Global Climate Governance in the Light of Geoengineering: A Shot in the Dark?*

Michael Finus *Department of Economics, Karl-Franzens-Universität Graz, e-mail: michael.finus@uni-graz.at*

Francesco Furini *Department of Socioeconomics, Universität Hamburg, e-mail: francesco.furini@uni-hamburg.de*

Abstract

Solar radiation management (SRM), as one form of geoengineering, has been proposed as a last exit strategy to address global warming. Even though SRM is expected to be a cheap, it may be risky and associated with high collateral damages. We analyze how SRM affects equilibrium mitigation strategies, the governance architecture of a climate agreement and whether and how signatories to a climate agreement can avoid that non-signatories deploy SRM. We show under which conditions the threat to deploy geoengineering can stabilize a large climate agreement. Results are derived in a cartel formation game and all qualitative conclusions are confirmed in a repeated game framework.

Keywords: mitigation-geoengineering game; solar radiation management; collateral damages; climate agreements.

JEL-Code: D71, D74, H41, Q54

* This paper is a completely revised and new version of an older version of the paper with Pamina Tomka.

1. Introduction

Addressing global warming is one of the major challenges of our society. Immediate and substantial actions are needed in order to keep the temperature increase within 1.5 to 2 degrees Celsius compared to pre-industrial levels such as to avoid the most serious impacts on the natural, economic and social system (IPCC 2018, 2021). However, mitigation (emission reduction) is a public good and because there is no supranational institution which could enforce targets, not much progress has been made to address global warming. Since the first consequences are already visible, adaptation has become part of the portfolio of climate change measures. Adaptation does not address the root of the problem, but reduces the impacts of global warming, as for instance building dykes against flooding, developing heat resistant crops, irrigation of farmland and installation of air conditioning to cope with the heat.[1] Given the danger of abrupt climate damages due to tipping points of environmental and climate systems, geoengineering has been proposed as a third option in recent years.

Two main categories of geoengineering can be identified: carbon dioxide removal (CDR) and solar radiation management (SRM) (Klepper and Rickels 2014, The Royal Society and Sheperd 2009). CDR extracts CO_2 from the air and deposits it in biomass or underground. It aims at slowing, or even reversing, the atmospheric accumulation of greenhouse gases (Meadowcroft 2013). Hence, CDR is similar to mitigation: it reduces the carbon concentration and needs a shared effort to produce tangible results. CDR is often seen as a complementary measure to achieve the net emission zero goal (IPCC 2018, Fuss et al. 2020). However, most CDR measures are currently too expensive and those which appear feasible, such as reforestation, afforestation and bioenergy, can only account for modest amounts of carbon sequestration (Sandler 2018).

[1] Strategic interaction between mitigation and adaptation in the context of agreement formation has been analyzed by Bayramoglu et al. (2018), Borrero and Rubio (2022), Finus et al. (2021) and Lazkano et al. (2016).

In contrast, SRM can be seen as a deliberate modification of the Earth's climate system in order to cool down the planet. The most frequently discussed proposal is stratospheric sulfur aerosols injections to reduce the incoming solar radiation. Alternative proposals include marine cloud brightening, which increases the reflectivity of clouds, and the use of reflective particles to increase the longevity of sea ice reflecting back solar radiation (Caldeira et al. 2013, IPCC 2014, National Academies of Sciences, Engineering and Medicine 2021). On the one hand, SRM offers a quick fix and may be seen as a last resort to solve the climate change problem at relatively low cost. On the other hand, it also comes with uncertain risks that cannot be underestimated (Barrett 2014, Bodansky 2013, Stephens et al. 2021). Therefore, SRM poses several governance challenges (Reynolds 2019). SRM does not address the underlying problem of high atmospheric carbon concentration and, by artificially modifying the climate, it may generate unpredictable environmental side effects (Parker and Irvine 2018).

In this paper, when we talk about geoengineering, we refer to solar radiation management. We are interested in the interplay between mitigation and geoengineering measures in a strategic setting of climate agreement formation. Does the possibility of deploying geoengineering reduce mitigation efforts? How does it affect the governance architecture of climate agreements and what are the prospects of avoiding the unilateral deployment of geoengineering?

Weitzman (2015) coined the term 'gob' (good or bad) to describe the nature of geoengineering. On the one hand, geoengineering produces a positive externality in the form of a reduction of climate change damages; on the other hand, it may generate a negative externality in the form of unexpected collateral damages. Additionally, the relatively cheap price of geoengineering and its potentially large benefits could lead individual countries to act unilaterally in order to solve the climate change problem (Barrett 2008, Blackstock and Long 2010). This 'free-driving' behaviour is difficult if not impossible to prevent.

Weitzman (2015) argues that free-driving can be successfully managed if countries would agree to take decisions about the level of geoengineering by a super majority rule. However, he does not provide answers as to whether and how countries would agree on such a governance rule and what could be done against those countries which do not join the majority.

Ricke et al. (2013) model the governance of geoengineering by considering a coalition formation game on geoengineering. They assume that the coalition agrees on the level of geoengineering by maximizing the aggregate welfare of its members. Countries which are not members of the agreement are assumed not to deploy geoengineering. The authors show that all countries will have an incentive to join the coalition, as this is the only way to participate in geoengineering decisions. However, the assumption that non-members can be excluded and prevented from taking decisions on geoengineering on their own is questionable, as it defines away the free-driving problem. Moreover, as Weitzman (2015), Ricke et al. (2013) also do not capture the interaction between geoengineering and other climate change measures.

This is different in Heyen et al. (2019), even though we would argue that the scenarios analyzed in this paper are a bit far-fetched, a mix of science fiction and war games. The authors consider the effects of counter-geoengineering, i.e., measures capable of negating the climate effects of geoengineering, on the free-driving problem. They find that counter-geoengineering prevents the free-driver outcome and leads to either a cooperative deployment of geoengineering or a non-cooperative escalation of opposing geoengineering interventions.

In this paper, we analyze the more obvious connection between geoengineering and mitigation. At first sight, one may expect that the availability of geoengineering would reduce countries' efforts to cut emissions. However, due to the possibility of high collateral damages, countries could actually increase their mitigation efforts in order to reduce or avoid the use of geoengineering (Moreno-Cruz 2015, Urpelainen 2012). This could also affect the outcome of climate agreements codifying the obligations of greenhouse gas emission reduction.

Fabre and Wagner (2020) show in a weakest-link game that more ambitious climate targets could be achieved in the presence of geoengineering. However, the authors do not answer the question why emission reductions can be viewed as a weakest-link game where it is typically categorized as a summation game (Barrett 2007, Sandler 2004).

Millard-Ball (2012) assumes for mitigation a summation technology and that the marginal benefits from geoengineering are reduced by mitigation. That is, mitigation and geoengineering are strategic substitutes. He considers a four-stage coalition formation game in which countries choose first their membership, then signatories and afterwards non-signatories choose their mitigation levels and finally geoengineering decisions are taken by all countries simultaneously. The analysis focuses on the stability of the grand coalition (i.e., all countries are signatories). Two main results may be highlighted. First, if collateral damages from geoengineering are perceived to be sufficiently high, countries will have an incentive to increase total mitigation up to a level at which the deployment of geoengineering is no longer attractive. Second, sufficiently high collateral damages can also stabilize the grand coalition if leaving an agreement triggers the deployment of geoengineering.

In this paper, we build on the model by Millard-Ball (2012) because of its analytical simplicity and provide four extensions. First, we offer a comprehensive analysis of the entire parameter space of the model. This is important for understanding and evaluating the economic and ecological relevance of different policy scenarios. Second, we extend the analysis beyond the grand coalition. Third, we correct several conceptual flaws in Millard-Ball's analysis, which lead to wrong conclusions. Even though we confirm that sufficiently large collateral damages are required to make a climate agreement robust to deviations, we also show that collateral damages cannot be too high, as otherwise the threat of punishment of deploying geoengineering would not be credible. Fourth, we derive our results in two conceptual frameworks frequently used in the literature on the game-theoretic analysis of international environmental agreements

(IEAs): a cartel formation game and a repeated game.[2] This is to test the robustness of qualitative conclusions. The main message of our paper is that an agreement on mitigation with a large participation can be self-enforcing in the light of the possibility of the deployment of geoengineering. However, it is not so simple as Millard-Ball (2012) suggests. Countries do not only need to perceive collateral damages to be sufficiently high (lower bound of collateral damages) to be deterrent, they also need to evaluate those collateral damages not too high (upper bound of collateral damages), as otherwise threats would not be credible.

In what follows, section 2 introduces the model, including three different equilibrium scenarios and six policy scenarios. Section 3 analyzes coalition stability in the cartel formation game and section 4 does this in a repeated framework. Section 5 concludes.

2 Model

2.1 Payoff Functions

We consider n symmetric countries, $i = 1, 2, ..., n$, with $n > 3$.[3] Following Millard-Ball (2012), each country can choose its individual mitigation level $q_i \geq 0$ along a continuous interval $q_i \in \left[0, q_i^{max}\right]$ with q_i^{max} some upper bound such that emissions are zero. The decision about the deployment of geoengineering z_i is a binary variable. If $z_i = 0$ geoengineering is not deployed whereas if $z_i = 1$ it is deployed. As it will become clear below, since countries are symmetric either all countries find it attractive to deploy geoengineering or none does. The individual payoff function is given by

---

2 The cartel formation game is part of the group of membership games and the repeated game is part of the group of compliance games. They have been the workhorse models in the literature on IEAs. See Finus and Caparros (2015) and Marrouch and Chaudhuri (2015) for an overview.

3 An externality game requires $n \geq 2$. We assume $n > 3$ to be able to present all subsequent results in a concise manner. Generally, it seems reasonable to assume a sufficiently large number of countries; there are roughly 200 countries world-wide of which for instance 197 countries have ratified the United Nations Framework Convention on Climate Change (UNFCCC).

$$\pi_i = bQ - \frac{c}{2}q_i^2 + \frac{1}{n}(g - Q)\cdot z_i - \frac{(n-1)}{n}d \cdot z_i \ . \qquad (1)$$

The first two elements in equation (1) constitute the mitigation part of the payoff function. bQ are the linear benefits arising from total mitigation $Q = \sum_{i=1}^{n} q_i$ and $\frac{c}{2}q_i^2$ are the quadratic costs from individual mitigation q_i, with b and c being strictly positive parameters.

The last two elements in equation (1) constitute the geoengineering part of the payoff function provided $z_i = 1$ (as otherwise they are zero). Millard-Ball (2012) assumes that geoengineering is conducted by a randomly selected country with probability $1/n$. Hence, country i enjoys $1/n$ of expected net benefits $\frac{1}{n}(g - Q)$ with g being a strictly positive parameter. That is, net benefits already include deployment costs, which may be so low that they can be approximately neglected. They also include possible collateral damages to country i. The marginal net benefits of geoengineering decrease in total mitigation Q. That is, mitigation and geoengineering are strategic substitutes.

Geoengineering produces collateral damages to all $n-1$ other countries, except to the country which deploys geoengineering. Hence, expected collateral damages are $\frac{(n-1)}{n}d$ with d being a strictly positive damage parameter.[4]

2.2 Three-Stage Game

The entire coalition formation game consists of three stages. The first stage is the membership stage. Out of n countries, k countries, $1 \le k \le n$, will join the climate coalition K. These k countries are called signatories (S), while the $n-k$ countries which remain outside the climate

[4] In section 3.3, we show that giving up the assumption of a randomly selected country and allowing all countries to deploy geoengineering does not change qualitative conclusions at all.

agreement are called non-signatories (NS). The second stage is the mitigation stage.[5] Signatories choose their mitigation levels by maximizing the aggregate welfare across all signatories, while non-signatories choose their mitigation levels by maximizing their individual welfare. The third stage is the geoengineering stage. Countries choose whether to deploy geoengineering. The game is solved by backward induction.

The size of stable coalitions is determined by applying the concept of internal and external stability, frequently applied in the game-theoretic literature on IEAs (see, e.g., Finus and Caparros 2015 and Marrouch and Chaudhuri 2015 for an overview and the literature cited there).

$$\pi_{i\in K}\left(k\right) \geq \pi_{i\notin K}\left(k-1\right) \tag{2}$$

$$\pi_{j\notin K}\left(k\right) \geq \pi_{j\in K}\left(k+1\right) \tag{3}$$

The notation makes already use of the symmetry assumption in that the payoff will only depend on the size k of coalition K and whether a country is a member of K or remains outside. Internal stability (2) implies that no signatory wants to leave the coalition and external stability (3) implies that no non-signatory wants to join the coalition. We denote the stable coalition of size k by k^*. If there is more than one stable coalition, we apply the Pareto-dominance criterion. If some players are strictly better off and others are not worse off in a coalition of size $k^* = \hat{k}$ than in a coalition of size $k^* = \breve{k}$, we select $k^* = \hat{k}$. In our analysis, it turns out that coalition of size $k^* = \hat{k}$ is always the larger coalition.

[5] Millard-Ball (2012) assumes a four-stage game in which signatories choose mitigation levels first and then non-signatories. This assumption is unnecessarily complicated for payoff function (1) with linear benefits from global mitigation, as countries have dominant mitigation strategies. Hence, the assumption of Stackelberg leadership of signatories does not play a role. The same outcome materializes if signatories and non-signatories chose their mitigation levels simultaneously, as we do, which has been called the Nash-Cournot assumption in the literature (e.g., Finus 2003).

2.3 Three Equilibria

We consider that generally for any coalition of size k, $1 \le k \le n$, three possible equilibria can arise, which we call the Geoengineering-equilibrium (G-equilibrium), the Mitigation-equilibrium (M-equilibrium) and the Avoidance-equilibrium (A-equilibrium). We note that the decision in the last stage whether to deploy geoengineering is symmetric for all countries. Since collateral damages caused by other countries are taken as given, geoengineering will (not) be deployed if marginal benefits $\frac{1}{n}(g-Q)$ are positive (negative). Hence, if $Q \ge g$, geoengineering does not pay and $z_i^* = 0$. If the reverse is true, i.e., $Q < g$, geoengineering is deployed and $z_i^* = 1$. Hence, in the last stage, there is no difference between signatories and non-signatories, i.e., there is no need for cooperation. This is different in the second stage.

Geoengineering-Equilibrium (G-equilibrium)

In the G-equilibrium, $z_i^* = 1$ in the last stage. Consequently, from the first order conditions in an interior equilibrium in the second stage, equilibrium mitigation levels for signatories are $q_S^{G^*}(k) = \frac{k(bn-1)}{nc}$ and for non-signatories $q_{NS}^{G^*} = \frac{bn-1}{nc}$. Hence, the aggregate mitigation level is $Q^{G^*}(k) = k \cdot q_S^{G^*} + (n-k) \cdot q_{NS}^{G^*} = \frac{(bn-1)(k^2+n-k)}{nc}$. For consistency, we need to require weakly positive mitigation levels, $q_S^{G^*}(k) \ge 0$ and $q_{NS}^{G^*} \ge 0$, which is condition C_1 below. Moreover, total mitigation must fall short of the level g, $Q^{G^*}(k) < g$, for every k, and in particular for $k = n$ because total mitigation increases in the coalition size k, as otherwise this would not be compatible with $z_i^* = 1$. This is condition C_2 below.

$$C_1 := b \ge \frac{1}{n} \quad \text{and} \quad C_2 := g > \frac{bn^2 - n}{c}$$

Thus, the mitigation benefit parameter b must be sufficiently large such that equilibrium mitigation is an interior solution and the net geoengineering benefit parameter g must be sufficiently large in order to render the deployment of geoengineering attractive at all. We note that signatories mitigate more than non-signatories in the G-equilibrium, i.e., $q_S^{G^*}(k) > q_{NS}^{G^*}$ for any coalition of size k, $1 < k < n$, as they internalize the positive externality of the public good mitigation among the group of signatories. Hence, non-signatories are better off than signatories, i.e., $\pi_{NS}^{G^*}(k) > \pi_S^{G^*}(k)$. All countries have the same benefits from mitigation, the same net benefits and collateral damages from geoengineering, but signatories have higher mitigation costs than non-signatories.

Mitigation-Equilibrium (M-equilibrium)

In the M-equilibrium countries abstain from geoengineering in the last stage, i.e., $z_i = 0$. From the first order conditions in an interior equilibrium, $q_S^{M^*}(k) = \frac{kb}{c}$ and $q_{NS}^{M^*} = \frac{b}{c}$ follow. Again, as observed in the G-equilibrium, signatories mitigate more than non-signatories, i.e., $q_S^{M^*}(k) > q_{NS}^{M^*}$ for any coalition of size k, $1 < k < n$, and for this reason non-signatories receive a higher payoff that non-signatories, $\pi_{NS}^{M^*}(k) > \pi_S^{M^*}$. Total mitigation is given by $Q^{M^*}(k) = k \cdot q_S^{M^*}(k) + (n-k) \cdot q_{NS}^{M^*} = \frac{b\left(k^2 + n - k\right)}{c}$ which increases in the coalition size k, $\partial Q^{G^*}(k) \big/ \partial k > 0$. Consequently, given a coalition of size k, $1 \le k \le n$, the M-equilibrium exists if $Q^{M^*}(k) \ge g$ and does not exist if $Q^{M^*}(k) < g$. Given $Q^{M^*}(k) = \frac{b\left(k^2 + n - k\right)}{c}$, and using the short-hand notation $Q^{M^*}(k) = \frac{b}{c}(k)$, we can state the following.

Lemma 1: Existence of the M-equilibrium

If $g \le \frac{b}{c}(k)$, *then the M-equilibrium exists for all* $k, k, k+1,...,n$, *but does not exist for all* k, $1, 2,..., k-1$.

Thus, the M-equilibrium may or may not exist. If it exists, it may only exist for larger coalitions for which total mitigation is above the threshold g. However, generally, it could also exist for all coalition sizes k, $1 \le k \le n$, or not at all. We will later comment on the plausibility of equilibria. At this stage, it is important to note that for analytical tractability and plausibility, we want that if the M-equilibrium exists, the same applies to the G-equilibrium. Given that $\frac{b}{c}(k)$ takes on its smallest value for $k = 1$, then existence of the M-equilibrium requires $g \le \frac{b}{c}(1) = \frac{bn}{c}$. Existence of the G-equilibrium was condition C_2, which stated $g > \frac{bn^2 - n}{c}$. Hence, compatibility requires that $\frac{bn}{c} \ge \frac{bn^2 - n}{c}$, which gives condition C_3.

$$C_3 := b \le \frac{1}{n-1}$$

As we will discuss below, this condition is just a normalization of the parameter space of our model. It allows us to separate the parameter space in terms of the geoengineering parameter g into three distinct ranges, called Range 1, Range 2 and Range 3 in Figure 1.

Avoidance-Equilibrium (A-equilibrium)

The A-equilibrium considers the possibility that the M-equilibrium does not exist for a coalition of size k because the M-equilibrium falls short of the total mitigation level g, i.e., $Q^{M^*}(k) < g$. In the A-equilibrium, $z_i^* = 0$ and for total mitigation $Q^{A^*} \ge g$ must hold by assumption. We assume that non-signatories do not contribute to the extra effort to avoid

geoengineering. That is, $q_{NS}^{A^*} = q_{NS}^{G^*}$. Hence, all effort to avoid geoengineering is exerted by signatories. Given that signatories would normally set the sum of marginal benefits of mitigation equal to their individual marginal mitigation cost in equilibrium for $z_i^* = 0$, signatories have no incentive to provide more mitigation than is just necessary to achieve the threshold g. Hence, $Q^{A^*} = g$ with $q_S^{A^*}(k) = \frac{Q^{A^*} - (n-k)q_{NS}^{A^*}}{k} = \frac{cgn - n^2 b + bkn + n - k}{ckn}$ and $q_{NS}^{A^*} = q_{NS}^{G^*} = \frac{bn-1}{nc}$.

Clearly, in the grand coalition, $k = n$, all countries share equally mitigation efforts in order to achieve $Q^{A^*} = g$. Hence, $q_S^{A^*}(n) = \frac{g}{n}$. In contrast, in any coalition which is smaller than the grand coalition, signatories mitigate more that non-signatories, i.e., $q_S^{A^*}(k) > q_{NS}^{A^*}$ for every k, $1 \le k < n$. Hence, payoffs of non-signatories exceed those of signatories, $\pi_{NS}^{A^*} > \pi_S^{A^*}(k)$.

2.4 Policy Scenarios

In order to derive possible policy scenarios, we proceed in three steps. In a first step, we identify for which parameter ranges the three equilibria exist. In a second step, we identify which equilibrium will be played for a coalition of size k. That is, if multiple equilibria exist, we determine which equilibrium materializes if a coalition of size k forms. Third, we predict which equilibrium will be played across all possible coalition sizes k, $1 \le k \le n$. This gives rise to six policy scenarios.

The first step follows immediately from the previous analysis in terms of the geoengineering net benefit parameter g. This can be illustrated with the help of the upper panel in Figure 1. On the horizontal axis, parameter g is displayed, with the smallest value defined by condition C_2. No upper bound needs to be established. The G-equilibrium exists in the entire range of g

for every coalition of size k, $1 \le k \le n$. The M-equilibrium exists in Range 1 for every coalition of size k, $1 \le k \le n$. It may also exist in Range 2 for coalitions larger than size k, $k, k+1,...,n$. More precisely, the M-equilibrium exists for all values of g and coalitions of size k or larger provided $g \le \frac{b}{c}(k)$ holds. It does not exist if the reverse is true, $g > \frac{b}{c}(k)$. In Range 3, it does not exist for any coalition, regardless of its size. Finally, the A-equilibrium always exists for any value of g, as this was true for the G-equilibrium.

Figure 1 about here

In the second step, we have to ask the question which equilibrium will be played, provided it exists. This is related to the collateral damage parameter d. For this Lemma 2 is helpful.

Lemma 2: The Choice of Equilibrium Behavior

i) *For any coalition of size k, $1 \le k \le n$, if the M-equilibrium exists, the A-equilibrium is never played for any coalition of size k or larger.*

ii) *Given a coalition of size k, if the M-equilibrium exists, the G-equilibrium is not played for any coalition of size k or larger if collateral damages are sufficiently large, i.e., $d \ge \overline{d}^{M}(k)$. If $d < \overline{d}^{M}(k)$, then the G-equilibrium is played for any coalition of size k or smaller.*

iii) *Given a coalition of size k, if the M-equilibrium does not exist, the A- and not the G-equilibrium materializes if collateral damages are sufficiently large, i.e., $d \ge \overline{d}^{A}(k)$. If $d < \overline{d}^{A}(k)$, then the G-equilibrium is played for any coalition of size k or smaller.*

iv) *Both, $\overline{d}^{A}(k)$ and $\overline{d}^{M}(k)$, decrease in coalition size k, i.e., $\frac{\partial \overline{d}^{A}}{\partial k} < 0$ and $\frac{\partial \overline{d}^{M}}{\partial k} < 0$, and $\overline{d}^{M}(k) < \overline{d}^{A}(k)$ for every k, $1 \le k \le n$.*

Proof: Appendix A.1.

If the M-equilibrium exists, it is always preferred to the A-equilibrium, as the A-equilibrium is a kind of second-best equilibrium (Lemma 2, i). That is, only if the M-equilibrium does not exist, will the A-equilibrium be relevant as an alternative to the G-equilibrium.

Moreover, if the M-equilibrium exists at k, it emerges as an equilibrium provided collateral damages from geoengineering are perceived to be sufficiently high, i.e., $d \geq \overline{d}^{M}(k)$ (Lemma 2, ii). If $d \geq \overline{d}^{M}(k)$, signatories prefer the M- over the G-equilibrium, $\pi_S^{M*}(k) \geq \pi_S^{G*}$, and, given that they choose $q_S^{M*}(k)$, non-signatories follow suit, their best-response is to choose q_{NS}^{M*} instead of q_{NS}^{G*}.

With reference to the lower panel in Figure 1, which shows on the horizontal axis the possible ranges of the collateral damage parameter d, we can identify 3 ranges. In Range A, even if the M-equilibrium exists, only the G-equilibrium would emerge for any coalition of size k, $1 \leq k \leq n$, as collateral damages are very low, $d < \overline{d}^{M}(n)$. In Range B, the G-equilibrium would emerge for all coalitions (strictly) smaller than size k if collateral damages are moderate, i.e., $\overline{d}^{M}(n) \leq d < \overline{d}^{M}(k)$, but the M-equilibrium would materialize for any (weakly) larger coalition than k. Only for a sufficient number of signatories k, which share the burden of higher mitigation efforts, it pays to avoid geoengineering. In this context, k can be viewed as the minimum coalition size above which geoengineering can be avoided. Moving along the d -axis to the right in Range B, the M-equilibrium becomes increasingly attractive, i.e., the minimum coalition size k drops (Lemma 2, iv). Finally, in Range C, the collateral damage is sufficiently high such that even if no coalition forms at all, the M-equilibrium is preferred to

the G-equilibrium for any coalition size k, even for $k = 1$, as collateral damages are above $\overline{d}^{M}(1)$.

If the M-equilibrium does not exist, the choice is between the A- and G-equilibrium (Lemma 2, iii). The analysis is almost symmetric, as explained for the choice between the M- and G-equilibrium. In Range A, only the G-equilibrium materializes, as collateral damages are low. In Range B, for any damage parameter within the range $\overline{d}^{A}(n) \leq d < \overline{d}^{A}(k)$, also the G-equilibrium emerges for all (strictly) smaller coalitions that size k but the A-equilibrium for any (weakly) larger coalition than size k. Again, only if enough signatories share the mitigation burden does it pay to avoid the deployment of geoengineering. The larger the collateral damage parameter d, the more attractive is the A-equilibrium: the minimum coalition size k above which the A-equilibrium pays decreases when moving along the d-axis to the right (Lemma 2, iv). In Range C, the A-equilibrium emerges for any coalition size k, even if no agreement forms at all, i.e., $k = 1$.

In step 3, we combine steps 1 and 2 to derive 6 policy scenarios. They are summarized in Table 1 and illustrated in Figure 2. Table 1 focuses on the exact parameter range which generates a particular scenario and briefly summarizes under "Description" the underlying economics. Figure 2 translates this into characterizing equilibria for different coalition sizes, starting at the top with the grand coalition, then displaying smaller coalitions when moving down until the singleton coalition $k = 1$ is reached.

Table 1 and Figure 2 about here

There are three "pure scenarios" which imply that the same equilibrium is played for every coalition size k, $1 \leq k \leq n$. Moreover, there are three mixed scenarios which imply that for coalitions of size k or larger one equilibrium emerges and another equilibrium for coalitions $k-1$ and smaller.

The pure scenarios (Case 1, 2 and 3) are further illustrated in Figure 3. They emerge in situations in which geoengineering does not pose any governance issue, though for very different reasons. On the one hand, in the pure M- and A-scenarios, countries would never deploy geoengineering, even in absence of cooperation, as collateral damages are so high (Range C). On the other hand, in the pure G-scenario, collateral damages from geoengineering are so low that it is always economically rational to deploy geoengineering, even if all countries would join a climate agreement (Range A). Hence, the pure scenarios are listed for completeness and to highlight the difference to the other scenarios, but they are less interesting from a policy perspective.

Figure 3 about here

The mixed scenarios (Case 4, 5 and 6) are illustrated in Figure 4. In particular, Case 4 and 5 represent the most interesting cases. In Case 4, geoengineering net benefits are very high, making geoengineering attractive in the first place (Range 3 in Figure 4). Geoengineering can only be avoided with additional mitigation efforts by signatories (A-equilibrium). Even in the grand coalition the M-equilibrium cannot deliver a sufficiently high level of global mitigation to make geoengineering unattractive. For every coalition of size k, $1 \le k \le n$, the choice is between the A- and the G-equilibrium. Collateral damages are sufficiently high to make avoidance pay for a coalition of size k ($\overline{d}^{A}(k) \le d$), but not for any coalition which is (strictly) smaller than k ($d < \overline{d}^{A}(k-1)$). This is a part of Range B in Figure 4.

Figure 4 about here

Case 5 is similar to Case 4, with the distinct difference that the net benefits from geoengineering benefits are smaller such that the M-equilibrium exists, at least for larger coalitions (Range 1 or parts of Range 2 in Figure 4). As in Case 4, collateral damages are in an intermediate range, such that it pays to avoid geoengineering for larger coalitions than k ($\overline{d}^{M}(k) \le d$) though not

for smaller coalitions than k ($d < \overline{d}^{M}(k-1)$ or $d < \overline{d}^{A}(k-1)$). This is a part of Range B in Figure 4. For Case 5 two sub-cases, a and b, must be distinguished according to Table 1. Case 5a assumes that if a signatory leaves the agreement with k signatories the total mitigation still exceeds the threshold level g, $Q^{M^*}(k-1) \geq g$, i.e., $g \leq \frac{b}{c}(k-1)$ (Range 1 and parts of Range 2 in Figure 4), but the M-equilibrium is not played because countries prefer the G-equilibrium, which is true as long as $d < \overline{d}^{M}(k-1)$. Alternatively, Case 5b assumes that after the deviation the M-equilibrium does not produce sufficient mitigation, $Q^{M^*}(k-1) < g$, i.e., $g > \frac{b}{c}(k-1)$ (parts of Range 2 in Figure 4). The A-equilibrium is not played at $k-1$ because countries prefer the G-equilibrium due to sufficiently low collateral damages, i.e., $d < \overline{d}^{A}(k-1)$.

Finally, Case 6 has similarities with the pure M- and A-scenarios: collateral damages are very high such that geoengineering is never deployed (Range C in Figure 4). Hence, Case 6 does not look as interesting as Case 4 and 5. For large coalitions, the M-equilibrium is possible because $g \leq \frac{b}{c}(k)$ and is preferred to the G-equilibrium. For smaller coalitions, the M-equilibrium does not deliver sufficiently large total mitigation because $g > \frac{b}{c}(k-1)$; only the A-equilibrium exists and dominates the G-equilibrium due to high collateral damages. Thus, Case 6 implies parts of Range 2 and Range C in Figure 4.

This exhaust all possible policy scenarios, except one. If the g and d parameters lie in the intermediate range (Range 2 and Range B), it is possible that for large coalitions the M-equilibrium is played. Then, as the coalition size decreases, the M-equilibrium disappears in favour of the A-equilibrium, which, as the M-equilibrium, is preferred to the G-equilibrium. Eventually, as the coalition size further decreases, the G-equilibrium could emerge. This mixed

scenario is the only one in which we could observe a double switch of equilibria. However, a detailed analysis of this additional case is redundant, as it is a composition of Case 4 and 6. Hence, the same results directly apply.

Taken together, Case 1, 2 and 3 are pure policy scenarios that provide a benchmark for the mixed scenarios but are less interesting from an economic and policy perspective. Either collateral damages are so high that geoengineering would never be seriously considered by any country or they are so small that no country would ever worry about the deployment of geoengineering, and, in both cases, this is true for signatories and non-signatories. This also applies to the mixed scenario Case 6 which only emerges if collateral damages from geoengineering are very high. The mixed scenarios Case 4 and Case 5 describe more interesting situations in which the deployment of geoengineering generates high or moderate net benefits but also causes only moderate collateral damages. Hence, only if participation in a climate agreement is sufficiently high, can the deployment of geoengineering be possibly avoided. Whether and under which conditions such an agreement is stable will be analyzed in section 3 in the cartel formation game.

As a normative benchmark, it is useful to confirm that the often-made implicit assumption that "larger are better than smaller agreements" is indeed true for our model.

Proposition 1: Normative Properties of the Coalition Formation Game

Consider an expansion of a coalition of size $k-1$ *to* k. *Global welfare (i.e., sum of payoffs over all signatories and non-signatories) increases strictly for each expansion, with* $1 < k \leq n$, *for each of the six policy scenarios described in Table 1 and depicted in Figure 2. The same is true for total mitigation (i.e., the sum of mitigation levels over all countries), though the increase is not always strict, as in the pure A-scenario total mitigation remains always constant.*

Proof: See Appendix A.2.

Proposition 1 implies that global mitigation and welfare obtain their maximum in the grand coalition, i.e., an agreement in which all countries participate. The only exception in Proposition 1 relates to the pure avoidance scenario (Case 3) in which the global mitigation level is by definition $Q^{A^*}(k) = g$ for every k, $1 \le k \le n$, even though global welfare strictly increases with the enlargement of the agreement: the effort to avoid the deployment of geoengineering is shared by more signatories which reduces mitigation costs. In Appendix A.2, we derive further properties, which describe the incentive to join or abstain from coalitions. Those other properties are also useful for proving Proposition 1 and subsequent propositions.

2.5 Millard-Ball's Approach

Millard-Ball (2012) focuses on the stability analysis of the grand coalition. He rules out the possibility of the M-equilibrium to arise for every k, $1 \le k \le n$, which implies $Q^{M^*}(k) < g$. Since we know that $Q^{M^*}(k)$ increases in k, he assumes that even in the grand coalition $Q^{M^*}(n) < g$, which translates into $g > \frac{b}{c}(n)$, which is Range 3 in Figure 4. Additionally, Millard-Ball (2012) assumes that in the grand coalition the A-equilibrium is played while once a country free-rides, the G-equilibrium emerges. Hence, his analysis focuses on a particular sub-case of Case 4 for which $k = n$. He then claims to test for internal stability in the cartel formation game and concludes that only if the collateral damage parameter d is sufficiently large, the grand coalition will be stable. However, the analysis of Millard-Ball (2012) is flawed in many respects, leading to wrong conclusions.

First, he does not establish the condition under which all countries will play the A-equilibrium in the grand coalition and the G-equilibrium after a deviation such that a coalition of size $n-1$ forms. We know that this situation would arise only if the collateral damage parameter d is in the range $\overline{d}^{A}(n) \le d < \overline{d}^{A}(n-1)$, which is a part of Range B in Figure 4. Consequently, he does

not establish a lower bound for the damage parameter d, which turns out to be less of a problem, as we will demonstrate below, but, more importantly, he ignores the upper bound.

Second, he does not correctly calculate the free-rider payoff $\pi_{NS}^{G^*}(n-1)$ when deriving the condition for internal stability of the grand coalition. The free-rider payoff in the cartel formation game must be calculated for mutual best responses if a coalition of size $n-1$ forms: $\pi_{NS}^{G^*}(n-1)=\pi_{NS}^{G^*}\left(q_S^{G^*}(n-1),q_{NS}^{G^*}(n-1)\right)$. In other words, free-riding is in terms of leaving the grand coalition after which signatories and non-signatories choose their "new" equilibrium strategies, given that one country has left the agreement. Instead, Millard-Ball (2012) assumes that for the coalition of size $n-1$, signatories stick to their equilibrium strategies chosen in the grand coalition with $k=n$, $q_S^{A^*}=\frac{g}{n}$ and the free-rider chooses its best-reply, which is its dominant strategy level $q_{NS}^{G^*}=\frac{bn-1}{nc}$. This type of assumption would be more in line with what is typically assumed regarding the temporary free-rider payoff in a repeated game.

We draw two conclusions from our observations. First, there is a need to correctly calculate the free-rider payoff in the cartel formation game, including a full characterization of the parameter space for which this policy scenario would occur. This also includes a generalization that also consider other coalitions than the grand coalition and other policy scenarios than the A-G-scenario. This is done in section 3. Second, we test and confirm that all qualitative conclusions which we derive in the cartel formation game carry over to a repeated game. This is done in section 4.

3. Stability Analysis in the Cartel Formation Game

In this section, we analyze stable coalitions in the cartel formation game for the six policy scenarios, which we have identified in section 2.

3.1 Pure Policy Scenarios (Case 1, 2 and 3)

In the pure scenarios (M-, G-, and A-scenario), a given equilibrium (M-, G- or A-) is played for every coalition of size k, $1 \leq k \leq n$.

We noted above that these pure policy scenarios are less interesting from a policy perspective but serve as benchmarks for the comparisons with the other mixed policy scenarios.

Proposition 2: Stable Coalitions in the Pure M-, G- and A-scenario

In the cartel formation game, in the pure M-scenario (Case 1) and in the pure G-scenario (Case 2) the stable coalition size is $k^{M*} = k^{G*} = 3$. *In the pure A-scenario (Case 3) no non-trivial coalition is stable, i.e.,* $k^{A*} = 1$.

Proof: See Appendix A.3.

In the pure mitigation scenario with linear benefits and quadratic costs from mitigation, the result is well-known in the literature on international environmental agreements (e.g., Barrett 1994 and 2005 and Finus 2003): cooperation does not achieve much if the total number of countries exposed to the externality problem is substantially larger than three countries. We find the same to be true in G-scenario. Finally, if countries always play the avoidance equilibrium, then no agreement that includes at least a minimum of two countries is internally stable. In an A-equilibrium, signatories will always provide enough mitigation to avoid geoengineering. Hence, it is always attractive to leave a coalition and let the remaining signatories bear the mitigation burden. Given Proposition 1, this failure of effective cooperation entails a global welfare loss in all three policy scenarios.

3.2 Mixed Policy Scenarios (Case 4, 5 and 6)

In the mixed scenarios, once a country leaves coalition k, a different equilibrium is played at $k-1$. In section 2, we identified three possibilities: the mixed scenario A-G (Case 4), in which the A-equilibrium is played for every coalition weakly larger than k, while the G-equilibrium

is played for every coalition strictly smaller than k; the mixed scenario M-G (Case 5), in which the M-equilibrium is played for every coalition weakly larger than k, while the G-equilibrium is played for every coalition strictly smaller than k; the mixed scenario M-A (Case 6), in which the M-equilibrium is played for coalitions weakly larger than k, while the A-equilibrium is played for every coalition strictly smaller than k.

3.2.1 Mixed A-G-Scenario (Case 4): The Millard-Ball-Scenario

In the A-G-scenario (Case 4), the net benefits of geoengineering are so high such that the M-equilibrium never exists, even for the grand coalition. The only alternative to the G-equilibrium is the A-equilibrium, requiring additional mitigation efforts by signatories to avoid the deployment of geoengineering. However, the avoidance equilibrium only pays if the number of signatories passes the threshold of k members. Below, the deployment of geoengineering is a rational choice for signatories as well as non-signatories. For coalitions strictly larger than k (if they exist, i.e., $k \neq n$) implementing the A-equilibrium, we can immediately conclude from the proof of Proposition 2 that internal stability fails (because all coalitions for which $k > 1$ holds are not internally stable). Conversely, it follows that all coalitions larger or equal to k implementing the A-equilibrium are externally stable. Testing for internal stability at k implies to test under which conditions

$$\pi_S^{A^*}(k) \geq \pi_{NS}^{G^*}(k-1) \tag{4}$$

holds. Since $\Omega^{A-G}(k) := \pi_S^{A^*}(k) - \pi_{NS}^{G^*}(k-1)$ increases in the collateral damage parameter d, we find a critical value denoted by $\hat{d}^{A-G}(k)$ such that internal stability holds if $d \geq \hat{d}^{A-G}(k)$ (but fails if $d < \hat{d}^{A-G}(k)$). In order to complete the analysis, we need to take two more steps. First, we need to analyze how this stability threshold $\hat{d}^{A-G}(k)$ relates to our feasibility range in Case 4, which requires in terms of parameter d $\overline{d}^{A}(k) \leq d < \overline{d}^{A}(k-1)$. Moreover, we need

to analyze whether smaller coalitions than k, which implement the G-equilibrium, could also be stable and, if so, whether they are Pareto-dominated by the stable A-equilibrium at k. We leave those details to Appendix A.4.

Proposition 3: Stable Coalitions in the Mixed A-G-scenario

Consider the mixed A-G-scenario in the cartel formation game.

i) Let $k > 3$. The stable coalition size is $k^* = k$ with no geoengineering being deployed (A-equilibrium) if $\overline{d}^A(k) < \hat{d}^{A-G}(k) \leq d < \overline{d}^A(k-1)$, otherwise the stable coalition size is $k^* = 3$ with geoengineering being deployed (G-equilibrium).

ii) Let $k \leq 3$, $k^* = k$ with no geoengineering being deployed (A-equilibrium).

Proof: See Appendix A.4.

Hence, we confirm the qualitative conclusion of Millard-Ball (2012) considering the case that the A-equilibrium is established in a coalition with more than three members. Stability only holds if collateral damages are perceived to be severe enough in order to deter free-riding. This is because the threshold value of parameter d which we derive for internal stability $\hat{d}^{A-G}(k)$ lies above the lower bound of the feasibility range for this policy scenario $\overline{d}^A(k)$. However, we add the upper bound of the feasibility range $\overline{d}^A(k-1)$. Only if $d < \overline{d}^A(k-1)$ will the G-equilibrium be played at $k-1$; if $\overline{d}^A(k-1) \leq d$, the threat of punishment is an empty threat, i.e., it is not credible, as the A-equilibrium would be played. This upper bound has been ignored by Millard-Ball (2012). This theme will be recurrent in the following analysis.

Our further results emerge from the fact that we do not only consider the grand coalition but also possibly smaller coalitions and that we recognize the possibility of multiple equilibria. They are technically important but less for our take-away message.

3.2.2 Mixed M-G-Scenario and M-A-Scenario (Case 5 and 6)

The mixed M-G-scenario (Case 5) has been identified as the second most interesting policy scenario after the A-G-scenario (Case 4). The collateral damages are not so high as to render the deployment of geoengineering an irrational choice even for smaller coalitions. The agreement with k members produces enough global mitigation and collateral damages are severe enough in order to render the G-equilibrium not being attractive, such that the M-equilibrium is established at k. However, smaller coalitions do not pass the hurdle and therefore implement the G-equilibrium. Again, we know from the proof of Proposition 2 that if k is sufficiently large but is not the grand coalition, all strictly larger coalitions which implement the M-equilibrium are not internally stable. Hence, all coalitions larger or equal size k are externally stable. Hence, internal stability of coalition of size k implies to test whether

$$\pi_S^{M^*}(k) \geq \pi_{NS}^{G^*}(k-1) \tag{5}$$

holds. In the same spirit as in Case 4, a threshold value $\hat{d}^{M-G}(k)$ can be established such that internal stability holds if $d \geq \hat{d}^{M-G}(k)$ and this threshold is related to the feasibility range in Case 5, accounting for possible additional stable coalitions below coalition size k.

In the mixed M-A-scenario (Case 6), we know that all coalitions larger than k which implement the M-equilibrium are internally unstable, provided $n \geq k > 3$. Hence, all coalitions equal or larger k are externally stable. Hence, we need to test whether

$$\pi_S^{M^*}(k) \geq \pi_{NS}^{A^*}(k-1) \tag{6}$$

holds for a coalition of size k.

Proposition 4: Stable Coalitions in the Mixed M-G-scenario and M-A-scenario

Consider the mixed M-G-scenario in the cartel formation game.

i) Let $k > 3$. The stable coalition size is $k^* = k$ with no geoengineering being deployed (M-equilibrium) if $\overline{d}^M(k) < \hat{d}^{M-G}(k) \leq d < \overline{d}^M(k-1)$ in Case 5a) and if $\overline{d}^M(k) < \hat{d}^{M-G}(k) \leq d < \overline{d}^A(k-1)$ in Case 5b), otherwise, in both sub-cases, the stable coalition size is $k^* = 3$ with geoengineering being deployed (G-equilibrium).

ii) Let $k \leq 3$. The stable coalition size is $k^* = k$ with no geoengineering being deployed (M-equilibrium).

Consider the mixed M-A-scenario in the cartel formation game.

i) Let $k \geq 3$. There is no non-trivial coalition which is stable. Hence, $k^* = 1$ with no geoengineering being deployed (A-equilibrium).

ii) Let $k < 3$. The stable coalition size is $k^* = 3$ with no geoengineering being deployed (M-equilibrium).

Proof: See Appendix A.5.

The results for Case 5 are qualitatively similar to Case 4 (Proposition 3). Provided collateral damages are sufficiently high to be deterrent, but not too high to be credible, a coalition of size k can be stable because a deviation implies the deployment of geoengineering.

Case 6 does not produce non-trivial stable agreements when $k \geq 3$. The punishment of playing the avoidance equilibrium in case a coalition member leaves the agreement is simply not deterrent. When $k < 3$, Case 6 is very similar (equal in case of $k = 1$) to Case 2 as the M-equilibrium is played for coalitions larger or equal to size k, with $k^* = 3$.

3.3 Deployment of Geoengineering by All Countries

In this subsection, we depart from the original assumption of Millard-Ball (2012) that a random country deploys geoengineering in order test whether this impacts on our qualitative conclusions. Hence, the modified payoff function (1) reads

$$\pi_i = bQ - \frac{c\left(q_i\right)^2}{2} + \left(g - Q\right)\cdot z_i - \left(n-1\right)d\cdot z_i \tag{7}$$

with $z_i = 0$ if geoengineering is not deployed and $z_i = 1$ if it is deployed. If $z_i = 0$, as in the M- and A-equilibrium, nothing changes and in the G-equilibrium, all countries deploy geoengineering receiving net benefits $g - Q$ and suffering from collateral damages $\left(n-1\right)d$ arising from the deployment of geoengineering by all $n-1$ other countries. Not surprisingly, all qualitative results obtained Propositions 1, 2, 3 and 4 continue to hold, as the net benefits and collateral damages from geoengineering have just been recalibrated. (Details are provided in our Online Appendix O.4.)

Viewing these propositions together, three conclusions can be drawn. First, the pure policy scenarios confirmed the small coalition paradox, as Nordhaus (2015) coined it. Self-enforcing agreements are small and therefore fall substantially short of providing socially optimal mitigation and welfare levels. Second, the mixed policy scenarios Case 4 and 5 may produce larger stable coalitions, including the grand coalition, through the threat to deploy geoengineering in case of deviations. However, it is not as simple as one may conjecture. The perception of collateral damages must not only be sufficiently large (lower bound), but cannot be too large (upper bound), as otherwise the threat loses its credibility. Third, also the conjecture that the larger collateral damages are, the larger stable agreements will be is wrong. We recall that for a coalition of size k to be stable we need for the collateral damage parameter d to be in the range $\overline{d}^{A}\left(k\right) < \hat{d}^{A-G}\left(k\right) \le d < \overline{d}^{A}\left(k-1\right)$ in Case 4 and $\overline{d}^{M}\left(k\right) < \hat{d}^{M-G}\left(k\right) \le d < \overline{d}^{M}\left(k-1\right)$ in Case 5a and $\overline{d}^{M}\left(k\right) < \hat{d}^{M-G}\left(k\right) \le d < \overline{d}^{A}\left(k-1\right)$ in Case 5b (Proposition 3 and 4). These ranges do not move upward but downward with the coalition size k (Lemma 2). This a variant of the second conclusion: the threat of deploying geoengineering must be credible. Larger and more successful stable climate agreements require not larger but smaller collateral

damages in order for countries to believe that a deviation really triggers the deployment of geoengineering.

4 Stability Analysis in the Repeated Game

4.1 Preliminaries

In this section, we analyze stability in a repeated game. On the one hand, this would be an alternative route to correct the mistake by Millard-Ball (2012). On the other hand, we can test the robustness of our qualitative conclusions in a conceptually different framework. We consider a simple subgame-perfect trigger strategy to drive home our results, being aware that more sophisticated equilibrium strategies are available.[6] As most papers which analyze repeated games, we focus on the stability of the grand coalition in order to keep the presentation short and simple.

There are three phases in a repeated game which establishes cooperation using a trigger strategy: cooperation during which all countries receive a per period payoff π_i^C, free-riding, giving the free-rider a temporary payoff π_i^F for one period until free-riding has been discovered and the response to free-riding, namely punishment, which gives a per period payoff π_i^P for the rest of the game. Cooperation is preferred if the discounted payoff from cooperation is at least as large as the discounted payoff from temporarily free-riding and subsequently being punished:

$$\frac{\pi_i^C}{1-\delta} \geq \pi_i^F + \frac{\delta \pi_i^P}{1-\delta} \;\Leftrightarrow\; \delta \geq \delta_{min} = \frac{\pi_i^F - \pi_i^C}{\pi_i^F - \pi_i^P} \tag{8}$$

[6] The requirements on the credibility of sanctions increase along the following sequence of equilibrium concepts (see, e.g., for an overview Finus 2003 and Finus and Caparros 2015): Nash equilibrium (NE), subgame perfect equilibrium (SPE), weakly renegotiation-proof (WRPE), strongly renegotiation-proof (SRPE) and strong subgame-perfect equilibrium (SSPE). Applications in the context of game-theoretic analyses of IEAs are provided for instance by Asheim et al. (2006), Asheim and Holtsmark (2009) and Finus and Rundshagen (1998).

with δ the discount factor by which time is discounted, $0 \le \delta \le 1$. We call δ_{min} the minimum discount factor requirement. The larger δ, the less countries discount time and the less attractive is the temporary free-rider gain followed by punishment compared to compliance with the terms of the agreement.

4.2 Stability of Policy Scenarios

We introduced six policy scenarios in section 2 in the context of the cartel formation game. In the context of the repeated game, the three pure policy scenarios can be directly applied. Thus, Table 1 and Figure 3 apply regarding the parameter range g and d.

In the pure M-scenario (Case 1), $\pi_i^C = \pi_S^{M^*}(n)$ and $\pi_i^P = \pi_{NS}^{M^*}(1)$ (with $\pi_{NS}^{M^*}(1) = \pi_S^{M^*}(1)$). Free-riding implies that signatories continue to mitigate $q_s^{M^*}(n) = \frac{nb}{c}$ whereas the free-rider choses its best-response, which is $q_{NS}^{M^*} = \frac{b}{c}$. This deviation leads to total mitigation $\tilde{Q}^{M^*}(n-1) = \frac{b(n^2-n+1)}{c}$. Introducing shorth-hand notation, it can be shown (Appendix A.6) $\frac{\tilde{b}}{c}(n-1) := \frac{b(n^2-n+1)}{c} > \frac{b}{c}(1) > g$, given the parameter range which defines Case 1 (Range 1, see Table 1 and Figure 3).[7] That is, a deviation does not lead to the implementation of geoengineering but to a kind of M-equilibrium with the free-rider payoff $\pi_i^F = \tilde{\pi}_i^{M^*}(q_S^{M^*}(n), q_{NS}^{M^*})$. This equilibrium is different from the M-equilibrium in the cartel formation game as signatories play $q_S^{M^*}(n)$ and not $q_S^{M^*}(n-1)$.

[7] $\frac{\tilde{b}}{c}(n)$ in the context of the repeated game has a similar interpretation as $\frac{b}{c}(k)$ in the context of the cartel formation game. If the parameter g is weakly below this threshold, the M-equilibrium exist and if it is above, it does not exist.

In the pure G-scenario (Case 2), $\pi_i^C = \pi_S^{G*}(n)$, $\pi_i^P = \pi_{NS}^{G*}(1)$ (with $\pi_{NS}^{G*}(1) = \pi_S^{G*}(1)$) and $\pi_i^F = \tilde{\pi}_i^{G*}(q_S^{G*}(n), q_{NS}^{G*})$. No complication arises. In the pure A-scenario (case 3), cooperation implies that every country contributes to achieve the level $Q^{A*} = g$ to avoid the deployment of geoengineering, i.e., $q_S^{A*}(n) = \frac{g}{n}$. A deviation will therefore inevitably lead to $\tilde{Q}^{A*} < g$, i.e., geoengineering is implemented. The free-rider anticipating this will consequently choose the optimal mitigation level q_{NS}^{G*} compatible with the implementation of geoengineering. The only punishment available in case 3 is to play the A-equilibrium, as, by assumption, collateral damages are so high that even if there is no cooperation, $k = 1$, the A-equilibrium is preferred to the G-equilibrium (Range C in Figure 3). However, playing the A-equilibrium during punishment is not an effective punishment.

Proposition 5: Stability of the Grand Coalition in the Pure M-, G- and A-Scenario

In a repeated game, in the pure M- and G-scenario, the grand coalition is stable for discount factors δ *for which* $\delta \geq \delta_{min}^M = \delta_{min}^G = \frac{1}{2}$ *holds. In the pure A-scenario, the grand coalition is never stable.*

Proof: See Appendix A.6.

In order to evaluate these results, it is useful to derive those for the mixed policy scenarios first. Compared to the cartel formation game, the parameter ranges for which these policy scenarios arise need to be slightly adjusted. This is shown in Figure 5 and summarized in Table A in Appendix A.7.

Figure 5 about here

Those mixed scenarios are constructed in the following way. First, the parameter ranges of parameter g and d need to be specified which guarantees that a particular equilibrium is

played in the grand coalition and during punishment. In the A-G-scenario (Case 4), this implies that in the cooperative phase the A-equilibrium and in the punishment phase the G-equilibrium is played. In the M-G-scenario (Case 5), the M-equilibrium is played during cooperation and the G-equilibrium during punishment. Finally, in the M-A-scenario, the M-equilibrium is played during cooperation and the A-equilibrium during punishment.

The logic which we apply to construct these mixed policy scenarios is the following. First, the equilibrium of a coalition of size k which is tested for stability in the cartel formation game is chosen to be the equilibrium implemented in the grand coalition in the repeated game. Second, we note that in the cartel formation game deviation and punishment coincide. We choose the equilibrium scenario which emerges in the cartel formation game if a player leaves a coalition of size k such that $k-1$ forms as the punishment equilibrium in the repeated game for $k=1$. Then we consider all possible scenarios which could emerge if a player takes a free-ride.

In the mixed A-G-scenario (Case 4) this is straightforward. We have $\pi_i^C = \pi_S^{A*}(n)$, $\pi_i^P = \pi_{NS}^{G*}(1)$ and $\pi_i^F = \tilde{\pi}_i^{G*}\left(q_S^{A*}(n), q_{NS}^{G*}\right)$. The A-equilibrium is implemented in the grand coalition because the benefits of geoengineering are high. That is, the M-equilibrium does not exist for any coalition size. With respect to parameter g, this is Range 3, as noted in the context of the cartel formation game. Moreover, collateral damages are sufficiently high such that in the grand coalition the A- is preferred to the G-equilibrium. Thus, $\bar{d}^A(n) \leq d$. This excludes Range A, as in the cartel formation game. In order to be able to punish with the G-equilibrium if $k=1$, we need to require $d < \bar{d}^A(1)$. Collateral damages are not too high. This excludes Range C as in the cartel formation game. Taken together, this gives the entire Range B with $\bar{d}^A(n) \leq d < \bar{d}^A(1)$. Compared to the cartel formation game, only the permissible d-range is larger in the repeated game. Finally, free-riding will inevitably lead to the deployment of

geoengineering. Even though all compliant signatories choose $q_S^{A*}(n) = \frac{g}{n}$, the free-rider will choose q_{NS}^{G*} if he/she deviates at all. Hence, total mitigation will drop below g which would be required to avoid geoengineering.

The mixed M-G-scenario (Case 5) is slightly more involved, as this was also the case in the cartel formation game. In the cartel formation game (see Table 1), we had to distinguish whether after the deviation from the M-equilibrium, a) total mitigation is still above the benchmark level g, such that there is no incentive to deploy geoengineering or whether b) total mitigation falls short of this level g and geoengineering is implemented. As discussed in the context of the pure M-scenario above, a) implies $g \leq \frac{\tilde{b}}{c}(n-1)$ and b) $g > \frac{\tilde{b}}{c}(n-1)$ with $\frac{bn^2 - n}{c} < \frac{b}{c}(1) < \frac{\tilde{b}}{c}(n-1) < \frac{b}{c}(n)$ along the g-range in the repeated game.

In all sub-cases of Case 5, in the grand coalition the M-equilibrium is played and this requires $g \leq \frac{b}{c}(n)$ and $\overline{d}^M(n) \leq d$. In all sub-cases of Case 5, the G-equilibrium is played during punishment and this requires either $d < \overline{d}^M(1)$ if the M-equilibrium exists for $k = 1$ but is inferior to the G-equilibrium (Case 5a,i) or $d < \overline{d}^A(1)$ if the M-equilibrium does not exists, but the A-equilibrium is dominated by the G-equilibrium (Case 5a,ii & Case 5b). Taken together, in Case 5, $\pi_i^C = \pi_S^{M*}(n)$ and $\pi_i^P = \pi_{NS}^{G*}(1)$. In sub-case 5a, $\pi_i^F = \tilde{\pi}_i^{M*}\left(q_S^{M*}(n), q_{NS}^{M*}\right)$ and in sub-case 5b $\pi_i^F = \tilde{\pi}_i^{G*}\left(q_S^{M*}(n), q_{NS}^{G*}\right)$.

The mixed M-A-scenario (Case 6) follows a very similar logic than in Case 5. Sub-case 6a assumes that after a deviation from the M-equilibrium, total mitigation remains at a level that avoids the implementation of geoengineering and sub-case 6b assumes the opposite.

Proposition 6: Stability of the Grand Coalition in the Mixed A-G, M-G and M-A-Scenarios

In a repeated game, the grand coalition is stable in the mixed A-G-scenario (Case 4) and the mixed M-G-scenario (Case 5a and 5b) for discount factors δ *for which* $\delta \geq \delta_{min}^{A-G}$, $\delta \geq \delta_{min}^{M-G}(a)$ *and* $\delta \geq \delta_{min}^{M-G}(b)$ *hold, respectively. The minimum discount factor* δ_{min} *decreases in parameter* d, $\frac{\partial \delta_{min}}{\partial d} < 0$ *in both cases.*

In the A-G-scenario, the grand coalition is stable if

i) $1 \geq \delta \geq \delta_{min}^{A-G} > \frac{1}{2}$ *provided:* $\overline{d}^{A}(n) \leq d < \tilde{d}^{A-G}$

ii) $\frac{1}{2} \geq \delta \geq \delta_{min}^{A-G} > 0$ *provided:* $\tilde{d}^{A-G} \leq d < \tilde{\tilde{d}}^{A-G}$

iii) is always stable *provided:* $\tilde{\tilde{d}}^{A-G} \leq d < \overline{d}^{A}(1)$,

with $\overline{d}^{A}(n) < \tilde{d}^{A-G} < \tilde{\tilde{d}}^{A-G} < \overline{d}^{A}(1)$.

In the M-G-scenario, Case 5a, the grand coalition is stable if

i) $1 \geq \delta \geq \delta_{min}^{M-G}(a) > \frac{1}{2}$ *provided:* $\overline{d}^{M}(n) \leq d < \tilde{d}^{M-G}(a) = \overline{d}^{M}(1)$

ii) $\frac{1}{2} \geq \delta \geq \delta_{min}^{M-G}(a) > 0$ *provided:* $\overline{d}^{M}(1) = \tilde{d}^{M-G}(a) \leq d < \overline{d}^{A}(1)$

with $\overline{d}^{A}(n) < \tilde{d}^{M-G}(a) = \overline{d}^{M}(1) < \overline{d}^{A}(1)$.

In the M-G-scenario, Case 5b, the grand coalition is stable if

i) $1 \geq \delta \geq \delta_{min}^{M-G}(b) > \frac{1}{2}$ *provided:* $\overline{d}^{M}(n) \leq d < \tilde{d}^{M-G}(b)$

ii) $\frac{1}{2} \geq \delta \geq \delta_{min}^{M-G}(b) > 0$ *provided:* $\tilde{d}^{M-G}(b) \leq d < \tilde{\tilde{d}}^{M-G}$

iii) is always stable *provided:* $\tilde{\tilde{d}}^{M-G} \leq d < \overline{d}^{A}(1)$,

with $\overline{d}^{A}(n) < \tilde{d}^{M-G}(b) < \tilde{\tilde{d}}^{M-G} < \overline{d}^{A}(1)$.

In the M-A-scenario, the grand coalition is never stable.

Proof: See Appendix A.7.

We would like to interpret our results for the six policy scenarios in the repeated game, as summarized in Proposition 5 and 6, by drawing directly on the comparison to those obtained in the cartel formation game, as summarized in Proposition 2, 3 and 4.

First, we recall that we argued that the pure policy scenarios (Case 1, 2 and 3) are less interesting from a policy point of view because either collateral damages are so high that geoengineering never poses a threat or they are so low that the implementation of geoengineering is always rational even if all countries cooperate. We also pointed out that the mixed M-A-policy scenario (Case 6) is also not really interesting because it would only materialize if collateral damages are very high. This incentive structure is exactly preserved in the repeated game.

Second, we found no stable coalition in case 3 and 6 in the cartel formation game and the same is true in the repeated game. Third, in case 1 and 2, we found the stable coalition in the cartel formation game to be always of size $k^* = 3$ and find in the repeated game a constant minimum discount factor of $\delta_{min} = \frac{1}{2}$. Fourth, in case 4 and 5, we found that the internal stability function increases in parameter d in the cartel formation game and find now that the minimum discount factor decreases in d in the repeated game. Fifth, in the cartel formation game, we found that stability of coalition requires that the collateral damage parameter d needs to be in some range. The damage must exceed a lower bound in order to make the A-equilibrium (Case 4) and the M-equilibrium (Case 5) attractive in a cooperative agreement and to avoid that geoengineering is deployed. The upper bound was needed to the make the threat to deploy geoengineering credible in case of a deviation. Exactly, the same is true in the repeated game. Sixth, in the cartel formation game, the mixed policy scenarios Case 4 and 5 could generate larger stable coalitions than the pure policy scenarios Case 1 and 2. Comparing Proposition 5 with 6, the same is

qualitatively true in the repeated game. Given our assumption to focus entirely on the grand coalition for simplicity in the repeated game, the relevant reference is δ_{min}. In the pure scenarios, $\delta_{min} = \frac{1}{2}$ and in mixed policy scenarios δ_{min} can be below $\frac{1}{2}$ if d is sufficiently large, even though d must be below the upper threshold $\overline{d}^{A}(1)$.

5 Conclusion

The possibility to use solar radiation management, as one form of geoengineering, can profoundly change the incentives to participate in a climate agreement, aiming at reducing greenhouse gas emissions (mitigation). This was analyzed in a simple model proposed by Millard-Ball (2012), which captures the interplay between mitigation and geoengineering. Depending on the level of the benefits and collateral damages from geoengineering, we identified different policy scenarios. In a cartel formation game, the threat of implementing geoengineering with its associated collateral damages can stabilize large climate agreements (including full participation) without the deployment of geoengineering. The most interesting scenario implied that signatories, in their own interest, go the extra mile and increase mitigation above a level such that the deployment of geoengineering was not attractive to all countries, in particular not to non-signatories. We called this the avoidance equilibrium. The second most interesting scenario implied that if a sufficient number of countries join a climate agreement, global mitigation is sufficiently high to render the deployment of geoengineering unattractive to all countries. We called this the mitigation equilibrium. In both cases, this was because the marginal benefits from geoengineering decrease in the level of global mitigation. Hence, if global mitigation is sufficiently high, geoengineering simply does not pay. In both cases, implementing a climate agreement and avoiding the deployment of geoengineering was only incentive compatible provided collateral damages from geoengineering are perceived to be sufficiently high. This delivered a lower bound on collateral damages. However, we also

showed that such climate agreements are only stable if the threat is credible that should a signatory leave the climate agreement, the deployment of geoengineering is in the interest of all countries. This was only the case if collateral damages are below a threshold, which gave us an upper bound on collateral damages. We argued that this upper bound was ignored by Millard-Ball (2012).

We then tested our qualitative conclusions in a different conceptual framework, namely a repeated game. On the one hand, we argued that the incorrectly calculated free-rider payoff in the cartel formation game by Millard-Ball (2012) had more of the spirit of what is typically assumed in a repeated game. One the other hand, we could test the robustness of our conclusions obtained in the cartel formation game. We found all qualitative conclusions confirmed.

The possibility of using geoengineering, proposed as a quick fix for the climate change problem, but with possibly high collateral damages, can be a game changer for the formation of climate agreements. The good news is that if countries are aware of possibly high collateral damages, this can enforce larege climate agreements aiming at reducing global greenhouse gases. The bad news is that if countries perceive those damages too high, then the game does not change: only climate agreements with low participation are stable because free-riding prevails.

Clearly, further research is needed in order to fully understand the incentives of climate agreement formation in the presence of geoengineering in the form of solar radiation management. Millard-Ball's model provides a good starting point, but it abstracts from several aspects which are important in climate change. The most obvious aspect relates to the fact that countries have a different perception of the benefits and collateral damages of geoengineering, face different mitigation costs and are differently affected by climate change damages, whereas we assumed symmetric countries. Another aspect is that we assumed the deployment of geoengineering to be a discrete choice where in a richer model this could be modelled as a continuous choice.

References

Asheim, G.B. and B. Holtsmark (2009), Renegotiation-Proof Climate Agreements with Full Participation: Conditions for Pareto-Efficiency. "Environmental and Resource Economics", vol. 43, pp. 519–533.

Asheim, G.B., C. Bretteville Froyn, J. Hovi and F.C. Menz, (2006), Regional versus global cooperation for climate control. "Journal of Environmental Economics and Management", Vol. 51(1), pp. 93-109.

Barrett, S. (1994), Self-Enforcing International Environmental Agreements. "Oxford Economic Papers", col. 46, pp. 878–894.

Barrett, S. (2005), The Theory of International Environmental Agreements. Chapter 28 in "Handbook of Environmental Economics", vol. 3, pp. 1457–1516.

Barrett, S. (2007), Why cooperate? The incentive to supply global public goods. New York: Oxford University Press.

Barrett, S. (2008), The Incredible Economics of Geoengineering. "Environmental and Resource Economics", vol. 39(1), pp. 45–54.

Barrett, S. (2014), Solar Geoengineering's Brave New World: Thoughts on the Governance of an Unprecedented Technology. "Review of Environmental Economics and Policy", vol. 8, pp. 249–269.

Bayramoglu, B., M. Finus and J.-F. Jacques (2018), Climate Agreements in Mitigation-Adaptation Game. "Journal of Public Economics", vol. 165, pp. 101-113.

Blackstock, J. and J. Long (2010). Climate Change. The Politics of Geoengineering. "Science", vol. 327(5965), pp. 527-527.

Bodansky, D. (2013). The Who, What, and Wherefore of Geoengineering Governance. "Climatic Change", vol. 121(3), pp. 539–551.

Borrero, M. and S.J. Rubio (2022), An Adaptation-Mitigation Game: Does Adaptation Promote Participation in International Environmental Agreements? "International Environmental Agreements: Politics, Law and Economics", https://doi.org/10.1007/s10784-021-09560-5.

Caldeira, K., G. Bala, and L. Cao (2013), The Science of Geoengineering. "Annual Review of Earth and Planetary Sciences", vol. 41(1), pp. 231–256.

Carraro, C. and D. Siniscalco (1993), Strategies for the International Protection of the Environment. "Journal of Public Economics", vol. 52(3), pp. 309–328.

Finus, M. (2003). Stability and design of international environmental agreements: the case of transboundary pollution, Edward Elgar Publishing Ltd, UK United Kingdom, pp. 82–158.

Finus, M. and Caparros, A. (2015). Handbook on Game Theory and International Environmental Cooperation: Essential Readings, The International Library of Critical Writings in Economics Series, Edward Elgar Publishing Ltd, UK United Kingdom.

Finus, M., F. Furini and A.V. Rohrer (2021), The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs Stackelberg leadership. "Journal of Environmental Economics and Management", vol. 109, 102461.

Finus, M. and B. Rundshagen (1998), Toward a Positive Theory of Coalition Formation and Endogenous Instrumental Choice in Global Pollution Control. "Public Choice", vol. 96, pp. 145-186.

Fuss, S., J.G. Canadell, P. Ciais, R.B. Jackson, C.D. Jones, A. Lyngfelt, G.P. Peters, D.P. Van Vuuren (2020), Moving Toward Net-Zero Emissions Requires New Alliances for Carbon Dioxide Removal. "One Earth", vol. 3(2), pp. 145-149.

Heyen, D., J. Horton and J. Moreno-Cruz (2019), Strategic Implications of Counter-Geoengineering: Clash or Cooperation? "Journal of Environmental Economics and Management", vol. 95, pp. 153-177.

IPCC (2014). "Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change"

IPCC (2018). "Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty"

IPCC (2021), Climate Change 2021: The Physical Science Basis. "Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change".

Klepper, G. and W. Rickels (2014), Climate Engineering: Economic Considerations and Research Challenges. "Review of Environmental Economics and Policy", vol. 8, pp. 270– 289.

Lazkano, I., W. Marrouch and B. Nkuiya (2016), Adaptation to Climate Change: How Does Heterogeneity in Adaptation Costs Affect Climate Coalitions? "Environment and Development Economics", vol. 21(06), pp. 812–838.

Marrouch, W. and A.R. Chaudhuri (2015), International Environmental Agreements: Doomed to Fail or Destined to Succeed? A Review of the Literature. "International Review of Environmental and Resource Economics", vol. 9, pp. 245–319.

Meadowcroft, J. (2013), Exploring Negative Territory Carbon Dioxide Removal and Climate Policy Initiatives. "Climatic Change", vol. 118, pp. 137–149.

Millard-Ball (2012). The Tuvalu Syndrome. Can geoengineering solve climate's collective action problem?, "Climatic Change", vol. 110, pp. 1047-1066.

Moreno-Cruz, J.B. (2015), Mitigation and the Geoengineering Threat. "Resource and Energy Economics", 41, pp. 248–263.

National Academies of Sciences, Engineering, and Medicine (2021), Reflecting Sunlight: Recommendations for Solar Geoengineering Research and Research Governance. Washington, DC: The National Academies Press.

Nordhaus, W.D. (2015), Climate clubs: Overcoming free-riding in international climate policy. "American Economic Review", 105(4), pp. 1339-1370.

Parker, A. and P.J. Irvine (2018). The Risk of Termination Shock From Solar Geoengineering. "Earth's Future", vol. 6(3), pp. 456-467.

Reynolds, J.L. (2019), The Governance of Solar Geoengineering: Managing Climate Change in the Anthropocene. Cambridge: Cambridge University Press.

Ricke, K.L., J.B. Moreno-Cruz, and K. Caldeira (2013), Strategic Incentives for Climate Geoengineering Coalitions to Exclude Broad Participation. "Environmental Research Letters", vol. 8, pp. 1-11.

Sandler, T. (2004), Global collective action. New York: Cambridge University Press.

Sandler, T. (2018), Collective Action and Geoengineering. "The Review of International Organizations, vol. 13, pp. 105–125.

Stephens, J.C., P. Kashwan, D. McLaren, and K. Surprise (2021), The Risks of Solar Geoengineering Research. "Science", vol. 372(6547), pp. 1161-1161.

The Royal Society and J. Shepherd (2009), Geoengineering the Climate: Science, Governance and Uncertainty. London: The Royal Society.

Urpelainen, J. (2012), Geoengineering and Global Warming: a Strategic Perspective. "International Environmental Agreements: Politics, Law and Economics", 12 (4), pp. 375–389.

Weitzman, M.L. (2015), A Voting Architecture for the Governance of Free-Driver Externalities, with Application to Geoengineering. "The Scandinavian Journal of Economics", 117 (4), pp. 1049–1068.

Appendix

In the following, we only present the general idea of our proofs. Due to space constraint, we refer to our Online Appendix for further details.

A.1 Proof of Lemma 2

Suppose the M-equilibrium exists. $\pi_S\left(q_S, q_{NS}^{M^*}\right) > \pi_S\left(q_S, q_{NS}^{G^*}\right)$ because $q_{NS}^{M^*} > q_{NS}^{A^*} = q_{NS}^{G^*}$. Hence, signatories have no incentive to depart from their best response in the M-equilibrium if the M-equilibrium exists. That is, $\pi_S\left(q_S^{M^*}, q_{NS}^{M^*}\right) \geq \pi_S\left(q_S, q_{NS}^{M^*}\right)$ if $q_S \neq q_S^{M^*}$. Furthermore, we can show that for any given coalition size signatories would prefer the M- to the A-equilibrium, i.e., $\pi_S^{M^*}\left(q_S^{M^*}, q_{NS}^{M^*}\right) > \pi_S^{A^*}\left(q_S^{A^*}, q_{NS}^{A^*}\right)$ for every k, $1 \leq k \leq n$. If signatories choose $q_S^{M^*}$, the best-response by non-signatories is $q_{NS}^{M^*}$, which follows from the equilibrium conditions in the M-equilibrium. Thus, if the M-equilibrium exists, the A-equilibrium will not be played.

Suppose the M-equilibrium exists. Signatories prefer the M- over the G-equilibrium if $\pi_S^{M^*}(k) - \pi_S^{G^*}(k) \geq 0$. If this condition holds, not only signatories but also non-signatories will be better off by moving from the G- to the M-equilibrium. This is true because $\pi_{NS}^{M^*}(k) - \pi_{NS}^{G^*}(k) > \pi_S^{M^*}(k) - \pi_S^{G^*}(k)$ holds. Moving from the G- to the M-equilibrium, both signatories and non-signatories will experience the same welfare effects with respect to the net benefits of geoengineering and collateral damages, both will have the same increase in mitigation benefits, but signatories will face a larger increase in mitigation costs compared to non-signatories as $q_S^{M^*} - q_S^{G^*} > q_{NS}^{M^*} - q_{NS}^{G^*}$ and mitigation cost functions are strictly convex. This inequality holds because it can be rewritten as $k \cdot q_{NS}^{M^*} - k \cdot q_{NS}^{G^*} > q_{NS}^{M^*} - q_{NS}^{G^*}$, using the fact that $q_S^{M^*} = k \cdot q_{NS}^{M^*}$ and $q_S^{G^*} = k \cdot q_{NS}^{G^*}$ from section 2. Hence, the choice between the M- and G-equilibrium will depend on whether $\pi_S^{M^*}(k) - \pi_S^{G^*}(k) \geq (<) 0$. Differentiating

$\pi_S^{M^*}(k) - \pi_S^{G^*}(k)$ with respect to d, it can be shown that the difference increases in d. Hence, we solve for d and obtain the critical damage level

$$\overline{d}^M(k) = -\frac{2bk^2n - 4bkn + 4bn^2 - 2gnc - k^2 + 2k - 2n}{2cn(n-1)}.$$

The M-equilibrium is played for $d \geq \overline{d}^M(k)$ and the G-equilibrium if $d < \overline{d}^M(k)$. Differentiating $\overline{d}^M(k)$, we find: $\frac{\partial \overline{d}^M(k)}{\partial k} = -\frac{(k-1)(2bn-1)}{cn(n-1)} < 0$ using condition C_1.

Suppose the M-equilibrium does not exist. Non-signatories choose the same mitigation level in the G- and A-equilibrium. Signatories increase their mitigation level from $q_S^{G^*}(k)$ to $q_S^{A^*}(k)$ and avoid geoengineering if $\pi_S^{A^*}(k) - \pi_S^{G^*}(k) \geq 0$ holds. Consequently, if this condition holds, also non-signatories must be better off, as they have not changed their mitigation levels, i.e., $\pi_{NS}^{A^*}(k) - \pi_{NS}^{G^*}(k) > \pi_S^{A^*}(k) - \pi_S^{G^*}(k)$. Moving from the G- to the A-equilibrium, both, signatories and non-signatories will experience the same welfare effects with respect to the net benefits of geoengineering and collateral damages, both will have the same increase in mitigation benefits, but signatories will face an increase in mitigation costs, which does not occur to non-signatories. Differentiating $\pi_S^{A^*}(k) - \pi_S^{G^*}(k)$ with respect to d, it can be shown that the difference increases in d. Hence, solving for d, we obtain the critical damage level

$$\overline{d}^A(k) = \frac{\left(bk^2n - bkn + bn^2 - gnc - k^2 + k - n\right)^2}{2cnk^2(n-1)}.$$

The A-equilibrium is played for $d \geq \overline{d}^A(k)$ and the G-equilibrium is played for $d < \overline{d}^A(k)$. We find: $\frac{\partial \overline{d}^A(k)}{\partial k} = \frac{\left(nbk^2 - bkn + bn^2 - gnc - k^2 + k - n\right)\left(nbk^2 - bn^2 + gnc - k^2 + n\right)}{cnk^3(n-1)} < 0$.

The denominator is clearly positive, while the first bracket is negative and the second bracket positive in the nominator. This can be shown by using conditions C_1 and C_2 in the text. Furthermore, $\overline{d}^{M}(k) < \overline{d}^{A}(k)$ for every k, $1 \le k \le n$ follows from the fact that $\pi_S^{M*}\left(q_S^{M*}, q_{NS}^{M*}\right) > \pi_S^{A*}\left(q_S^{A*}, q_{NS}^{A*}\right)$ for every k, $1 \le k \le n$ and from the derivation of $\overline{d}^{M}$ and $\overline{d}^{A}(k)$.

Further details are reported in Online Appendix O.1.

A.2 Proof of Propositions 1

We define the following properties.

Consider an expansion of a coalition of size $k-1$ to k:

i) Positive Externality Property (PEP): $\pi_{NS}^{*}(k) > \pi_{NS}^{*}(k-1)$ for $1 < k \le n$.

ii) Positive Internalization Property (PIP): $\pi_{S}^{*}(k) > \pi_{S}^{*}(k-1)$ for $1 < k \le n$

iii) Superadditivity (SAD): $k \cdot \pi_{S}^{*}(k) > [k-1] \cdot \pi_{S}^{*}(k-1) + \pi_{NS}^{*}(k-1)$ for $1 < k \le n$.

iv) Welfare Cohesiveness (W-COH):

$$k \cdot \pi_{S}^{*}(k) + [n-k] \cdot \pi_{NS}^{*}(k) > [k-1] \cdot \pi_{S}^{*}(k-1) + [n-k+1] \cdot \pi_{NS}^{*}(k-1) \text{ for } 1 < k \le n.$$

v) Mitigation Cohesiveness (M-COH)

$$k \cdot q_{S}^{*}(k) + [n-k] \cdot q_{NS}^{*}(k) > [k-1] \cdot q_{S}^{*}(k-1) + [n-k+1] \cdot q_{NS}^{*}(k-1) \text{ for } 1 < k \le n.$$

If inequalities do not hold strictly, we say that the property holds only weakly.

Pure M-scenario (Case 1)

For analyzing PIP, PEP and M-COH, we can treat k as a continuous variable. We find $\frac{\partial \pi_S^M}{\partial k} = \frac{b^2(k-1)}{c} > 0$, $\frac{\partial \pi_{NS}^M}{\partial k} = \frac{b^2(2k-1)}{c} > 0$ and $\frac{\partial Q^M}{\partial k} = \frac{b(2k-1)}{c} > 0$ where for the first derivative we need to assume $k > 1$, as otherwise no signatory exists. Otherwise, $k \ge 1$.

Rearranging the SAD-condition above, SAD holds if $k \cdot \pi_S^M(k) - [k-1] \cdot \pi_S^M(k-1)$ $-\pi_{NS}^M(k-1) > 0$ is true. Substituting the appropriate payoffs, we find $\frac{b^2 k(k-1)}{2c} > 0$ for $k > 1$.

Finally, we note that SAD and PEP are sufficient that W-COH holds.

Pure G-scenario (Case 2)

We find $\frac{\partial \pi_S^G}{\partial k} = \frac{(bn-1)^2(k-1)}{cn^2} > 0$ for $k > 1$, $\frac{\partial \pi_{NS}^G}{\partial k} = \frac{(bn-1)^2(2k-1)}{cn^2} > 0$ and $\frac{\partial Q^G}{\partial k} = \frac{(bn-1)(2k-1)}{cn} > 0$ for $k \geq 1$. SAD holds, provided $k \cdot \pi_S^G(k) - [k-1] \cdot \pi_S^G(k-1)$ $-\pi_{NS}^G(k-1) > 0$ is true. We find $\frac{k(bn-1)^2(k-1)}{2cn^2} > 0$. W-COH follows from SAD and PEP.

Pure A-scenario (Case 3)

In the A-scenario, individual mitigation level of non-signatories and total mitigation do not change as the coalition size increases. Thus, M-COH holds only weakly. Individual mitigation of signatories decreases in k as there are more countries sharing the effort to achieve the avoidance equilibrium with $Q^{A*} = g$. Hence, π_S^A increases in k and PIP strictly holds. Non-signatories' payoff, π_{NS}^A, remains constant with size k of coalition K. Hence, PEP holds only weakly.

For SAD, we require $k \cdot \pi_S^A(k) - [k-1] \cdot \pi_S^A(k-1) - \pi_{NS}^A(k-1) > 0$ and find: $\frac{(bn-cg-1)^2}{2(k-1)ck} > 0.$ Finally, W-COH follows from SAD and PEP.

Mixed A-G-scenario (Case 4)

For coalition size $1,...,k-1$ we have the G-scenario and for coalition size $k,...,n$ we have the A-scenario. Hence, properties need only to be established for the move from $k-1$ to k. M-

COH is obvious because $Q^{A^*}(k) = g > Q^{G^*}(k) > Q^{G^*}(k-1)$. We have: $\pi_S^{A^*}(k) > \pi_S^{G^*}(k) > \pi_S^{G^*}(k-1)$ where the first inequality follows from the fact that the A- and not the G-equilibrium is played at k in this scenario and the second follows from PIP in the pure G-scenario. Hence, PIP holds in the A-G-scenario. By the same token $\pi_{NS}^{A^*}(k) > \pi_{NS}^{G^*}(k) > \pi_{NS}^{G^*}(k-1)$ where the first inequality follows from the fact that the A- and not the G-equilibrium is played at k and the second from PEP in the pure G-scenario. Hence, PEP holds in the A-G-scenario. SAD would be $k \cdot \pi_S^{A^*}(k) > [k-1] \cdot \pi_S^{*G}(k-1) + \pi_{NS}^{G^*}(k-1)$ and this holds because we have: $k \cdot \pi_S^{A^*}(k) > k \cdot \pi_S^{G^*}(k) > [k-1] \cdot \pi_S^{*G}(k-1) + \pi_{NS}^{G^*}(k-1)$ where the first inequality follows that the A- and not the G-equilibrium is played at k and the second inequality follows from SAD in the pure G-scenario. PEP and SAD give W-COH.

Mixed M-G-scenario (Case 5)

Again, we only need to establish properties for the move from $k-1$ to k. M-COH is obvious because $Q^{M^*}(k) \geq g > Q^{G^*}(k) > Q^{G^*}(k-1)$ where the first inequality follows from the fact that the M-equilibrium exists at k in this scenario and the second inequality has been established above in the pure G-scenario. We have: $\pi_S^{M^*}(k) > \pi_S^{G^*}(k) > \pi_S^{G^*}(k-1)$ where the first inequality follows from the fact that the M- and not the G-equilibrium is played at k in this scenario and the second from PIP in the pure G-scenario. Hence, PIP holds in the M-G-scenario. By the same token $\pi_{NS}^{M^*}(k) > \pi_{NS}^{G^*}(k) > \pi_{NS}^{G^*}(k-1)$ where the first inequality follows from the fact that the M- and not the G-equilibrium is played at k and the second from PEP in the pure G-scenario. Hence, PEP holds in the M-G-scenario. SAD would be $k \cdot \pi_S^{M^*}(k) > [k-1] \cdot \pi_S^{*G}(k-1) + \pi_{NS}^{G^*}(k-1)$ and this holds because we have: $k \cdot \pi_S^{M^*}(k) > k \cdot \pi_S^{G^*}(k) > [k-1] \cdot \pi_S^{*G}(k-1) + \pi_{NS}^{G^*}(k-1)$ where the first inequality follows

from the fact that the M- and not the G-equilibrium is played at k and the second inequality follows from SAD in the pure G-scenario. PEP and SAD give W-COH.

Mixed M-A-scenario (Case 6)

We only need to establish properties for the move from $k-1$ to k. M-COH must be true because $Q^{A*}(k-1)=Q^{A*}(k)=g\le Q^{M*}(k)$ by assumption if the M-equilibrium exists. PIP follows from $\pi_S^{M*}(k)>\pi_S^{A*}(k)>\pi_S^{A*}(k-1)$ where the first inequality is always true if the M-equilibrium exists and the second inequality follows from PIP in the pure A-scenario. PEP and SAD are readily proved. Hence, W-COH follows.

Further details are reported in Online Appendix O.2.

A.3 Proof of Proposition 2

In the M-scenario, internal stability holds if $\Omega^M(k)=\pi_S^{M*}(k)-\pi_{NS}^{M*}(k-1)\ge 0$ holds. We find:

$$\Omega^M(k)=-\frac{b^2(k-1)(k-3)}{2c}$$

which is strictly negative for $k>3$ and zero for $k=3$ and $k=1$, and strictly positive for $k=2$. Hence, $k=\{2,3\}$ (and larger coalitions) are externally stable, but not $k=1$. Thus, $k^*=\{2,3\}$. It can be easily checked that $k^*=3$ Pareto-dominates $k^*=2$.

In the G-scenario, we obtain for $\Omega^G(k)=\pi_S^{G*}(k)-\pi_{NS}^{G*}(k-1)$:

$$\Omega^G(k)=-\frac{(k-1)(k-3)(bn-1)^2}{2cn^2}$$

with exactly the same conclusion as above. Thus, $k^*=\{2,3\}$ where the larger coalition Pareto-dominates the smaller coalition.

In the A-scenario, $\Omega^{A}(k)=\pi_{S}^{A^*}(k)-\pi_{NS}^{A^*}(k-1)\geq 0$ must hold for internal stability. We know that $\pi_{NS}^{A}(k)>\pi_{S}^{A}(k)$ and $\pi_{NS}^{A}(k)=\pi_{NS}^{A}(k-1)$ hold. Consequently, $\Omega^{A}(k)=\pi_{S}^{A}(k)-\pi_{NS}^{A}(k-1)<0$ for all $k\geq 2$.

Details of all calculations can be found in Online Appendix O.3.

A.4 Proof of Proposition 3

In Case 4, we have $g>\frac{bn^2}{c}$ (Range 3) and $\overline{d}^{A}(k)\leq d<\overline{d}^{A}(k-1)$ (parts of Range B) from the feasibility conditions. We compute $\Omega^{A-G}(k)=\pi_{S}^{A^*}(k)-\pi_{NS}^{G^*}(k-1)$ where $\Omega^{A-G}(k)$ increases in the parameter d, with the threshold value of d for which $\Omega^{A-G}(k)=\pi_{S}^{A^*}(k)-\pi_{NS}^{G^*}(k-1)=0$ holds, given by

$$\hat{d}^{A-G}(k)=\frac{1}{2cnk^2(n-1)}(2b^2k^4n^2-6b^2k^3n^2+2n^3b^2k^2-2bcgk^2n^2+4b^2k^2n^2-2n^3b^2k+b^2n^4 \\ +2bcgkn^2-2bcgn^3-4nbk^4+c^2g^2n^2+12nbk^3-4bk^2n^2+2ncgk^2-8nbk^2 \\ +4bkn^2-2bn^3-2ncgk+2cgn^2+2k^4-6k^3+2nk^2+4k^2-2nk+n^2)$$

Hence, if $d\geq\hat{d}^{A-G}(k)$, internal stability holds and if the reverse is true, $d<\hat{d}^{A-G}(k)$, it does not hold.

Let $k\leq 3$. One can show that $\hat{d}^{A-G}(k)\leq\overline{d}^{A}(k)$ for $k\leq 3$. Recall that $d\geq\overline{d}^{A}(k)$ holds in Case 4. If at $k=3$ stability strictly holds, $k-1$ is not externally stable. If it weakly holds, this implies $\pi_{NS}^{A^*}(3)=\pi_{S}^{G^*}(2)$. Moreover, $\pi_{S}^{A^*}(3)\geq\pi_{S}^{G^*}(3)>\pi_{S}^{G^*}(2)$ where the first weak inequality follows from $d\geq\overline{d}^{A}(k)$ and the second from the property PIP (see Appendix 2). At $k=2$ stability always strictly holds and $k-1$ is not externally stable. Hence, for $k=\{2,3\}$, $k^*=k$ is the unique equilibrium.

Let $k > 3$. If $k \neq n$, all coalitions strictly larger than k are not internally stable from the proof of Proposition 2 in Appendix A.3. Consequently, k is externally stable. (If $k = n$, the grand coalition is externally stable by definition.) One can show that $\hat{d}^{A-G}(k) \geq \overline{d}^{A}(k)$ holds. Additionally, $\hat{d}^{A-G}(k) < \overline{d}^{A}(k-1)$. Hence, stability holds if $\overline{d}^{A}(k) \leq \hat{d}^{A-G}(k) < \overline{d}^{A}(k-1)$.

Apart from this stable coalition, there is a second smaller stable coalition with the G-equilibrium implemented at $k^* = 3$, which follows from Proposition 2. Hence, we have to show that the first equilibrium, implementing the A-equilibrium at $k^* = \hat{k}$ Pareto-dominates the second equilibrium, implementing the G-equilibrium at $k^* = \breve{k} = 3$. For signatories, we have $\pi_S^{A*}(\hat{k}) \geq \pi_S^{G*}(\hat{k})$ because $d \geq \overline{d}^{A}(\hat{k})$. Furthermore, from Appendix 2, we have $\pi_S^{G*}(\hat{k}) > \pi_S^{G*}(\breve{k})$ due to the property PIP. Hence, $\pi_S^{A*}(\hat{k}) > \pi_S^{G*}(\breve{k})$. From internal stability at $k = \hat{k}$, we have: $\pi_S^{A}(\hat{k}) \geq \pi_{NS}^{G}(\hat{k}-1)$. Furthermore, $\pi_{NS}^{A}(\hat{k}) > \pi_{NS}^{G}(\hat{k}-1) > \pi_{NS}^{G}(\breve{k})$ from the property PEP in Appendix 2. In case internal stability does not hold for coalition $k^* = \hat{k}$, then $k^* = \breve{k} = 3$.

Further details are reported in Online Appendix O.3.

A.5 Proof of Proposition 4

In Case 5, we compute $\Omega^{M-G}(k) = \pi_S^{M*}(k) - \pi_{NS}^{G*}(k-1)$. It can be shown that $\Omega^{M-G}(k)$ increases in d with the threshold value $\hat{d}^{M-G}(k)$ for which $\Omega^{M-G}(k) = 0$. We find:

$$\hat{d}^{M-G}(k) = \frac{bk^2 - 4b^2k - 4bk^2 + 3b^2 + 12bk - 4bn + 2cg + 2k^2 - 6b - 6k + 2n + 3}{2c(n-1)}.$$

Let $k \leq 3$. Then, $\hat{d}^{M-G}(k) < \overline{d}^{M}(k)$ holds. Hence, in both Cases 5i) and 5ii), $k = \{2,3\}$ are strictly internally stable and $k^* = k$, is the unique stable coalition size.

Let $k > 3$. Then it can be shown that $\hat{d}^{M-G}\left(\hat{k}\right) \geq \overline{d}^{M}\left(\hat{k}\right)$ holds. Additionally, $\hat{d}^{A-G}(k) < \overline{d}^{M}(k-1) < \overline{d}^{A}(k-1)$. Hence, $\overline{d}^{M}(k) \leq \hat{d}^{M-G}(k) < \overline{d}^{M}(k-1) < \overline{d}^{A}(k-1)$. Stability of coalition k is achieved if $\hat{d}^{M-G}(k) \leq d < \overline{d}^{M}(k-1)$ in Case 5i) and if $\hat{d}^{M-G}(k) \leq d < \overline{d}^{A}(k-1)$ in Case 5ii).

If $k > 3$, implementing the A-equilibrium, is stable, then there is a second smaller stable coalition which implements the G-equilibrium at $k^{*} = 3$, as emerges from Proposition 2. In the spirit of Appendix A.4 for Case 4 it can be shown that this smaller coalition is Pareto-dominated by the larger coalition. Only if internal stability does not hold for $k > 3$, this G-equilibrium will emerge.

In Case 6, internal stability requires that $\Omega^{M-A}(k) = \pi_{S}^{M^{*}}(k) - \pi_{NS}^{A^{*}}(k-1) \geq 0$ holds. It can be shown that $\Omega^{M-A}(k)$ decreases in the geoengineering benefit parameter g. Hence, internal stability holds if

$$g \leq \hat{g}^{M-A}(k) = \frac{b^{2}k^{2} - 2b^{2}k + 2b^{2}n + b^{2} - 2b + 1}{2bc}$$

holds. Substituting the lowest possible value for g which is permissible in Case 6, i.e., $g = \frac{b}{c}(k-1)$, we find that $g > \hat{g}^{M-A}(k)$ always holds. Hence, k is not internally stable.

Let $k \geq 3$. No larger and no smaller coalition than k which implements the A-equilibrium is internally stable, as we know from Proposition 2. Let $k < 3$. $k = 2$ can also not be internally stable, as we have shown above. However, with $k < 3$, the M-equilibrium is played for all coalitions larger than k and from Proposition 2 we know that $k^{*} = 3$ with the M-equilibrium emerging.

Further details are provided in Online Appendix O.3.

A.6 Proof of Proposition 5

Consider the pure M-scenario (Case 1). In the free-riding phase, signatories keep the cooperative mitigation level $q_s^{M^*}(n)=\frac{nb}{c}$ whereas the free-rider can choose either its best-response for the M-equilibrium, $q_{NS}^{M^*}=\frac{b}{c}$, or its best-response for the G-equilibrium, $q_{NS}^{G^*}$. Suppose the free-rider chooses $q_{NS}^{M^*}=\frac{b}{c}$. This deviation leads to total mitigation $\frac{\tilde{b}}{c}(n-1):=\tilde{Q}^{M^*}(n-1)=[n-1]\cdot q_S^{M^*}(n)+q_S^{M^*}=\frac{b(n^2-n+1)}{c}$. It is straightforward to show that $\frac{\tilde{b}}{c}(n-1)>\frac{b}{c}(n-1):=Q^{M^*}(n-1)=[n-1]\cdot q_S^{M^*}(n-1)+q_S^{M^*}$ as individual mitigation of signatories $q_S^{M^*}(k)=\frac{kb}{c}$ increases in the coalition size k. Hence, we can establish $\frac{b}{c}(1)<\frac{b}{c}(n-1)<\frac{\tilde{b}}{c}(n-1)<\frac{b}{c}(n)$. If $\frac{\tilde{b}}{c}(n-1)\geq g$, then $q_{NS}^{M^*}$ represents the optimal deviation of the free-rider and geoengineering is not deployed. If $\frac{\tilde{b}}{c}(n-1)<g$, then the free-rider deviates by choosing $q_{NS}^{G^*}$ and geoengineering is deployed.

In Case 1, we consider Range 1 with respect to parameter g, i.e., $g\leq\frac{b}{c}(1)$. Hence, we can establish $g\leq\frac{b}{c}(1)<\frac{b}{c}(n-1)<\frac{\tilde{b}}{c}(n-1)<\frac{b}{c}(n)$. Thus, the pure M-scenario delivers sufficient mitigation to avoid the deployment of geoengineering in all the three phases of the repeated game. Inserting payoff levels $\pi_i^C=\pi_S^{M^*}(n)$, $\pi_i^P=\pi_{NS}^{M^*}(1)$ and $\pi_i^F=\tilde{\pi}_i^{M^*}\left(q_S^{M^*}(n),q_{NS}^{M^*}\right)$ in equation (8) gives $\delta_{min}^M=\frac{1}{2}$.

In the pure G-scenario (Case 2), inserting payoff levels $\pi_i^C = \pi_S^{G*}(n)$, $\pi_i^P = \pi_{NS}^{G*}(1)$ and $\pi_i^F = \tilde{\pi}_i^{G*}\left(q_S^{G*}(n), q_{NS}^{G*}\right)$ in equation (8) gives $\delta_{min}^G = \frac{1}{2}$.

In the pure A-scenario (case 3), given that the free-rider has already reduced its mitigation level in the previous phase, in the punishment phase it will be one of the other former signatories to provide the mitigation level which is needed to achieve $Q^{A*} = g$. In the punishment phase the free-rider contributes $q_{NS}^{G*} < q_S^{A*}(n) = \frac{g}{n}$ while total mitigation remains the same as in the cooperative phase. Hence, it follows that $\pi_i^P = \pi_{NS}^{A*}(1) > \pi_i^C$. The punishment is not effective to deter free-riding and the grand coalition cannot be stable.

Further details are provided in Online Appendix O.5.

A.7 Proof of Proposition 6

Table A displays the parameter ranges for which the three mixed policy scenarios arise in the repeated game.

Table A about here

Mixed A-G-scenario

In the A-G-scenario, the range to be considered is Range 3, $g > \frac{bn^2}{c}$, and Range B, $\overline{d}^A(n) \le d < \overline{d}^A(1)$. These conditions guarantee that the A-equilibrium is played in the grand coalition while the G-equilibrium is played in non-cooperation equilibrium (punishment phase). In the free-riding phase, total mitigation drops below g and geoengineering is deployed. We have $\pi_i^C = \pi_S^{A*}(n)$, $\pi_i^P = \pi_{NS}^{G*}(1)$ and $\pi_i^F = \tilde{\pi}_i^{G*}\left(q_S^{A*}(n), q_{NS}^{G*}\right)$. Inserting these payoffs in equation (8) gives $\delta_{min}^{A-G} = -\frac{b^2n^2 - 2bcgn + c^2g^2 - 2dcn^2 + 2dcn - 2bn + 2gc + 1}{2(n-1)(bn-1)(bn-gc-1)}$.

Differentiating δ_{min}^{A-G} with respect to the geoengineering collateral damages parameter d, we find $\frac{\partial \delta_{min}^{A-G}}{\partial d} = \frac{cn}{(bn-cg-1)(bn-1)} < 0$, as the first bracket in the denominator is negative due to condition C_2 and the second bracket in the denominator is positive due condition C_1. Hence, δ_{min}^{A-G} decreases in d.

We solve for $\delta_{min}^{A-G} = \frac{1}{2}$ for d in order to find the threshold value $\tilde{d}^{A-G}$:

$$\tilde{d}^{A-G} = \frac{1}{2}\frac{\left(bn^2-cg-n\right)(bn-cg-1)}{(n-1)nc}.$$

We solve for $\delta_{min}^{A-G} = 0$ for d in order to find the threshold value $\approx\!\tilde{d}^{A-G}$:

$$\approx\!\tilde{d}^{A-G} = \frac{(bn-gc-1)^2}{2cn(n-1)}.$$

We compare the two threshold values with the upper and lower bound of the collateral damage parameter d in the A-G-scenario, $\bar{d}^{A}(n) \le d < \bar{d}^{A}(1)$. We find:

$$\tilde{d}^{A-G} - \bar{d}^{A}(n) = -\frac{1}{2}\frac{(bn-1)\left(bn^2-cg-n\right)}{nc} > 0 \text{ and}$$

$$\bar{d}^{A}(1) - \approx\!\tilde{d}^{A-G} = \frac{(n+1)(bn-gc-1)^2}{2nc} > 0.$$

For the first difference, the first bracket in the nominator is positive by condition C_1, while the second bracket is negative by condition C_2. Hence, we established $\bar{d}^{A}(n) < \tilde{d}^{A-G} < \approx\!\tilde{d}^{A-G} < \bar{d}^{A}(1)$.

Mixed M-G-scenario

Case 5a

For Case 5a, we have $\frac{n(bn-1)}{c} < g \leq \frac{\tilde{b}}{c}(n-1)$. In subcase i) $\overline{d}^{M}(n) \leq d < \overline{d}^{M}(1)$ and $g \leq \frac{b}{c}(1)$. In subcase ii) $\overline{d}^{M}(n) \leq d < \overline{d}^{A}(1)$ and $g > \frac{b}{c}(1)$. The M-equilibrium is played in the grand coalition, the G-equilibrium is played in non-cooperation (punishment) and in the free-riding phase geoengineering is not deployed. We have $\pi_i^C = \pi_S^{M*}(n)$, $\pi_i^P = \pi_{NS}^{G*}(1)$ and $\pi_i^F = \tilde{\pi}_i^{M*}\left(q_S^{M*}(n), q_{NS}^{M*}\right)$. Inserting these payoffs in equation (8) gives $\delta_{min}^{M-G}(a) = \frac{b^2 n^2 (n-1)^2}{2b^2n^4 - 4b^2n^3 + 2b^2n^2 + 2cdn^2 + 4bn^2 - 2cdn - 2cgn - 2bn - 2n + 1}$. Differentiating $\delta_{min}^{M-G}(a)$ with respect to the geoengineering collateral damages parameter d, we find $\frac{\partial \delta_{min}^{M-G}(a)}{\partial d} = -\frac{2b^2cn^3(n-1)^3}{\left(2b^2n^4 - 4b^2n^3 + 2b^2n^2 + 2cdn^2 + 4bn^2 - 2cdn - 2cgn - 2bn - 2n + 1\right)^2} < 0$.

Hence, $\delta_{min}^{A-G}(a)$ decreases in d.

It is straightforward to show that $\delta_{min}^{A-G}(a) > 0$ as $\pi_i^F = \tilde{\pi}_i^{M*}\left(q_S^{M*}(n), q_{NS}^{M*}\right) > \pi_i^C = \pi_S^{M*}(n) > \pi_i^P = \pi_{NS}^{G*}(1)$ in this scenario.

We solve for $\delta_{min}^{M-G}(a) = \frac{1}{2}$ for d in order to find the threshold value $\tilde{d}^{M-G}(a)$:

$$\tilde{d}^{M-G}(a) = -\frac{1}{2}\frac{4bn^2 - 2cgn - 2bn - 2n + 1}{(n-1)nc}.$$

Now, we compare this threshold value with the upper and lower bound of the collateral damage parameter d in the M-G-scenario. In Case 5a, (i) $\overline{d}^{M}(n) \leq d < \overline{d}^{M}(1)$ and (ii) $\overline{d}^{M}(n) \leq d < \overline{d}^{A}(1)$. We find:

$$\tilde{d}^{M-G}(a)-\overline{d}^{M}(n)=\frac{1}{2}\frac{(n-1)(2bn-1)}{nc}>0 \text{ and}$$

$$\overline{d}^{M}(1)-\tilde{\tilde{d}}^{M-G}(a)=0.$$

The first difference is positive as the second bracket in the nominator is positive by condition C_1. Hence, we established $\overline{d}^{M}(n)<\tilde{d}^{M-G}(a)=\overline{d}^{M}(1)<\overline{d}^{A}(1)$, where the last inequality has been proved in Appendix A.1.

Case 5b

For Case 5b, we have $\frac{\tilde{b}}{c}(n-1)<g\leq\frac{b}{c}(n)$ and $\overline{d}^{M}(n)\leq d<\overline{d}^{A}(1)$. The M-equilibrium is played in the grand coalition, the G-equilibrium is played in non-cooperation (punishment) and in the free-riding phase geoengineering is deployed. We have $\pi_i^C=\pi_S^{M*}(n)$, $\pi_i^P=\pi_{NS}^{G*}(1)$ and $\pi_i^F=\tilde{\tilde{\pi}}_i^{G*}\left(q_S^{M*}(n),q_{NS}^{G*}\right)$. Inserting these payoffs in equation (8). gives $\delta_{min}^{M-G}(b)=\frac{b^2n^4-2b^2n^3+b^2n^2-2bn^3-2cdn^2+2bn^2+2cdn+2cgn-2bn+1}{2(n-1)(bn-1)\left(bn^2-bn-1\right)}$. Differentiating $\delta_{min}^{M-G}(b)$ with respect to the geoengineering collateral damages parameter d, we find $\frac{\partial\delta_{min}^{M-G}(b)}{\partial d}=-\frac{cn}{\left(bn^2-bn+1\right)(bn-1)}<0$, as the first and second bracket in the denominator is positive by condition C_1. Hence, $\delta_{min}^{M-G}(b)$ decreases in d.

We solve for $\delta_{min}^{M-Gb}=\frac{1}{2}$ for d in order to find the threshold value $\tilde{d}^{M-G}(b)$:

$$\tilde{d}^{M-G}(b)=-\frac{1}{2}\frac{bn^2+bn-2cg-1}{(n-1)nc}.$$

We solve for $\delta_{min}^{M-G}(b)=0$ for d in order to find the threshold value $\tilde{\tilde{d}}^{M-G}(b)$:

$$\approx^{M-G}_{d}(b)=\frac{1}{2}\frac{b^2n^4-2b^2n^3+b^2n^2-2bn^3+2bn^2+2cgn-2bn+1}{cn(n-1)}\ .$$

Now, we compare the two threshold values with the upper and lower bound of the collateral damage parameter d in the M-G-scenario Case 5b, $\bar{d}^{M}(n)\le d<\bar{d}^{A}(1)$. We find:

$$\tilde{d}^{M-G}(b)-\bar{d}^{M}(n)=\frac{1}{2}\frac{(bn-1)}{c}>0 \text{ and}$$

$$\bar{d}^{M}(1)-\approx^{M-G}_{d}(b)=-\frac{b^2n^3-b^2n^2-2bn^2+4bn-2}{2nc}>0\,.$$

The first difference is positive by condition C_1. In the second difference, the sign depends on the sign of nominator. One can show that the nominator is a convex function of the mitigation benefit b and that it is negative for both the lower and the upper bound of parameter b, given by conditions C_1 and C_3. Hence, the nominator is negative and the difference is positive. Hence, we established $\bar{d}^{M}(n)<\tilde{d}^{M-G}(b)<\approx^{M-G}_{d}(b)<\bar{d}^{M}(1)<\bar{d}^{A}(1)$, where the last inequality has been proved in Appendix A.1. Details are reported in the Online Appendix O.5.

Mixed M-A-scenario

In the M-A-scenario, in both sub-cases 6a and 6b, the M-equilibrium is played in the cooperation phase giving $\pi_i^C=\pi_S^{M^*}(n)$ and the A-equilibrium is played in the punishment phase giving $\pi_i^P=\pi_{NS}^{A^*}(1)$. We show that the punishment cannot be effective as $\pi_i^P>\pi_i^M(n)$. For this, we need to prove $\pi_S^{M^*}(n)-\pi_{NS}^{A^*}(1)<0$. Recall that in the A-scenario the payoff of non-signatories does not change with the coalition size. Hence, $\pi_{NS}^{A^*}(n-1)=\pi_{NS}^{A^*}(1)$. Thus, it suffices to show: $\pi_S^{M^*}(n)-\pi_{NS}^{A^*}(n-1)<0$. In the cartel formation game, we established $\Omega^{M-A}(k)=\pi_S^{M^*}(k)-\pi_{NS}^{A^*}(k-1)<0$, by noticing that $\Omega^{M-A}(k)$ decreases in the parameter g (see Appendix A.5) with $\Omega^{M-A}(k)<0$ for $g>\hat{g}^{M-A}(k)$. For Case 6 of the repeated game, the

lowest possible value for g is $g = \frac{b}{c}(1)$ which is larger than $\hat{g}^{M-A}(k)$. Hence, $\Omega^{M-A}(n) < 0$, $\pi_i^P > \pi_i^M$ and the punishment is not effective.

Further details are reported in the Online Appendix O.5.

Figure 1: Geoengineering Net Benefit and Collateral Damage Parameter Space

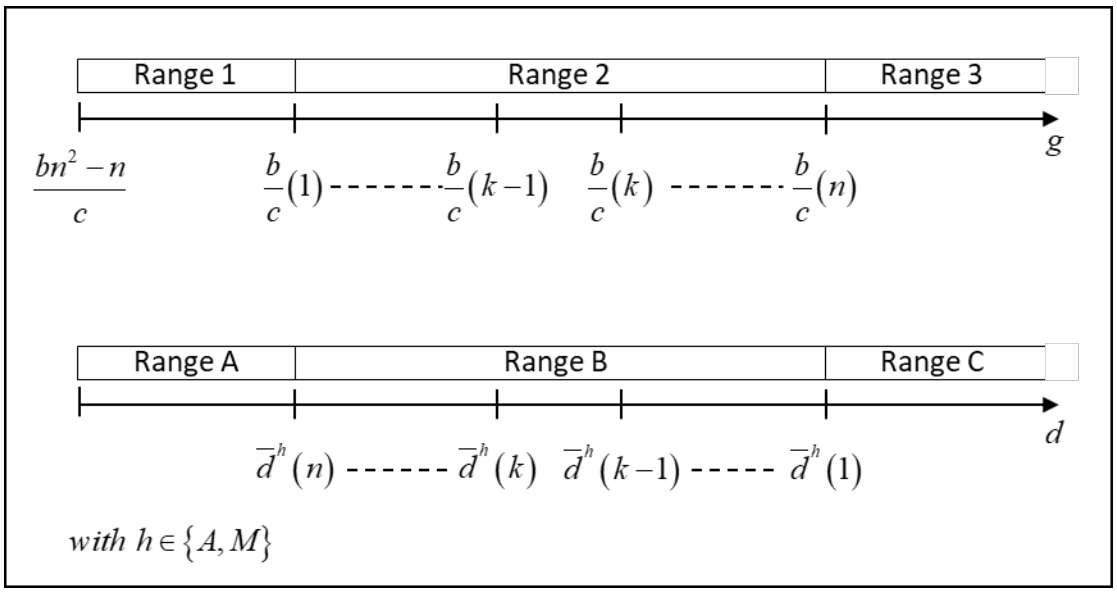

Figure 2: Policy Scenarios

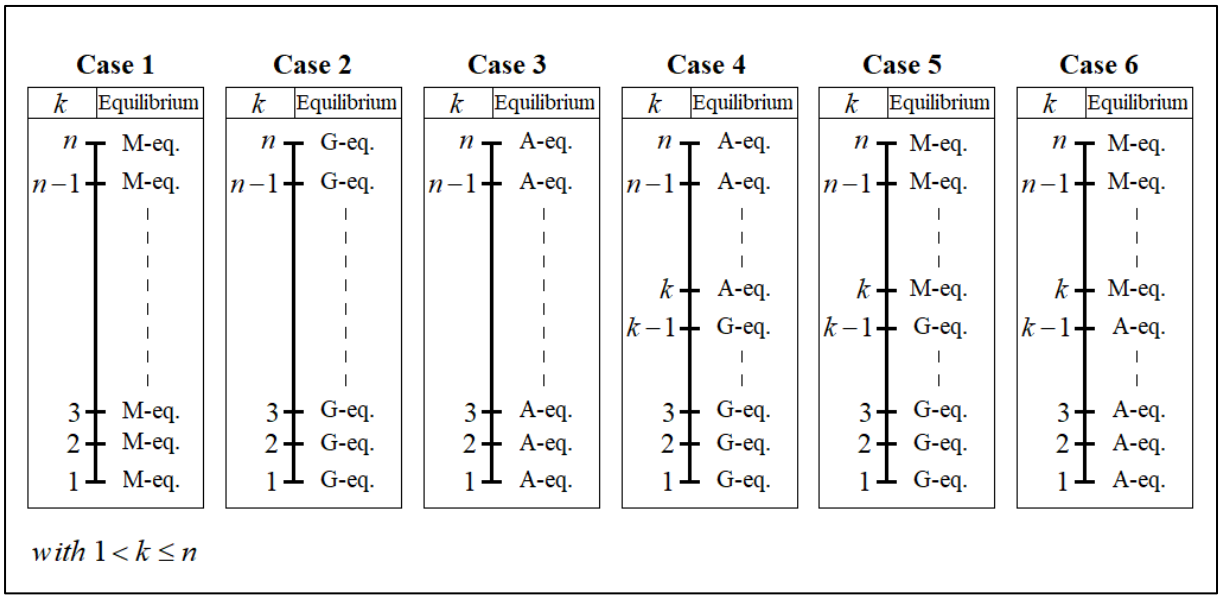

Figure 3: Pure Policy Scenarios (Case 1, 2 and 3)

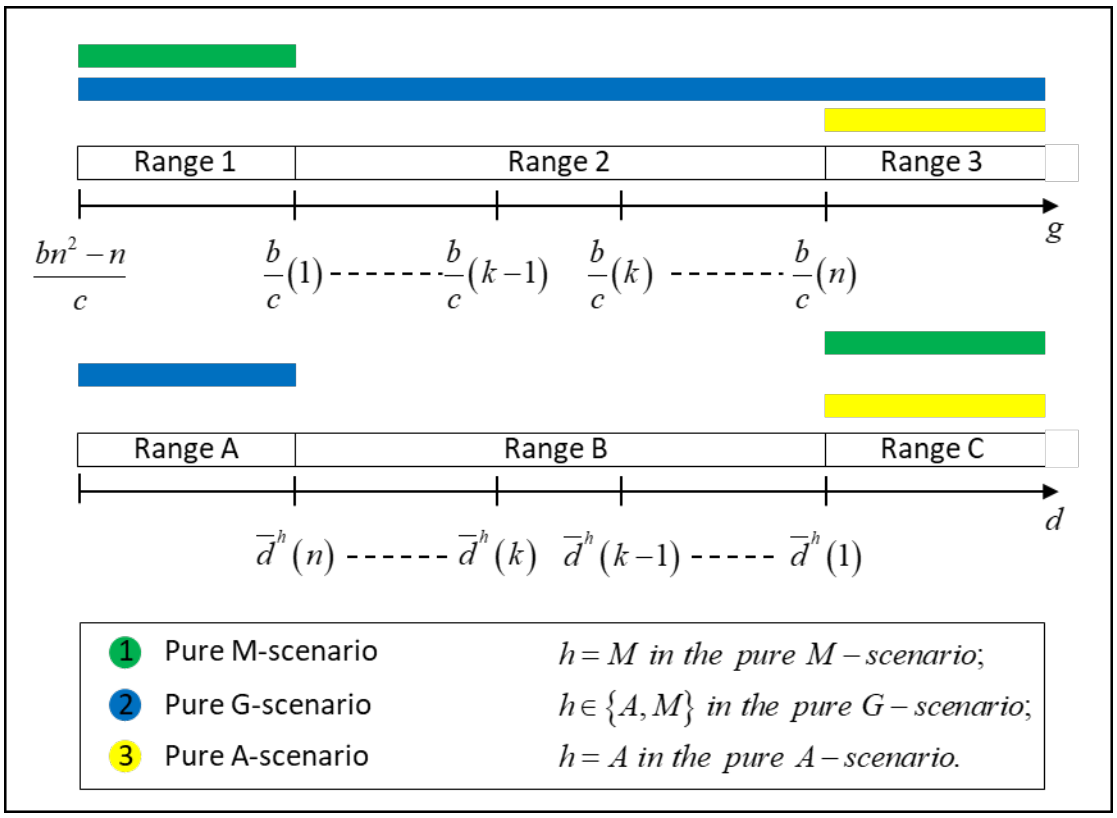

Figure 4: Mixed Policy Scenarios (Case 4, 5 and 6)

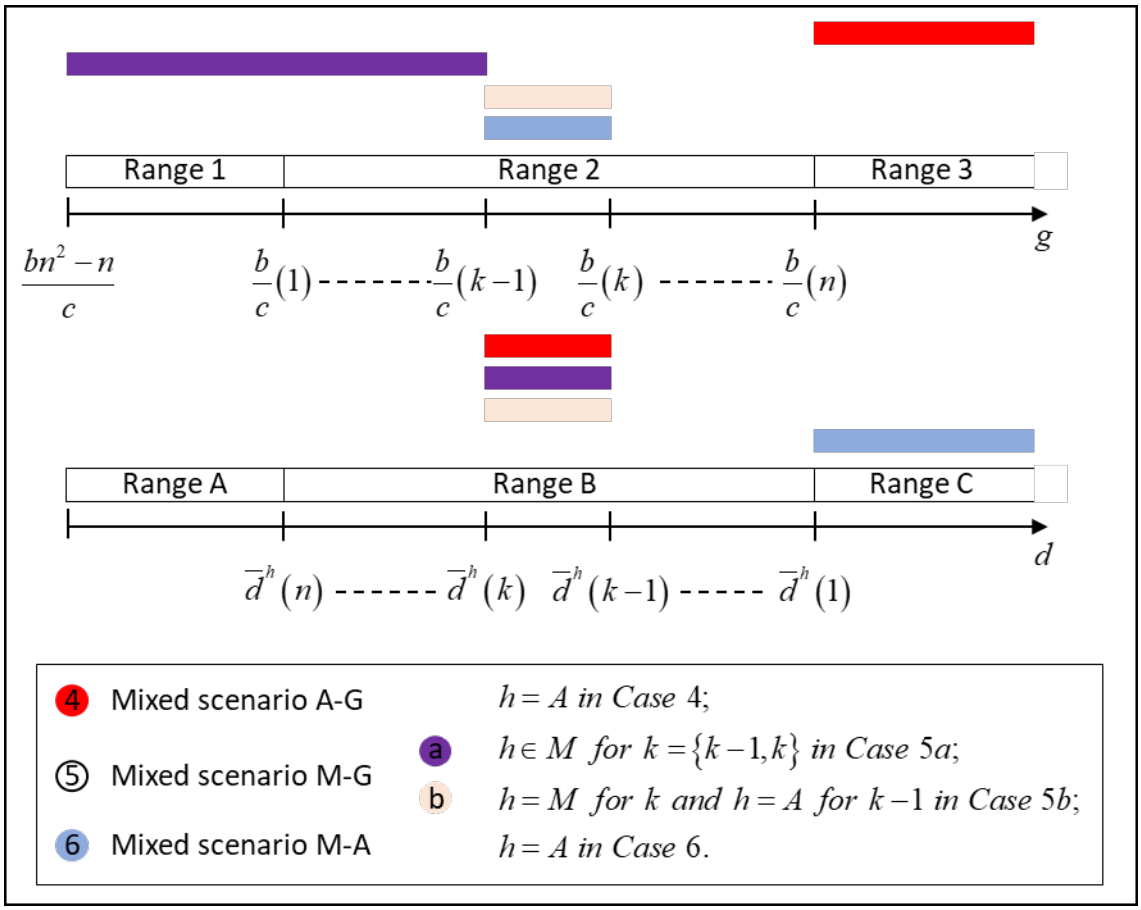

Figure 5: Mixed Policy Scenarios in the Repeated Game (Case 4, 5 and 6)

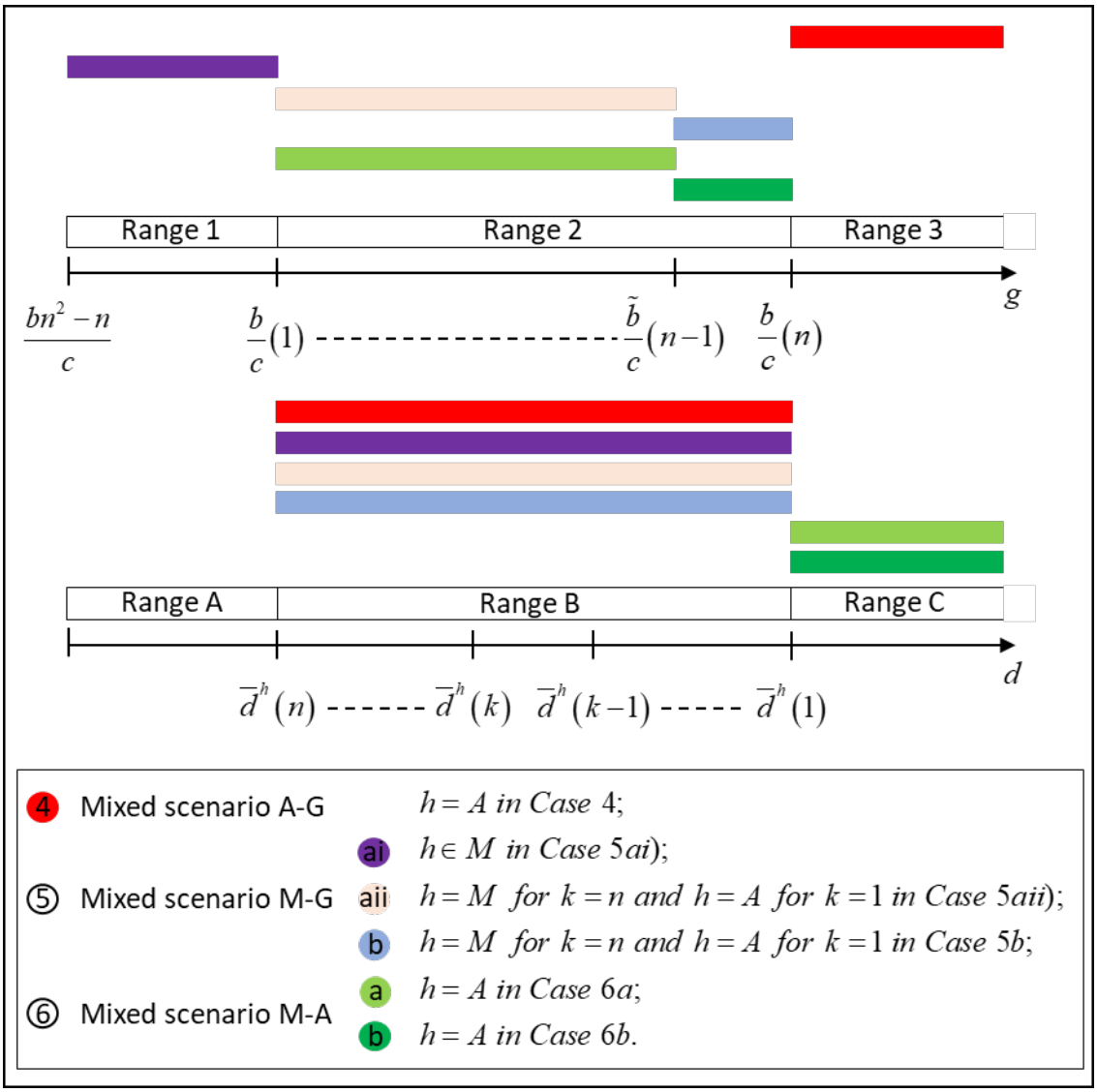

Table 1: Policy Scenarios

Case	Policy Scenario	Parameter Range	Description
Case 1	Pure M-scenario	Range 1 & Range C with $h = M$	Deployment of geoengineering is not a problem; the M-eq. produces sufficiently high global mitigation, even for $k = 1$, such that geoengineering does not pay; the benefits of geoengineering are very low. Collateral damages are very high, making geoengineering never attractive.
Case 2	Pure G-scenario	Range 1, 2 and 3 & Range A with $h \in \{M, A\}$	Deployment of geoengineering is not a problem; abstaining from geoengineering is never rational, as collateral damages are very low.
Case 3	Pure A-scenario	Range 3 & Range C with $h = A$	Deployment of geoengineering is not a problem; even though the "second-best" A-eq. is the only option to avoid the deployment of geoengineering which generate high net benefits, collateral damages are so high as to make geoengineering never attractive.
Case 4	Mixed A-G-scenario	Range 3 & parts of Range B with $\bar{d}^A(k) \le d < \bar{d}^A(k-1)$	Geoengineering generates high net benefits (large values g); collateral damages are not too high as to make geoengineering never attractive for smaller coalitions (intermediate values d).
Case 5	Mixed M-G-scenario	a) Range 1 or parts of Range 2 with $g \le \frac{b}{c}(k-1)$ & parts of Range B with $\bar{d}^M(k) \le d < \bar{d}^M(k-1)$ b) Parts of Range 2 with $\frac{b}{c}(k-1) < g \le \frac{b}{c}(k)$ & parts of Range B with $\bar{d}^M(k) \le d < \bar{d}^A(k-1)$	Geoengineering generates low or moderate net benefits (low or intermediate values g); collateral damages are not too high as to make geoengineering never attractive for smaller coalitions (intermediate values d).
Case 6	Mixed M-A-scenario	Parts of Range 2 with $\frac{b}{c}(k-1) < g \le \frac{b}{c}(k)$ & Range C with $h = A$	Geoengineering generates moderate net benefits (intermediate values g); collateral damages are very high, making geoengineering never attractive (large values d).

Table A: Mixed Policy Scenarios in the Repeated Game

	Policy Scenario	**Parameters Range**
Case 4	A-G deviation leads to G-equilibrium	Range 3; entire Range B with $\bar{d}^{A}(n) \le d < \bar{d}^{A}(1)$
Case 5	M-G a) deviation leads to M-equilibrium b) deviation leads to G-equilibrium	a) i) Range 1 & entire Range B with $\bar{d}^{M}(n) \le d < \bar{d}^{M}(1)$ ii) Parts of Range 2 with $\frac{b}{c}(1) < g \le \frac{\tilde{b}}{c}(n-1)$; entire Range B with $\bar{d}^{M}(n) \le d < \bar{d}^{A}(1)$ b) Parts of Range 2 with $\frac{\tilde{b}}{c}(n-1) < g \le \frac{b}{c}(n)$; entire Range B with $\bar{d}^{M}(n) \le d < \bar{d}^{A}(1)$
Case 6	M-A a) deviation leads to M-equilibrium b) deviation leads to G-equilibrium	a) Parts of Range 2 with $\frac{b}{c}(1) < g \le \frac{\tilde{b}}{c}(n-1)$; Range C $(h = A)$ b) Parts of Range 2 with $\frac{\tilde{b}}{c}(n-1) < g \le \frac{b}{c}(n)$; Range C $(h = A)$