individual player in the first period may thus shy away from playing the “risky strategy” of contributing nothing in the first period.

Also note, that in each of these two variants of the game, there is no clear trend in the contributions to be observed across the 10 games. Hence, both the prevalence of intertemporal free-riding, and the observation that this effect is stronger in the absence of uncertainty, do not seem to vanish when participants gain more experience in playing the game. Also the occurrence of catastrophes shows no clear trend in the variants of the game without cooperation. This is illustrated in Figure 6, that shows the number of catastrophes in each game (red bars), along with the average total contribution of all players over both periods (blue bars). The average is again taken over the contributions in the three groups (groups A,B,C) in each game.

Let us now discuss the results for the two variants of the game where players had a possibility to cooperate (variants 2 and 4). Figure 7 shows the average coalition sizes in these games (the averages are taken over the coalition sizes of the three groups in each
We observe that coalition sizes are on average larger in variant 4 (with uncertainty about the location of the threshold) than in variant 2 (without uncertainty).

![Figure 7: Average coalition size (averaged over the three groups)](image)

Intuitively, what may have caused the larger participation levels in the case with uncertainty is that players could not be sure to be able to shift their costs to other players when lowering their own contribution in the first period. This makes the strategy of contributing one in each period relatively more attractive. But then players may as well join the coalition, that usually prescribes such contributions for all of its members. Furthermore, while cooperation usually could not eliminate the intertemporal free-riding problem among the non-signatories in variant 2 (see Figure 5, top right), average contributions were usually as high in period 1 as in period 2 in variant 4 (Figure 5, bottom right). Hence, uncertainty even acted as a driver of cooperation, and not an obstacle.

Nevertheless, comparing the variants with a possibility to cooperate (variants 2 and 4) with the corresponding variants without cooperation (variants 1 resp. 3), we observe that cooperation did not significantly lower the frequency of catastrophes (see Figure 6). It did, however, reduce the strength of the intertemporal free-rider problem. To see this, compare the difference in the average contributions in period 1 and period 2 in variant 2 with variant 1 (Figure 5, top), as well as those in variant 4 with variant 3 (Figure 5, bottom). We observe that both in the case without uncertainty, and in the case with uncertainty about the location of the threshold, cooperation reduced the degree of intertemporal free-riding. This explains why the average (virtual) payoffs of players were larger in these variants of the game than in the ones without a possibility to cooperate. However, although cooperation could reduce the overall abatement costs by allocating the abatement efforts more efficiently across periods, it did not have a sizable impact upon the frequency of catastrophes that occurred in the experiment. Only

---

29On average, signatories achieved higher (virtual) payoffs than the non-signatories. In variant 2, this value was 6.76 for signatories and 5.82 for non-signatories. In variant 4, signatories achieved on average 6.33 and non-signatories 5.94.

30The average (virtual) payoff per game and player in variant 2 (6.24) was slightly larger than in variant 1 (6.15). This difference is more pronounced between variant 4 (6.22) and variant 3 (5.88). Note, that in variant 4, only 7 catastrophes occurred while there were 8 in variant 2.
in variant 4 (with uncertainty and with a possibility to cooperate), a weak tendency towards more cooperation can be observed across the 10 games, and as a result, the frequency of catastrophes declined.

4 Conclusion

There is little hope that mother nature alone will provide a solution for the global warming problem, that the international community has so far failed to address adequately. Neither a scarcity of fossil fuels, nor the threat of catastrophic climate change will discipline countries that fail to cooperate in their efforts to reduce greenhouse gas emissions. Tipping points, however, have the potential to improve the prospects of cooperation.

This paper shows that even if the location of a threshold in the climate system is not precisely known, cooperation can play an important role. In particular, in a dynamic context there are additional welfare gains from cooperation. Non-cooperative countries tend to postpone their abatement efforts “until the last minute”. By contrast, if signatories can commit to their future abatement efforts early on, they will allocate them efficiently over time. As a result, in many cases a stable coalition forms that achieves significant welfare gains.

Appendix

Proof of Proposition 1. Consider first the case where an interior solution exists in period 1 that fulfills \( Q_1 \geq \bar{Q} - \frac{NX}{\Delta} \), which implies a corner solution in period 2. For this case we derived \( Q_1 = \bar{Q}/(N+1) \), so that this solution can only exist if

\[
\bar{Q} \leq (N+1)X/(c\Delta).
\]

Furthermore, welfare per country must be at least as large as \(-X\), for otherwise, countries switch to \( Q = 0 \). As in the case where \( \Delta \leq \sqrt{X/(2c)} \), this leads to condition (9). This implies that this type of solution exists iff

\[
\bar{Q} \leq \min\left\{ \frac{N(N+1)}{\sqrt{N^2+1}} \sqrt{\frac{2X}{c}}, (N+1)X/(c\Delta) \right\}. \tag{13}
\]

Equalizing both cutoff-values in the min-operator yields

\[
\Delta = \frac{\sqrt{N^2+1}}{N} \sqrt{\frac{X}{2c}}, \tag{14}
\]

which can easily be shown to lie between \( \sqrt{X/(2c)} \) and \( \sqrt{X/c} \) (for later reference).
Now consider the case where an interior solution exists in period 1 that fulfills $Q_1 < \bar{Q} - \frac{NX}{\Delta}$, which implies an interior solution also in period 2. For this case we found that $Q_1 = NX/(c\Delta)$, which implies that this type of solution can only exist if

$$\bar{Q} > 2NX/(c\Delta).$$

Again, we must check when this solution yields a welfare at least as large as $-X$. Using (12), this leads to the condition

$$\bar{Q} \leq \Delta + (2N - 1)X/(c\Delta).$$

Combining both conditions, we find that this type of solution exists iff

$$2NX/(c\Delta) < \bar{Q} \leq \Delta + (2N - 1)X/(c\Delta). \quad (15)$$

This interval for $\bar{Q}$ is non-empty if

$$\Delta > \sqrt{X/c}.$$

Let us now analyze the possible outcomes outside the ranges of parameter values (see (13) and (15)) where a $C_2$ - outcome or an $I$ - outcome is obtained. Consider the following interval for $\bar{Q}$:

$$(N + 1)X/(c\Delta) < \bar{Q} \leq 2NX/(c\Delta).$$

Clearly, in this range neither a $C_2$ - outcome nor an $I$ - outcome is obtained. We now show that in this range of parameter values, either $Q = 0$ obtains, or a $C_1$ - outcome is reached that entails a corner solution in the first and an interior solution in the second period (see the main text).

To see that this is an equilibrium outcome (provided that the resulting welfare per country is at least as large as $-X$), note that if country $i$ deviates to a higher quantity in period 1 then under a corner solution in period 2, its best choice of $q_{1,i}$ is determined by (7), which leads to

$$q_{1,i} = (\bar{Q} - (N - 1)Q_1^* / N) / (N^2 + 1)$$

if the other countries stick to their equilibrium strategy $q_{1,j \neq i} = Q_1^*/N$ where $Q_1^* = \bar{Q} - NX/(c\Delta)$. After rearranging, the condition $q_{1,i} < Q_1^*/N$, then, yields $\bar{Q} > (N+1)X/(c\Delta)$, which is fulfilled by (13) if (14) holds. By continuity, country $i$'s welfare, thus, decreases in $q_{1,i}$ for all $q_{1,i} > Q_1^*/N$. The deviation is, thus, unprofitable. If (14) fails to hold,

\footnote{Note, that this interval is non-empty for any $N \geq 3$.}
we have already shown that there is a direct transition from the $C_2$ - outcome towards $Q = 0$. 

Now consider a deviation to a lower $q_{1,i}$ in period 1. Such deviation is also unprofitable. To see this, note that country $i$’s best choice of $q_{1,i}$ is given by (11) under an interior solution in period 2. But this is larger than the equilibrium value $Q_1^*/N$ if (after rearranging) $\bar{Q} < 2N X/(c\Delta)$, which holds by (15) since we are outside the range (lower values of $\bar{Q}$) where the $I$ - outcome exists. Hence, the deviation is unprofitable.

To fully characterize the range of parameter values where the $C_1$ - outcome exists, we must check when this solution delivers a welfare per country of at least $-X$. The equilibrium welfare in this case is given by

$$\pi_i = -c(\bar{Q}/N - X/(c\Delta) )^2/2 - c(X/(c\Delta))^2/2.$$ 

This is at least as large as $-X$ if

$$\bar{Q} \leq N(\sqrt{2X/c - X^2/(c\Delta)^2} + X/(c\Delta)).$$

References


