Finus, Michael; Furini, Francesco

On the credibility of threats to avoid the deployment of solar geoengineering

Environmental Economics and Policy Studies

Provided in Cooperation with:
Springer Nature

Suggested Citation: Finus, Michael; Furini, Francesco (2024) : On the credibility of threats to avoid the deployment of solar geoengineering, Environmental Economics and Policy Studies, ISSN 1867-383X, Springer Japan, Tokyo, Vol. 27, Iss. 1, pp. 1-21, https://doi.org/10.1007/s10018-024-00407-2

This Version is available at:
https://hdl.handle.net/10419/315002

Environmental Economics and Policy Studies (2025) 27:1–21
https://doi.org/10.1007/s10018-024-00407-2

RESEARCH ARTICLE

On the credibility of threats to avoid the deployment of solar geoengineering

Michael Finus[1] · **Francesco Furini**[2]

Received: 14 November 2023 / Accepted: 19 June 2024 / Published online: 17 July 2024

Abstract

We analyze how geoengineering in the form of solar radiation management (SRM), associated with the potential of high collateral damages, affects the governance architecture of climate agreements. We investigate under which conditions signatories to a climate agreement can avoid the deployment of SRM and implement a climate agreement on mitigation. We show that a climate agreement with all countries can be stable with the threat to deploy SRM in case a country free-rides. The threat is deterrent if collateral damages are perceived to be sufficiently high (lower threshold), but only credible if those damages are not too high (upper threshold). SRM deployment is the only threat available to signatories if they choose mitigation levels simultaneously with non-signatories (Nash–Cournot scenario). However, if signatories choose mitigation levels before non-signatories (Stackelberg scenario), an additional punishment option arises. Then if collateral damages are sufficiently large, signatories can reduce their mitigation levels and impose a heavier burden on non-signatories that would find it profitable to avoid the deployment of SRM. We show that our results are robust in two analytical frameworks frequently employed in the game-theoretic analysis of international environmental agreements.

Keywords Coalition stability · Mitigation · Solar radiation management · Collateral damages · Credibility

JEL Classification D71 · D74 · H41 · Q54

✉ Francesco Furini
francesco.furini@uni-hamburg.de

[1] Department of Economics, Karl-Franzens-Universität Graz, Universitätsplatz 3, 8010 Graz, Austria

[2] Department of Socioeconomics, Universität Hamburg, Welckerstraße 8, 20354 Hamburg, Germany

1 Introduction

Due to slow progress in reducing carbon emissions, solar radiation management (SRM), as one form of geoengineering, has been proposed as an option to address global warming.[1] SRM aims at increasing the Earth's reflectivity to attenuate the effect of incoming solar radiation, thereby cooling down the planet. Among different SRM technologies, stratospheric sulfur aerosol injection is the primary technology currently discussed (Caldeira et al. 2013).[2] The effectiveness of aerosol injection in reducing temperatures has been observed after volcanic eruptions. For instance, the injection of about 20 mega tonnes of sulfur dioxide into the tropical stratosphere due to the eruption of Mount Pinatubo in 1991 provoked a global cooling effect of about a half degree Celsius in the following years (Crutzen 2006). Following the same logic, SRM offers an apparently fast and pragmatic solution to moderate the effects of human-induced global warming over the coming few decades. The costs of deployment are low—especially if compared to mitigation costs—and the potential benefits are estimated to be high (Aldy and Zeckhauser 2020). However, it also comes with a risk, as collateral damages may be very high (Aldy et al. 2021; Barrett 2014; Barrett et al. 2014; Bodansky 2013; Reynolds 2019; Robock et al. 2009; Stephens et al. 2021). SRM will have impacts on precipitation patterns and might increase the exposure to droughts in many regions (Haywood et al. 2013; Irvine et al. 2019; Kravitz et al. 2014; Ricke et al. 2010). The risk of the depletion of stratospheric ozone will increase (Moreno-Cruz and Keith 2013) and the deposition of sulfur particles could lead to acid rain and soil acidification with detrimental consequences for agriculture and health (Crutzen 2006; Kravitz et al. 2009; Visioni et al. 2020). Finally, SRM cannot easily be reversed as its interruption could cause a termination shock with sudden temperature increase (Irvine et al. 2012; Parker and Irvine 2018). Given these features, SRM poses several governance challenges. On the one hand, countries which are most exposed to climate damages may "free-drive" by individually deploying SRM, with a high risk for the rest of the world (Weitzman 2015). Thus, the issue may not be the underprovision of the public good "mitigation" (i.e., emissions reduction), but the overprovision of SRM. On the other hand, without SRM, the notorious failure of substantial mitigation efforts in the past may lead to an unprecedented warming of the planet that seriously threatens life and the environment.

Governance issues of geoengineering (in the form of SRM) have been analyzed for instance by Weitzman (2015), Rickels et al. (2020) and Ricke et al. (2013) who assume that only members to an agreement can set the level of the global thermostat through SRM, which incentivizes countries to join the agreement. However, the assumption that non-members can be prevented from the use of SRM is defining

[1] The term geoengineering refers to any action deliberately modifying the Earth's climate to counteract the effects of climate change. Apart from SRM, the other main geoengineering option is carbon dioxide removal (CDR), aiming at capturing CO_2 from the atmosphere and depositing it in biomass or underground (Shepherd et al. 2009).

[2] Alternative proposals include marine cloud brightening and the use of reflective particles to increase the longevity of sea ice. These options increase the reflectivity of clouds and sea ice respectively, reflecting back solar radiation (National Academies of Sciences, Engineering and Medicine 2021).

away the very problem of free-driving, which has been identified by these authors as the major governance issue in the first place. Moreover, Parker et al. (2018) and Heyen et al. (2019) consider the possibility for counter-geoengineering, i.e., measures capable of negating the climate effects of SRM.

All of these papers focus exclusively on SRM, but do not consider the interaction with other climate policies, like mitigation. For instance, as SRM represents a relatively inexpensive alternative to mitigation, one would expect that SRM reduces mitigation efforts by countries. However, as shown by Cherry et al. (2023), Moreno-Cruz (2015) and Urpelainen (2012), due to the possibility of high collateral damages, countries could increase their mitigation efforts in order to reduce or avoid the deployment of SRM technologies. How the interaction between mitigation and SRM could influence the formation of climate agreements is not studied in these papers.

Millard-Ball (2012) proposes a coalition formation model which studies not only the strategic interaction between SRM and mitigation, but also captures the free-rider incentive which hampers the stability of large climate agreements on mitigation. Two main results may be highlighted. First, if collateral damages from SRM are perceived to be sufficiently high, countries will have an incentive to increase total mitigation up to a level at which the use of SRM is no longer attractive. Second, sufficiently high collateral damages can stabilize the grand coalition if signatories behave as Stackelberg leaders. A country leaving the grand coalition is faced with the option to increase its mitigation efforts in order to avoid the deployment of SRM because Stackelberg leaders reduce their mitigation efforts. This is an unattractive option for the free-rider, and works as a credible threat provided collateral damages from SRM are sufficiently high.

Despite Millard-Ball's model provides an excellent starting point and offers a new perspective, some aspects remain underdeveloped and some parts of his analysis are incomplete. Finus and Furini (2023) use a similar model and expand the analysis in different directions, including the consideration of different possible coalition formation scenarios and the analysis of stability of all possible coalition sizes, not only of the grand coalition. They find that large coalitions, including the grand coalition, can be stabilized if signatories use the threat to deploy geoengineering once a country leaves the agreement. They show that this strategy works if collateral damages from SRM are sufficiently high, similar to Millard-Ball (2012), but also sufficiently small in order for the threat to be credible. In other words, collateral damages must lie in some range in order for a climate agreement without SRM deployment to be stable. However, Finus and Furini (2023) assume Nash–Cournot behavior by countries and wrongly argue that Stackelberg leadership would not make a difference in the Millard-Ball's model, as countries have dominant geoengineering strategies.

In this paper, we provide an analysis which considers and compares both assumptions, the Nash–Cournot and the Stackelberg assumption. We focus on various punishment options, their credibility and their effectiveness in stabilizing the grand coalition which cooperates on mitigation and avoids SRM. In order to test the robustness of our conclusions, we consider not only the cartel formation game which Millard-Ball (2012) and Finus and Furini (2023) have used in their analysis, but also consider a repeated game, a second framework, which has been frequently

employed to study stability of international environmental agreements.[3] We show that all our qualitative results hold under both frameworks.

We assume, as Millard-Ball (2012), symmetric countries and focus on the grand coalition for simplicity. We highlight that the different punishment options available under Nash–Cournot and Stackelberg behavior are not necessarily exclusive, can complement each other and are useful for different levels of collateral damages.

In the following, we set out the cartel formation game and derive equilibrium mitigation and SRM strategies in Sect. 2 and analyze stability in Sect. 3. Section 4 considers the alternative framework of a repeated game. Section 5 concludes.

2 The mitigation-solar geoengineering cartel formation game

2.1 The model

There are $n \geq 3$ symmetric countries, $i = 1, 2, ..., n$, with the set of all countries denoted by N, which play the cartel formation game which consists of three stages. In the first stage, countries decide whether to join a climate agreement and become signatories (S) or to stay outside as non-signatories (NS). In the second stage signatories choose their mitigation levels by maximizing the aggregate payoff of coalition members whereas non-signatories choose their mitigation levels by maximizing their individual payoff. In the third stage, all countries individually choose whether to deploy SRM.

We consider two versions for the second stage of the game. In the Nash–Cournot (NC-) scenario, signatories and non-signatories simultaneously set their mitigation levels; in the Stackelberg (ST-) scenario, signatories choose mitigation before non-signatories, taking into account the best-response of non-signatories.

The game is solved by backward induction. Based on the payoff function suggested by Millard-Ball (2012), the following decisions are taken.

In the last stage, all countries face a discrete choice whether to deploy SRM technologies, with no deployment $z_i = 0$ and deployment $z_i = 1$. The net benefit of SRM to country i is given by $(g - Q) \cdot z_i$. The net benefit includes the benefit from geoengineering in the form of reduced climate damages, minus the cost of production (which may be neglected anyway as the cost of SRM are generally viewed to be very low[4]) minus the collateral damage to country i. The marginal net benefit of SRM decreases in total mitigation Q with g being a strictly positive parameter. If $Q \geq g$, SRM does not pay, and if the reverse is true, i.e., $Q < g$, SRM pays. SRM produces collateral damages d to all $n - 1$ other countries, except to the country which uses SRM. It is clear that nothing would change if we acknowledge that collateral damages are uncertain by their nature and talk about expected collateral damages, as long all countries have the same expectations.[5]

[3] See the overview articles by Finus and Caparros (2015), Hovi et al. (2015) and Marrouch and Chaudhuri (2016).

[4] For instance, the models by Ricke et al. (2013) and Weitzman (2015) ignore production cost.

[5] See Finus and Furini (2023) for asymmetric collateral damages.

It is worthwhile to note that in this model, for the decision of a country whether to deploy SRM, expected collateral damages do not matter. This is because all countries are assumed take this decision non-cooperatively.[6] What matters is the level of total mitigation, as this determines whether SRM pays. Moreover, the decision is the same for all symmetric countries. Following Millard-Ball (2012), we also assume that if all countries deploy SRM (because $Q < g$), this is done by a randomly selected country with the probability $1/n$.[7] Hence, in case $z_i = 1 \; \forall \; i \in N$, the expected net benefit and collateral damages of SRM is given by $\frac{1}{n}(g - Q) - \frac{(n-1)}{n}d$ (and if $z_i = 0 \; \forall \; i \in N$, it is 0).

In the second stage, signatories and non-signatories choose their individual mitigation levels $q_i \geq 0$ along a continuous interval. The individual payoff function of country i is given by

$$\pi_i\big(z_i = 0 \;\; \forall \; i \in N\big) = bQ - \frac{c}{2}q_i^2 \tag{1a}$$

$$\pi_i\big(z_i = 1 \;\; \forall \; i \in N\big) = bQ - \frac{c}{2}q_i^2 + \frac{1}{n}(g - Q) - \frac{(n-1)}{n}d \tag{1b}$$

where bQ are linear benefits from total mitigation $Q = \sum_{i=1}^{n} q_i$ and $\frac{c}{2}q_i^2$ are quadratic costs from individual mitigation q_i, with b and c being strictly positive parameters. Whereas non-signatories behave selfishly, maximizing their own payoff (1) with respect to their own mitigation levels, signatories behave cooperatively, maximizing the sum of payoffs of all signatory countries, $\sum_{i \in S} \pi_i$, with respect to all mitigation levels among their group. For a given coalition of size $1 \leq k \leq n$, equilibrium mitigation levels of non-signatories and signatories can be derived, depending on whether SRM technologies will be deployed and whether signatories behave as Stackelberg leaders.[8] Accordingly, the respective payoffs for non-signatories and signatories can be computed. Details are provided in Sect. 2.2.

In the first stage of the game, k out of n countries, $1 \leq k \leq n$, are signatories while $n - k$ countries remain outside the agreement as non-signatories. In the cartel formation game, a coalition of size k is called stable if it is internally and externally stable:

$$\pi_S(k) \geq \pi_{NS}(k-1) \tag{2}$$

$$\pi_{NS}(k) \geq \pi_S(k+1). \tag{3}$$

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.

[6] See Finus and Furini (2023) who discuss the possibility that signatories cooperate on the deployment of SRM.

[7] In Finus and Furini (2023) it is shown that this assumption is not crucial and can be replaced by all countries deploying SRM if $Q < g$.

[8] For a trivial coalition with $k = 1$, there is no difference between signatories and non-signatories (all players behave non-cooperatively) and if $k = n$, there are no non-signatories (all players behave cooperatively).

2.2 Mitigation and solar radiation management equilibria

In the last stage, country i will abstain from SRM $\left(z_i = 0\right)$ if $g - Q \leq 0$ and will use SRM $\left(z_i = 1\right)$ if $g - Q > 0$. Due to symmetry, either all countries will find it attractive to use SRM or none does.

In the absence of SRM, i.e., $z_i^* = 0$, the individual payoff function is (1a), with equilibrium mitigation levels in the pure "Mitigation-equilibrium" (M-equilibrium) being given by $q_S^{M*}(k) = \frac{kb}{c}$, $q_{NS}^{M*} = \frac{b}{c}$ and $Q^{M*}(k) = \left(k^2 + n - k\right)\left(\frac{b}{c}\right)$ with $\partial Q^{M*}(k)/\partial k > 0$. Given linear benefits from mitigation, countries have dominant strategies in the M-equilibrium. Hence, there is no difference between the Stackelberg assumption and Nash–Cournot assumption.

In the "Geoengineering-equilibrium" (G-equilibrium), SRM technologies are deployed, i.e., $z_i^* = 1$.[9] Consequently, based on payoff function (1b), equilibrium individual and total mitigation levels are given by $q_S^{G^*}(k) = k\left(\frac{b}{c} - \frac{1}{nc}\right)$, $q_{NS}^{G^*} = \frac{b}{c} - \frac{1}{nc}$ and $Q^{G^*}(k) = \left(k^2 + n - k\right)\left(\frac{b}{c} - \frac{1}{nc}\right)$. Given linear benefits from mitigation, countries have dominant strategies in the G-equilibrium. Hence, the Stackelberg assumption results in the same equilibrium mitigation levels as the Nash–Cournot assumption. In order to guarantee non-negative mitigation levels, we impose condition $C_1 := b \geq \frac{1}{n}$. Inserting equilibrium mitigation levels in payoff function (1b), we derive payoffs of signatories and non-signatories $\pi_S^{G^*}(k)$ and $\pi_{NS}^{G^*}(k)$, respectively.

Millard-Ball (2012) assumes that the global mitigation level in the M-equilibrium always falls short of the level g such that SRM always pays, i.e., even in the grand coalition, i.e., $Q^{M^*}(n) < g$. This gives condition $C_2 := g > \frac{bn^2}{c}$. This leads to the consideration of an "Avoidance-equilibrium" (A-equilibrium), in which countries increase their mitigation efforts in order to avoid the deployment of SRM.

In the A-equilibrium, $Q^{A^*}(k) \geq g$ must hold by assumption, such that $z_i^* = 0$. In this equilibrium, it now makes a difference whether the NC- or the ST-scenario is assumed. This has been overlooked by Finus and Furini (2023).

In the NC-scenario, we assume that non-signatories do not contribute to the extra effort to make SRM unattractive. That is, non-signatories stick to their dominant strategy, $q_{NS}^{A^{NC*}} = q_{NS}^{G^*}$. Hence, all effort must be exerted by signatories. Given that signatories, according to their first order conditions in an interior equilibrium, would normally set the sum of marginal benefits equal to their individual marginal costs of mitigation, signatories have no incentive to provide more mitigation than is required to render SRM an unattractive choice. Hence, $Q^{A^{NC*}}(k) = g$. Consequently, $q_S^{A^{NC*}}(k) = \frac{g-(n-k)q_{NS}^{A^{NC*}}}{k} = \frac{cgn-n^2b+bkn+n-k}{ckn}$ and $q_{NS}^{A^{NC*}} = q_{NS}^{G^*} = \frac{b}{c} - \frac{1}{nc}$. Inserting equilibrium mitigation levels in payoff function (1a), gives payoffs $\pi_S^{A^{NC*}}(k)$ and $\pi_{NS}^{A^{NC*}}(k)$ in the A-equilibrium in the NC-scenario.

[9] We use the term "Geoengineering-Equilibrium" and not "Solar Radiation Management- Equilibrium" in order to save on notation.

In the ST-scenario, signatories can use their leadership role in order to shift part of the mitigation burden to non-signatories. They will choose the lowest possible contribution level $q_S^{A^{ST*}}$ such that non-signatories will still find it attractive to provide the additional mitigation required to reach the total mitigation level $Q^{A^*} = g$ in order to avoid the deployment of geoengineering.

Consider the coalition of size $n-1$, which is relevant when testing for the stability of the grand coalition. The individual non-signatory, the free-rider, who has left the grand coalition, needs to provide $q_{NS}^{A^{ST*}} = g - (n-1)q_S^{A^{ST*}}$, such that the level $Q^{A^{ST*}}(n-1) = g$ in the A-equilibrium is achieved and the SRM technology is not deployed. This Stackelberg strategy of signatories only works if the non-signatory prefers contributing $q_{NS}^{A^{ST*}}$ rather than only $q_{NS}^{G^*}$ as a response to $q_S^{A^{ST*}}$ where in the former case the SRM-technology is not deployed and in the latter case it is deployed in the last stage. Hence, in order for the A-equilibrium to arise in the Stackelberg scenario, the following inequality must hold:

$$\pi_{NS}^{A^{ST*}}(n-1) = bg - \frac{c}{2}\left(g-(n-1)q_S^{A^{ST*}}\right)^2 \geq b\left((n-1)q_S^{A^{ST*}} + q_{NS}^{G^*}\right) \tag{4}$$
$$- \frac{c}{2}q_{NS}^{G^{*2}} + \frac{1}{n}\left(g - \left((n-1)q_S^{A^{ST*}} + q_{NS}^{G^*}\right)\right) - \frac{(n-1)}{n}d = \tilde{\pi}_{NS}^{G*}(n-1).$$

Note that the G-equilibrium on the right-hand side is more attractive to non-signatories than the "original" G-equilibrium as $q_S^{A^{ST*}} > q_S^{G^*}(n-1)$ holds, i.e., $\tilde{\pi}_{NS}^{G*}(n-1) > \pi_{NS}^{G*}(n-1)$. We can show that the inequality above is more likely to hold the larger $q_S^{A^{ST*}}$. Given that signatories want to contribute as little as possible to maximize their payoff, we find the smallest contribution level $\underline{q}_S^{A^{ST*}} = \frac{gnc-bn+1-\sqrt{2dcn(n-1)}}{(n-1)cn}$ such that Eq. (4) holds with strict equality. Note that $\underline{q}_S^{A^{ST*}}$ decreases in d. Hence, if signatories want to establish the Stackelberg A-equilibrium at $n-1$, they can choose a lower mitigation level and consequently impose a greater burden on non-signatories the larger collateral damages from SRM are. The larger collateral damages are, the more attractive it is for non-signatories to stick to the A-equilibrium.

The details of all calculations in this section are provided in Appendix A.1.

3 Stability of the grand coalition in the cartel formation game

3.1 Nash–Cournot scenario

In the Nash–Cournot scenario, for any generic coalition k, the decision of whether to play the A- or the G-equilibrium is taken by signatories. Non-signatories choose the same mitigation level in both equilibria, but signatories go the extra mile in the A-equilibrium and provide additional mitigation in order to reach the total mitigation level $Q^{A^*} = g$ such that the deployment of SRM technologies

does not pay. Hence, the A-equilibrium is preferred over the G-equilibrium if $\pi_S^{A^{NC^*}}(k) - \pi_S^{G^*}(k) \geq 0$, and if $\pi_S^{A^{NC^*}}(k) - \pi_S^{G^*}(k) < 0$, the reverse is true, signatories play their dominant strategy and the G-equilibrium is implemented.

Lemma 1 (Choice of equilibrium strategies in the NC-scenario)
For any coalition size k, $1 \leq k \leq n$*, the A-equilibrium is played if* $d \geq \overline{d}(k)$*, while the G-equilibrium is played if* $d < \overline{d}(k)$*.* $\overline{d}(k)$ *decreases in the coalition size k, i.e.,* $\frac{\partial \overline{d}(k)}{\partial k} < 0$*.*

Proof See Appendix A.2.

Clearly, if $d < \overline{d}(k)$, it would be perfectly rational (individually and globally) to use SRM in a coalition of size k in the first place, as expected collateral damages falls short of expected benefits from SRM. Thus, an agreement which does not use SRM (A-equilibrium) is only individually and globally rational if $d \geq \overline{d}(k)$. The A-equilibrium becomes more attractive with the number of signatories, as each of them needs to contribute less to achieve the benchmark mitigation level $Q^{A*} = g$ in order to make it unattractive to use SRM.

In the grand coalition, $\overline{d}(n) \leq d$ is needed such that it is rational to play the A- and not the G-equilibrium. If one country leaves the A-agreement, i.e., the coalition is of size $n-1$, either the A- or the G-equilibrium can be played. If the A-equilibrium is played, then it can easily be shown that the grand coalition is not stable. Hence, in the NC-scenario, only the G-equilibrium can serve as a deterrent punishment to stabilize the grand coalition. For this to be credible, we need $d < \overline{d}(n-1)$ such that the G- and not the A-equilibrium is played by all countries if one country leaves the agreement. Together, we require $\overline{d}(n) \leq d < \overline{d}(n-1)$.

The grand coalition is internally stable, and, hence, stable (as the grand coalition is externally stable by definition), if $\pi_S^{A^*}(n) \geq \pi_{NS}^{G^*}(n-1)$ holds, which gives the condition $\hat{d}(n) \leq d$.

Altogether, we can state the following.

Proposition 1 (Stability of the grand coalition in the Nash–Cournot scenario)
In the cartel formation game and Nash–Cournot scenario, it can be individually and globally rational to implement the A-equilibrium in the grand coalition. If this is the case, then the A-equilibrium is stable if and only if after a deviation the G-equilibrium is played for which for the collateral damage parameter d the following must hold:

$$\overline{d}(n) < \hat{d}(n) \leq d < \overline{d}(n-1).$$

That is, collateral damages must lie between a lower $\hat{d}(n)$ *and an upper bound* $\overline{d}(n-1)$*.*

Proof See Appendix A.3.

Hence, in the NC-scenario, only the G-equilibrium can be used as punishment. The punishment is effective (stability holds) if the collateral damage parameter d

is sufficiently large. The punishment is credible if and only if signatories prefer the G-equilibrium over the A-equilibrium once a country has left their agreement for which the collateral damage parameter cannot be too large. Conceptually, it is important to establish under which conditions the A-equilibrium is attractive in the first place, i.e., $\overline{d}(n) \leq d$. This has been ignored by Millard-Ball (2012), even though it turns out that this condition is not binding as the internal stability condition $\hat{d}(n) < d$ is more demanding and we have $\overline{d}(n) < \hat{d}(n)$.

3.2 Stackelberg scenario

When analyzing the stability of the grand coalition playing the A-equilibrium, the ST-scenario differs from the NC-scenario only in terms of possible punishment after a country has left the agreement. Stackelberg leadership increases the options with which signatories can punish the deviating country.

In the grand coalition, all countries are signatories and hence Stackelberg leadership does not play a role. The A-equilibrium is implemented in the first place if $\overline{d}(n) \leq d$, as observed above already (see Lemma 1). However, at coalition of size $n-1$, Stackelberg leadership may play a role. Obviously, Stackelberg leaders can always replicate their Nash–Cournot strategy. Consequently, one punishment option is to play the G-equilibrium and we recall from Sect. 2.1 that in the G-equilibrium mitigation strategies are not affected by Stackelberg leadership as countries have dominant strategies. Hence, the G-equilibrium is an effective punishment that stabilize the grand coalition for collateral damage levels for which $\overline{d}(n) < \hat{d}(n) \leq d < \overline{d}(n-1)$ is true, as found for the NC-scenario. This possibility has been overlooked by Millard-Ball (2012). He focuses exclusively on the second punishment option with an A-equilibrium, which is not an option in the NC-scenario.

The A-equilibrium can constitute a punishment if signatories contribute less than what they do in the grand coalition, i.e., if $q_S^{A^{ST^*}} \leq \frac{g}{n}$. If this is the case, it follows that the individual non-signatory (i.e., the free-rider) will have to contribute $q_{NS}^{ST^*} \geq \frac{g}{n}$ and the free-rider will be worse off than in the grand coalition. Hence, the A-equilibrium constitutes an effective punishment that stabilizes the grand coalition. From Sect. 2.2 we know that the free-rider accepts the A-equilibrium punishment only if $q_S^{A^{ST^*}}$ is equal or above $\underline{q}_S^{A^{ST^*}}$ where $\underline{q}_S^{A^{ST^*}}$ decreases in d. We also know that the free-rider is more likely to accept the A-equilibrium punishment the larger $q_S^{A^{ST^*}}$. Hence, we can compute a lower bound for the collateral damage parameter, $\overline{d}_{\min}^{ST}(n-1)$, such that if $d \geq \overline{d}_{\min}^{ST}(n-1)$, the A-equilibrium constitutes an effective punishment that stabilizes the grand coalition. Because $q_{NS}^{A^{ST^*}}\left(q_S^{A^{ST^*}}(n-1)\right) \geq \frac{g}{n}$ and $q_S^{A^{ST^*}}(n-1) \leq \frac{g}{n}$, we have internal stability: $\pi_S^{A^*}(n) \geq \pi_{NS}^{A^{ST^*}}(n-1)$; at the same time the punishment is credible because signatories are better off in the Stackelberg A-equilibrium at $n-1$ than in the A-agreement in the grand coalition, i.e., $\pi_S^{A^*}(n) \leq \pi_S^{A^{ST^*}}(n-1)$.

(Moreover, one can show that $\pi_S^{A^{ST^*}}(n-1) > \pi_S^{G^*}(n-1)$, signatories prefer the A-equilibrium to the G-equilibrium at $n-1$.)

Proposition 2 (Stability of the grand coalition in the Stackelberg scenario) *In the cartel formation game and Stackelberg scenario, it can be individually and globally rational to implement the A-equilibrium in the grand coalition. If this is the case, then the A-equilibrium can be stable if after a deviation either the A- or the G-equilibrium is played:*

(i) The A-equilibrium constitutes an effective and credible punishment provided that for the collateral damage parameter d the following holds:

$$\overline{d}(n) < \hat{d}(n) < \overline{d}^{ST}_{\min}(n-1) \leq d.$$

(ii) If instead $d < \overline{d}^{ST}_{\min}(n-1)$*, the G-equilibrium can be used as an effective and credible punishment, provided that for the collateral damage parameter d the following holds:*

$$\overline{d}(n) < \hat{d}(n) \leq d < \overline{d}(n-1).$$

Proof See Appendix A.4.

Thus, with Stackelberg leadership signatories have a stronger position to enforce a stable climate agreement than under the NC-scenario which shows up in a larger parameter space for which the grand coalition can establish a stable A-equilibrium. For larger expected collateral damages, the A-equilibrium punishment works and for lower expected collateral damages the G-equilibrium punishment works. In both cases, punishment is deterrent and credible.

4 Stability of the grand coalition in the repeated game

In a repeated game, a free-rider gains a temporary free-rider gain before the deviation is discovered and punishment follows. Following many others, we consider the simple trigger punishment strategy to drive home our result.[10] Once signatories discover free-riding, they will stop cooperating and play a punishment in all subsequent periods. In an infinitely repeated game, this trigger strategy is subgame-perfect if the punishment constitutes a Nash-equilibrium (or Stackelberg-equilibrium) of the static game. Consistent with our previous assumption, we assume that the grand coalition implements the A-equilibrium as all countries prefer the A-equilibrium over the G-equilibrium, $\pi_S^{A^*}(n) \geq \pi_S^{G^*}(n)$ because $\overline{d}(n) \leq d$, where equilibrium mitigation levels and payoffs are those derived in the cartel formation game (Sect. 2.2).

[10] For an overview on repeated games and different punishment strategies see, e.g., Finus (2001, and 2003).

Furthermore, we denote free-rider payoff by π_{NS}^{F} and denote the punishment payoff by $\pi_{NS}^{P^*}$. The A-agreement is stable, provided

$$\frac{\pi_S^{A^*}(n)}{1-\delta} \geq \pi_{NS}^{F} + \frac{\delta \pi_{NS}^{P^*}}{1-\delta} \quad \Rightarrow \quad \delta \geq \delta_{\min} := \frac{\pi_{NS}^{F} - \pi_S^{A^*}(n)}{\pi_{NS}^{F} - \pi_{NS}^{P^*}} \tag{5}$$

holds. Countries comply if the net present value of complying is larger than the net present value of taking a free-ride and subsequently being punished. That is, if the discount factor by which countries discount time δ exceeds a value $\delta_{\min}$, the threat of punishment is deterrent and the agreement is stable. Since $\delta = \frac{1}{1+r}$ with r the discount or time preference rate, $0 \leq \delta \leq 1$. Obviously, the harsher the punishment, the lower $\pi_{NS}^{P^*}$ and the lower the threshold $\delta_{\min}$.

In a first instance, we need to determine the free-rider payoff. Free-riding implies that the defector chooses the dominant mitigation level $q_{NS}^{G^*}$ in the G-equilibrium and individually deploys SRM, receives the benefits but and does not suffer from collateral damages. All other $n-1$ compliant countries continue choosing the avoidance mitigation level $q_S^{A^*}(n) = \frac{g}{n}$. Hence, the free-rider payoff is given by:

$$\pi_{NS}^{F} = b\big(g - q_S^{A^*}(n) + q_{NS}^{G^*}\big) - \frac{c}{2}\big(q_{NS}^{G^*}\big)^2 + \big(g - \big(g - q_S^{A^*}(n) + q_{NS}^{G^*}\big)\big). \tag{6}$$

This exhaust the common assumption of the NC- and ST-scenario, as different punishment options arise in these two scenarios.

4.1 Nash–Cournot scenario

The only option for signatories to punish the free-rider in the NC-scenario is to stop cooperating and play the non-cooperative G-equilibrium with $q_{NS}^{G^*}(1) = \frac{b}{c} - \frac{1}{nc}$ and $z_i^* = 1$ as derived in subSect. 2.2. All countries receive the payoff $\pi_{NS}^{G^*}(1)$. Essentially, punishment implies that the grand coalition breaks apart and the G-equilibrium is played among single players. For the G-equilibrium to be preferred over the A-equilibrium such that the G-punishment is a credible punishment, we need $d < \overline{d}(1)$ from Lemma 1. Together with the condition that the A- is preferred over the G-equilibrium in the grand coalition, $\overline{d}(n) \leq d$, we require $\overline{d}(n) \leq d < \overline{d}(1)$. Again, as in the cartel formation game, the A-equilibrium cannot be used as punishment. All countries comply if

$$\frac{\pi_S^{A^*}(n)}{1-\delta} \geq \pi_{NS}^{F} + \frac{\delta \pi_{NS}^{G^*}(1)}{1-\delta} \quad \Rightarrow \quad \delta \geq \delta_{\min}^{G} := \frac{\pi_{NS}^{F} - \pi_S^{A^*}(n)}{\pi_{NS}^{F} - \pi_{NS}^{G^*}(1)} \tag{7}$$

Given that $\pi_{NS}^{G^*}(1)$ decreases in the collateral damage d and the other payoffs in (7) are not affected by collateral damages, we can conclude that the minimum required discount factor $\delta_{\min}^{G}$ decreases with d.

Proposition 3 (Stability of the grand coalition in the repeated game and NC-scenario) *In the repeated game and Nash–Cournot scenario, the grand coalition implementing the A-equilibrium and punishing with a G-equilibrium can be stable (subgame perfect Nash equilibrium) if for the collateral damage parameter of SRM d the following holds:*

$$\overline{d}(n) < d < \overline{d}(1).$$

That is, expected collateral damages must lie between a lower $\overline{d}(n)$ *and an upper bound* $\overline{d}(1)$.

The agreement is stable if for the discount factor $\delta \geq \delta^{G}_{\min}$ *holds, with* $0 < \delta^{G}_{\min} < 1$, $\delta^{G}_{\min}$ *decreasing in the collateral damage level d and* $\delta^{G}_{\min}$ *as defined in Eq.* (7).

Proof See Appendix A.5.

Thus, the qualitative conclusion in Proposition 3 for the repeated game is the same as derived for the cartel formation game in Proposition 1. The collateral damage must be sufficiently large such that punishment is deterrent, but cannot be too large, as otherwise the threat to punish with the deployment of SRM technologies is not credible. Within the feasibility range, $\overline{d}(n) < d < \overline{d}(1)$, the higher the collateral damage, the lower the minimum discount factor required for stability. Again, it should be noted that if SRM were not to pay even if there is no agreement at all, i.e., $\overline{d}(1) \leq d$, the grand coalition implementing the A-equilibrium would not be stable, as no deterrent punishment would be available.

4.2 Stackelberg scenario

Compared to the NC-scenario, and similarly to what we found for the cartel formation game, in the ST-scenario signatories have an additional punishment option. Given the G-equilibrium is not affected by Stackelberg leadership, this can also be used as a punishment under the same condition as in the NC-scenario, i.e., if $\overline{d}(n) < d < \overline{d}(1)$. Additionally, the A-equilibrium can also be used as punishment if the $n-1$ remaining signatories act as Stackelberg leaders and implement a A-equilibrium in which they contribute $q_S^{A^{ST}} \leq \frac{g}{n}$. As shown previously in Proposition 2, this is possible if $\overline{d}(n) < \overline{d}^{ST}_{\min}(n-1) \leq d$. This punishment will be credible as $\pi_S^{A^{ST^*}}(n-1) \geq \pi_S^{A^*}(n) \geq \pi_S^{G^*}(n) > \pi_{NS}^{G^*}(1)$ holds for $\overline{d}(n) < \overline{d}^{ST}_{\min}(n-1) \leq d$. Given that the A-equilibrium is played in the grand coalition and the A-equilibrium is an effective punishment, for which $\overline{d}(n) < \overline{d}^{ST}_{\min}(n-1) \leq d$ is required such that the free-rider accepts this punishment, all countries will comply in an infinitely repeated game, provided

$$\delta \geq \delta^{A}_{\min} := \frac{\pi^{F}_{NS} - \pi^{A^*}_{S}(n)}{\pi^{F}_{NS} - \pi^{A^{ST^*}}_{NS}(n-1)} \tag{8}$$

holds. As we know from Sect. 2.2, the punishment payoff $\pi^{A^{ST^*}}_{NS}(n-1)$ depends on signatories' mitigation level $q^{A^{ST^*}}_{S}$. The harshest possible punishment coincides with the lowest possible mitigation level $\underline{q}^{A^{ST^*}}_{S}$ which decreases in the collateral damage parameter d. Given that the other payoffs in (8) are not affected by parameter d, we can conclude that the minimum discount factor $\delta^{A}_{\min}$ decreases in d. Additionally, we note that, given $\overline{d}(n) < \overline{d}^{ST}_{\min}(n-1) \leq d$ and $\pi^{A^{ST^*}}_{NS}(n-1) > \pi^{G^*}_{NS}(n-1) > \pi^{G^*}_{NS}(1)$ hold, it follows that the minimum discount factor requirement is larger with the A- than the G-punishment, i.e., $\delta^{A}_{\min} > \delta^{G}_{\min}$. That is, the A-equilibrium is a weaker punishment than the G-equilibrium. Hence, when both punishment options establish cooperation, signatories will choose the G-equilibrium punishment.

Proposition 4 (Stability of the grand coalition in the repeated game and ST-scenario) *In the repeated game and Stackelberg scenario, the grand coalition implementing the A-equilibrium can be stable (subgame perfect Nash equilibrium) under two punishment options*:

(i) The G-equilibrium is used as punishment if for the collateral damage parameter d $\overline{d}(n) < d < \overline{d}(1)$ *holds.*

That is, expected collateral damages must lie between a lower and an upper bound. The agreement is stable if for the discount factor $\delta \geq \delta^{G}_{\min}$ *holds, with* $0 < \delta^{G}_{\min} < 1$, $\delta^{G}_{\min}$ *decreasing in the collateral damage parameter d and* $\delta^{G}_{\min}$ *as defined in Eq.* (7).

(ii) The A-equilibrium is used as punishment if for the collateral damage parameter d $\overline{d}(n) < \overline{d}^{ST}_{\min} < \overline{d}(1) \leq d$ *holds.*

That is, expected collateral damages must be sufficiently large.

The agreement is stable if for the discount factor $\delta \geq \delta^{A}_{\min}$ *holds, with* $0 < \delta^{A}_{\min} \leq 1$, $\delta^{A}_{\min}$ *decreasing in the collateral damage level d*, $\delta^{A}_{\min}$ *as defined in Eq.* (8) *and* $\delta^{A}_{\min} > \delta^{G}_{\min}$.

Proof See Appendix A.6.

In line with what we find in the cartel formation game, Stackelberg leadership increases the punishment options available to signatories in the repeated game too. When the G-equilibrium is a credible punishment, it will be used by signatories as it is the harsher punishment than the A-equilibrium punishment. If the G-equilibrium

does not work as a punishment, then the A-equilibrium can be an effective punishment, even though it is a weaker punishment, resulting in a larger minimum discount factor requirement.

4.3 Renegotiation-proof punishment

An interesting twist of our previous results emerges if we consider renegotiation-proofness which is a refinement of subgame-perfection. The version which we consider has been proposed by Farrell and Maskin (1989) and has been applied for instance by Asheim et al. (2006), Asheim and Holtsmark (2009), Barrett (1994, 2002), Finus and Rundshagen (1998a, b) in the context of international environmental agreements. A punishment is only credible if the players conducting the punishment are not worse off. Consequently, as $\pi_S^{A^*}(n) \geq \pi_S^{G^*}(n) > \pi_{NS}^{G^*}(1)$, where the first inequality follows from $\overline{d}(n) \leq d$ and the second inequality follows from the fact that full cooperation delivers higher payoffs than the Nash-equilibrium, the harsher punishment with the G-equilibrium is not renegotiation-proof. However, the weaker punishment with the A-equilibrium is renegotiation-proof as $\pi_S^{A^*}(n) \leq \pi_{NS}^{A^{ST^*}}(n-1)$ if $\overline{d}(n) < \overline{d}_{\min}^{ST}(n-1) \leq d$. This punishment option is only available under Stackelberg leadership. Hence, without Stackelberg leadership, the grand coalition implementing the A-equilibrium cannot be supported as a renegotiation-proof equilibrium and with Stackelberg leadership, only the A-equilibrium punishment is renegotiation proof. In Appendix A.6, we provide further details.

5 Conclusion

We analyzed the governance structure of a self-enforcing climate agreement aiming at reducing greenhouse gases in the light of geoengineering in the form of solar radiation management (SRM) with the possibility of high collateral damages. In particular, we focused on the design and credibility of threats that can be used to deter free-riding under two scenarios, the Nash–Cournot and Stackelberg scenario. We investigated stability in two settings, the cartel formation game and the repeated game. Hence, we qualify, extend and compare the results of Millard-Ball (2012) and Finus and Furini (2023).

We showed that a climate agreement on mitigation can be designed that avoids the deployment of SRM by increasing mitigation levels such that SRM is not attractive. Under the Nash–Cournot scenario, the only effective threat available to signatories is the deployment of SRM. This threat works if expected collateral damages are high enough such that SRM is avoided in the grand coalition and leaving the coalition is not profitable, but not too high so that the threat of SRM deployment once a country leaves the agreement is credible. Thus, the threat to deploy SRM can be a powerful tool to stabilize a climate agreement; it can lead to large stable climate agreements with a high global mitigation effort, but it is not as simple as Millard-Ball (2012) finds for the Stackelberg scenario. If collateral damages are expected to be very high,

then this stabilisation strategy does not work. Thus, if SRM was never considered a rational strategy, climate agreements which exclusively focus on mitigation will not achieve much, a feature confirmed by past climate agreements. However, this is different if signatories can choose mitigation levels before non-signatories. In the Stackelberg scenario, an additional threat is available. Signatories can now decrease their mitigation efforts and leave to non-signatories the mitigation burden in order to avoid the use of SRM. This threat is credible and effective if collateral damages are high enough, and no upper bound is required, in line with Millard-Ball (2012). However, even if collateral damages are below this threshold, signatories still have the possibility to stabilize the grand coalition using the deployment of SRM as a threat, as found in the NC-scenario. Hence, Stackelberg leadership increases the threat options available for signatories and expands the collateral damage parameter range under which stability can be achieved. Now, successful agreements providing high mitigation levels that render the use of SRM unattractive can be stable even when SRM is never considered a rational strategy due to very high expected collateral damages. Of course, the reverse is also true for our model: if expected collateral damages are expected to be very low, none of the threats works.

For future research, many issues come to mind. However, we believe the most important extensions would deal with more general payoff functions than those considered by Millard-Ball (2012) and Finus and Furini (2023), the analysis whether a moratorium on research on SRM technologies would be enforceable and effective and how adaptation in addition to mitigation affects the governance structure in climate change in the light of SRM.

Appendix A

Detailed proofs are available from the authors upon request.

A.1 Mitigation Levels in the G-, M- and A-equilibrium

In the M-equilibrium, $z_i^* = 0$ in the last stage. In the second stage of the game, non-signatories choose mitigation to maximize their individual payoff. Non-signatories' first order conditions in an interior equilibrium are given by $b - cq_{NS}^M = 0$ from which the optimal mitigation levels $q_{NS}^{M^*} = \frac{b}{c}$ follow. Signatories choose mitigation to maximize the aggregate payoff to the coalition. The resulting first order conditions in an interior equilibrium are given by $kb - cq_{NS}^M = 0$ from which the optimal mitigation levels $q_S^{M^*}(k) = \frac{kb}{c}$ follow. Total mitigation is given by $Q^{M^*}(k) = k \cdot q_S^{M^*}(k) + (n-k) \cdot q_{NS}^{M^*} = \left(k^2 + n - k\right)\left(\frac{b}{c}\right)$. Given dominant strategies, the mitigation (and payoff) levels in the M-equilibrium do not differ between the NC- and the ST-scenario.

In the G-equilibrium, $z_i^* = 1$ in the last stage. In the second stage, non-signatories choose mitigation to maximize their individual payoff. Non-signatories'

first order conditions in an interior equilibrium are given by $b - cq_{NS}^{G} - \frac{1}{n} = 0$ from which the equilibrium mitigation levels $q_{NS}^{G^*} = \frac{b}{c} - \frac{1}{nc}$ follow. Signatories choose mitigation levels to maximize the aggregate payoff to the coalition. The resulting first order conditions in an interior equilibrium are given by $kb - cq_{S}^{G} - k\frac{1}{n} = 0$ from which the optimal mitigation levels $q_{S}^{G^*}(k) = k\left(\frac{b}{c} - \frac{1}{nc}\right)$ follow. Hence, the aggregate mitigation level is given by $Q^{G^*}(k) = k \cdot q_{S}^{G^*} + (n-k) \cdot q_{NS}^{G^*} = \left(k^2 + n - k\right)\left(\frac{b}{c} - \frac{1}{nc}\right)$. Given dominant strategies, the mitigation (and payoff) levels in the G-equilibrium do not differ between the NC- and the ST-scenario.

In the A-equilibrium, $z_i^* = 0$ in the last stage. Second stage equilibrium levels in the NC-scenario have been derived in the text. In the ST-scenario, we focus on the coalition of size $k = n-1$, where the individual non-signatory needs to provide $q_{NS}^{A^{ST^*}} = g - (n-1)q_{S}^{A^{ST^*}}$, such that the level $Q^{A^{ST^*}}(n-1) = g$ is achieved. The signatories' mitigation level $q_{S}^{A^{ST^*}}$ needs to satisfy Eq. (4), which can be rewritten as

$$\pi_{NS}^{A^{ST^*}}(n-1) - \tilde{\pi}_{NS}^{G^*}(n-1) = \left(bg - \frac{c}{2}\left(g - (n-1)q_{S}^{A^{ST*}}\right)^2\right) - \left(b\left((n-1)q_{S}^{A^{ST^*}} + q_{NS}^{G^*}\right) - \frac{c}{2}q_{NS}^{G^{*2}} + \frac{1}{n}\left(g - \left((n-1)q_{S}^{A^{ST^*}} + q_{NS}^{G^*}\right)\right) - \frac{(n-1)}{n}d\right) \geq 0.$$

Differentiating $\pi_{NS}^{A^{ST^*}}(n-1) - \tilde{\pi}_{NS}^{G^*}(n-1)$ with respect to $q_{S}^{A^{ST^*}}$ gives $-\frac{(n-1)\left(q_{S}^{A^{ST^*}} cn^2 - gcn - q_{S}^{A^{ST^*}} cn + bn - 1\right)}{n}$. The sign depends on the sign of $-\left(q_{S}^{A^{ST^*}} cn^2 - gcn - q_{S}^{A^{ST^*}} cn + bn - 1\right)$. This term decreases in $q_{S}^{A^{ST^*}}$. Since we look for signatories' mitigation levels $q_{S}^{A^{ST^*}} \leq \frac{g}{n}$, replacing the largest possible $q_{S}^{A^{ST^*}} = \frac{g}{n}$, the term reads $-(bn - cg - 1)$, which is positive given condition $C_2 := g > \frac{bn^2}{c}$. Hence, the sign is always positive and the difference $\pi_{NS}^{A^{ST^*}}(n-1) - \tilde{\pi}_{NS}^{G^*}(n-1)$ increases in $q_{S}^{A^{ST^*}}$. Given that signatories want to contribute as little as possible, we solve Eq. (4) imposing strict equality. This gives two solutions: $\frac{gnc - bn + 1 - \sqrt{2dcn(n-1)}}{(n-1)cn}$ and $\frac{gnc - bn + 1 + \sqrt{2dcn(n-1)}}{(n-1)cn}$. It can be shown that the second solution is always larger than $\frac{g}{n}$ and hence it can be ruled out. The smallest contribution level of signatories is $\underline{q}_{S}^{A^{ST^*}} = \frac{gnc - bn + 1 - \sqrt{2dcn(n-1)}}{(n-1)cn}$. It follows $\overline{q}_{NS}^{A^{ST^*}} = g - (n-1)\underline{q}_{S}^{A^{ST^*}} = \frac{bn + \sqrt{2dcn(n-1)} - 1}{cn}$ and $Q^{A^{ST^*}} = g$.

A.2 Proof of Lemma 1

Consider the NC-scenario. If signatories find it attractive to implement the A-equilibrium, non-signatories will be better off by moving from the G- to the

A-equilibrium as well. This is true because $\pi_{NS}^{A^{NC^*}}(k) - \pi_{NS}^{G^*}(k) > \pi_S^{A^{NC^*}}(k) - \pi_S^{G^*}(k)$ holds. Moving from the G- to the A-equilibrium, both signatories and non-signatories will experience the same payoff effects with respect to the net benefits of SRM and collateral damages, both will have the same increase of the benefits of mitigation, but signatories will face an increase in mitigation costs, which does not occur to non-signatories, as their mitigation remains the same.

Signatories decide to implement the A-equilibrium if their payoff would be weakly larger than in the G-equilibrium, $\pi_S^{A^{NC^*}}(k) - \pi_S^{G^*}(k) \geq 0$. Otherwise, the G-equilibrium is played. Differentiating $\pi_S^{A^{NC^*}}(k) - \pi_S^{G^*}(k)$ 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 $\overline{d}(k) = \frac{\left(bk^2n - bkn + bn^2 - gnc - k^2 + k - n\right)^2}{2cnk^2(n-1)}$ above which $\left(d \geq \overline{d}(k)\right)$ the A-equilibrium and below which $\left(d < \overline{d}(k)\right)$ the G-equilibrium is played. Moreover, we find:

$$\frac{\partial \overline{d}(k)}{\partial k} = \frac{\left(nbk^2 - bkn + bn^2 - gnc - k^2 + k - n\right)\left(nbk^2 - bn^2 + gnc - k^2 + n\right)}{cnk^3(n-1)} < 0.$$

The denominator is clearly positive, while in the nominator the term in the first brackets is negative and the term in the second brackets is positive. This can be shown by using condition $C_1 := b \geq \frac{1}{n}$ and condition $C_2 := g > \frac{bn^2}{c}$, as derived in the paper.

A.3 Proof of Proposition 1

We consider internal stability of the grand coalition in the NC-scenario assuming that the A-equilibrium is played, while the G-equilibrium is played if a member leaves the coalition. That is, $\overline{d}(n) \leq d < \overline{d}(n-1)$ must hold. Substituting respectively $k = n$ and $k = n - 1$ in $\overline{d}(k)$ as derived above, we obtain:

$$\overline{d}(n) = \frac{\left(bn^2 - cg - n\right)^2}{2cn(n-1)} \text{ and } \overline{d}(n-1) = \frac{\left(bn^3 - 2bn^2 - gnc + 2bn - n^2 + 2n - 2\right)^2}{2cn(n-1)^3}.$$

Given $\overline{d}(n) \leq d < \overline{d}(n-1)$, we apply the internal stability condition which reads $\pi_S^{A^*}(n) - \pi_{NS}^{G^*}(n-1) \geq 0$. It can be shown that $\pi_S^{A^*}(n) - \pi_{NS}^{G^*}(n-1)$ increases in d. Thus, internal stability holds if.

$$d \geq \hat{d}(n) = \frac{2b^2n^4 - 4b^2n^3 - 2bcgn^2 + 3b^2n^2 - 4bn^3 + c^2g^2 + 8bn^2 + 2gnc - 6bn + 2n^2 - 4n + 3}{2cn(n-1)}.$$

Comparing $\overline{d}(n)$ with $\hat{d}(n)$ gives

$$\hat{d}(n) - \overline{d}(n) = \frac{(n-3)(bn-1)^2}{2nc} \geq 0$$

for $n \geq 3$ which we assume to hold. Hence, the condition for the stability of the grand coalition implementing the A-equilibrium is $\hat{d}(n) \leq d < \overline{d}(n-1)$.

A.4 Proof of Proposition 2

We consider internal stability of the grand coalition in the ST-scenario assuming that the A-equilibrium is played. The G-equilibrium can be used as effective punishment under the same conditions identified for the NC-scenario and derived in Appendix A.3. In the ST-scenario, an additional punishment option arises, with the A-equilibrium being played if a member leaves the coalition. As explained in the text, the punishment is effective and credible if $\underline{q}_S^{A^{ST^*}} = \frac{gnc-bn+1-\sqrt{2dc\,n(n-1)}}{(n-1)cn} \leq \frac{g}{n}$ from which it follows $\overline{q}_{NS}^{A^{ST^*}} \geq \frac{g}{n}$ and $\pi_S^{A^*}(n) \geq \pi_{NS}^{A^{ST^*}}(n-1)$. Differentiating $\underline{q}_S^{A^{ST^*}}$ with respect to d gives $-\frac{\sqrt{2}}{2\sqrt{(n-1)dcn}} < 0$. Hence, $\underline{q}_S^{A^{ST^*}}$ decreases in the collateral damage d. We can solve $\underline{q}_S^{A^{ST^*}} = \frac{g}{n}$ for d to find the lowest possible collateral damage level for which $\underline{q}_S^{A^{ST^*}} \leq \frac{g}{n}$ holds. This gives $\overline{d}_{\min}^{ST}(n-1) = \frac{(bn-cg-1)^2}{2cn(n-1)}$. Hence, the A-equilibrium constitutes an effective and credible threat if $\overline{d}_{\min}^{ST}(n-1) \leq d$. Comparing $\overline{d}_{\min}^{ST}(n-1)$ with $\hat{d}(n)$ gives

$\overline{d}_{\min}^{ST}(n-1) - \hat{d}(n) = -\frac{(bn-1)\left(bn^2-bn-cg-n+1\right)}{cn} > 0$ given condition $C_1 := b \geq \frac{1}{n}$ and condition $C_2 := g > \frac{bn^2}{c}$, as derived in the paper. Hence, the threshold ranking is $\overline{d}(n) < \hat{d}(n) < \overline{d}_{\min}^{ST}(n-1)$.

A.5 Proof of Proposition 3

We consider stability of the grand coalition for the NC-scenario in the repeated game assuming that the A-equilibrium is played during full cooperation, while the G-equilibrium is played as punishment in the non-cooperative equilibrium. That is, $\underline{d}(n) \leq d < \overline{d}(1)$ must hold. $\underline{d}(n)$ is the same as reported in Proposition 1, while substituting $k = 1$ in $\overline{d}(k)$, we have $\overline{d}(1) = \frac{(bn-gc-1)^2 n}{2c(n-1)}$.

Inserting payoff levels in Eq. (7), gives the minimum required discount factor for stability $\delta_{\min}^G = \frac{(bn-cg-1)(bn-cg-2n+1)}{2n(n-1)\left(bcg-nb^2+cd+b\right)}$ which decreases in parameter d as both nominator and denominator are positive due to condition $C_2 := g > \frac{bn^2}{c}$ and the denominator increases in d.

A.6 Proof of Proposition 4 and Renegotiation-Proof Punishment

We consider stability of the grand coalition for the ST-scenario in the repeated game assuming that the A-equilibrium is played during full cooperation. The G-equilibrium can be used as effective punishment under the same conditions identified for the NC-scenario and derived in Appendix A.5, giving stability for $\delta \geq \delta_{\min}^G$. In the ST-scenario, an additional punishment option arises, with the A-equilibrium being

played and the $n-1$ former signatories acting as Stackelberg leaders. As showed in Appendix A.4, this punishment can work only if the free-rider is now contributing more than $\frac{g}{n}$. This is the case if $\overline{d}^{ST}_{\min}(n-1) \leq d$, with the punishment payoff being $\pi^{A^{ST^*}}_{NS}(n-1)$. Comparing $\overline{d}^{ST}_{\min}(n-1)$ with $\overline{d}(1)$ gives.

$\overline{d}(1) - \overline{d}^{ST}_{\min}(n-1) = \frac{(n+1)(bn-cg-1)^2}{2cn} > 0$. Hence, the threshold ranking is $\overline{d}(n) < \overline{d}^{ST}_{\min}(n-1) < \overline{d}(1)$.

Inserting payoff levels in Eq. (8), gives the minimum required discount factor for stability

$$\delta^{A}_{\min} = \frac{(bn-cg-1)(bn-cg-2n+1)}{2\left(b^2n^2+cdn^2-bcgn-bn^2+\sqrt{2cdn(n-1)}(bn-1)-cdn+cgn-bn+n\right)}$$

Differentiating $\delta^{A}_{\min}$ with respect to the collateral damage parameter d, we find $\frac{\partial \delta^{A}_{\min}}{\partial d} = \frac{(cg-bn+1)(bn-cg-2n+1)cn(n-1)\left(\sqrt{2}(bn-1)+2\sqrt{cdn(n-1)}\right)}{4\left(b^2n^2+cdn^2-bcgn-bn^2+\sqrt{2cdn(n-1)}(bn-1)-cdn+cgn-bn+n\right)^2\sqrt{cdn(n-1)}} < 0$. The denominator is positive. The nominator is negative due to condition $C_1 := b \geq \frac{1}{n}$ and condition $C_2 := g > \frac{bn^2}{c}$. Hence, $\delta^{A}_{\min}$ decreases in d.

The grand coalition implements the A-equilibrium if $\overline{d}(n) \leq d$. According to Farrell and Maskin (1989), the following three conditions need to hold in order to support the grand coalition as a renegotiation-proof equilibrium.

(1) $\frac{\pi^{A^*}_{S}(n)}{1-\delta} \geq \pi^{F}_{i} + \frac{\delta \pi^{A^{ST^*}}_{NS}(n-1)}{1-\delta}$.

(2) $\pi^{A^{ST^*}}_{NS}(n-1) \geq \pi^{G^*}_{NS}(1)$.

(3) $\pi^{A^*}_{S}(n) \leq \pi^{A^{ST^*}}_{S}(n-1)$.

The first condition requires that the weak punishment with A-equilibrium is deterrent. The second condition requires that once the weak punishment starts, it is preferred by the punished player to the harsh punishment with the G-equilibrium. The third condition requires that the players conducting the punishment are not worse off during punishment than during cooperation.

Condition 1 is condition (8) in the text with $\delta \geq \delta^{A}_{\min}$. The second condition holds because $\pi^{A^{ST^*}}_{NS}(n-1) \geq \tilde{\pi}^{G^*}_{NS}(n-1)$ according to Eq. (4) in the text if $\overline{d}^{ST}_{\min}(n-1) \leq d$ and $\tilde{\pi}^{G^*}_{NS}(n-1) \geq \pi^{G^*}_{NS}(n-1) > \pi^{G^*}_{NS}(1)$ where the first inequality follows from the definition of $\tilde{\pi}^{G^*}_{NS}(n-1)$ in Eq. (4) and the second inequality follows from the fact that it can be shown that non-signatories' payoff in the G-equilibrium increases with the coalition size k. The third condition holds as argued in the text because $\underline{q}^{A^{ST^*}}_{S}(n-1) \leq \frac{g}{n}$ and $\overline{q}^{A^{ST^*}}_{NS}(n-1) \geq \frac{g}{n}$ if $\overline{d}^{ST}_{\min}(n-1) \leq d$.

Acknowledgements We would like to thank one anonymous reviewer. Their constructive comments helped improving the quality of the manuscript. Financial support for research visits by the University of Graz and Hamburg is greatly acknowledged.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Research involving human participants and/or animals Not applicable.

Informed consent Not applicable.

References

Aldy JE, Zeckhauser R (2020) Three prongs for prudent climate policy. South Econ J 87(1):3–29

Aldy JE, Felgenhauer T, Pizer WA, Tavoni M, Belaia M, Borsuk ME, Gosh A, Heutel G, Heyen D, Horton J, Keith D, Merk C, Moreno-Cruz J, Reynolds JL, Ricke K, Rickels W, Shayegh S, Smith W, Tilmes S, Wagner G, Wiener JB (2021) Social science research to inform solar geoengineering. Science 374(6569):815–818

Asheim GB, Holtsmark B (2009) Renegotiation-proof climate agreements with full participation: conditions for pareto-efficiency. Environ Resour Econ 43:519–533

Asheim GB, Froyn CB, Hovi J, Menz FC (2006) Regional versus global cooperation for climate control. J Environ Econ Manag 51(1):93–109

Barrett S (1994) Self-enforcing international environmental agreements. Oxf Econ Pap 46:878–894

Barrett S (2002) Consensus treaties. J Inst Theor Econ 158(4):529–547

Barrett S (2014) Solar geoengineering's brave new world: thoughts on the governance of an unprecedented technology. Rev Environ Econ Policy 8:249–269

Barrett S, Lenton T, Millner A, Tavoni A, Carpenter S, Anderies JM, Chapin FS III, Crépin A, Daily G, Ehrlich P, Folke C, Galaz V, Hughes T, Kautsky N, Lambin EF, Naylor R, Nyborg K, Polasky S, Scheffer M, Wilen J, Xepapadeas A, de Zeeuw A (2014) Climate engineering reconsidered. Nat Clim Change 4:527–529

Bodansky D (2013) The who, what, and wherefore of geoengineering governance. Clim Change 121(3):539–551

Caldeira K, Bala G, Cao L (2013) The science of geoengineering. Annu Rev Earth Planet Sci 41(1):231–256

Cherry TL, Kroll S, McEvoy DM, Campoverde D, Moreno-Cruz JB (2023) Climate cooperation in the shadow of solar geoengineering: an experimental investigation of the moral hazard conjecture. Environ Polit 32(2):362–370

Crutzen P (2006) Albedo enhancement by stratospheric sulfur injections: a contribution to resolve a policy dilemma? Clim Change 77(3–4):211–220

Farrell J, Maskin E (1989) Renegotiation in repeated games. Games Econ Behav 1:327–360

Finus M, 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

Finus M, Furini F (2023) Global climate governance in the light of geoengineering: a shot in the dark? J Environ Econ Manag 122:102854
Finus M, Rundshagen B (1998a) Renegotiation-proof equilibria in a global emission game when players are impatient. Environ Resour Econ 12(3):275–306
Finus M, Rundshagen B (1998b) Toward a positive theory of coalition formation and endogenous instrumental choice in global pollution control. Public Choice 96(1–2):145–186
Haywood JM, Jones A, Bellouin N, Stephenson D (2013) Asymmetric forcing from stratospheric aerosols impacts Sahelian rainfall. Nat Clim Change 3(7):660–665
Heyen D, Horton J, Moreno-Cruz J (2019) Strategic implications of counter-geoengineering: clash or cooperation? J Environ Econ Manag 95:153–177
Hovi J, Ward H, Grundig F (2015) Hope or despair? formal models of climate cooperation. Environ Resour Econ 62:665–688
Irvine PJ, Sriver RL, Keller K (2012) tension between reducing sea-level rise and global warming through solar-radiation management. Nat Clim Change 2(2):97–100
Irvine PJ, Emanuel K, He J, Horowitz L, Vecchi G, Keith D (2019) Halving warming with idealized solar geoengineering moderates key climate hazards. Nat Clim Change 9:295–299
Kravitz B, Robock A, Oman L, Stenchikov G, Marquardt AB (2009) Sulfuric acid deposition from stratospheric geoengineering with sulfate aerosols. J Geophys Res Atmos 114(14):1–7
Kravitz B, MacMartin DG, Robock A, Rasch PJ, Ricke KL, Cole JNS, Curry CL, Irvine PJ, Ji D, Keith DW, Kristjansson JE, Moore JC, Muri H, Singh B, Tilmes S, Watanabe S, Yang ST, Yoon JH (2014) A multi-model assessment of regional climate disparities caused by solar geoengineering. Environ Res Lett 9:074013
Marrouch W, Chaudhuri AR (2016) International environmental agreements: doomed to fail or destined to succeed? A review of the literature. Int Rev Environ Resour Econ 9:245–319
Millard-Ball, (2012) The Tuvalu Syndrome. Can geoengineering solve climate's collective action problem? Clim Change 110:1047–1066
Moreno-Cruz JB (2015) Mitigation and the Geoengineering Threat. Resour Energy Econ 41:248–263
Moreno-Cruz JB, Keith DW (2013) Climate policies under uncertainty: a case for solar geoengineering. Clim Change 121:431–444
National Academies of Sciences, Engineering, and Medicine (2021) Reflecting Sunlight: Recommendations for Solar Geoengineering Research and Research Governance. The National Academies Press, Washington, DC
Parker A, Irvine PJ (2018) The risk of termination shock from solar geoengineering. Earth's Future 6:456–467
Parker A, Horton JB, Keith DW (2018) stopping solar geoengineering through technical means: a preliminary assessment of counter-geoengineering. Earth's Future 6(8):1058–1065
Reynolds JL (2019) The governance of solar geoengineering: managing climate change in the anthropocene. Cambridge University Press, Cambridge
Ricke K, Morgan G, Allen M (2010) regional climate response to solar radiation management. Nat Geosci 3:537–541
Ricke KL, Moreno-Cruz JB, Caldeira K (2013) Strategic incentives for climate geoengineering coalitions to exclude broad participation. Environ Res Lett 8:1–11
Rickels W, Quaas MF, Ricke K, Quaas J, Moreno-Cruz JB, Smulders S (2020) Who turns the global thermostat and by how much? Energy Econ. 91:104852
Robock A, Marquardt A, Kravitz B, Stenchikov G (2009) Benefits, risks, and costs of stratospheric geoengineering. Geophys Res Lett 36(19):L19703
Shepherd J, Caldeira K, Haigh J, Keith D, Launder B, Mace G, MacKerron G, Pyle J, Rayner S, Redgwell C (2009) Geoengineering the climate: science, governance and uncertainty. The Royal Academy, London
Stephens JC, Kashwan P, McLaren D, Surprise K (2021) The risks of solar geoengineering research. Science 372(6547):1161–1161
Urpelainen J (2012) Geoengineering and global warming: a strategic perspective. Int Environ Agreem Polit Law Econ 12(4):375–389
Visioni D, Slessarev E, MacMartin DG, Mahowald NM, Goodale CL, Xia L (2020) What goes up must come down: impacts of deposition in a sulfate geoengineering scenario. Environ Res Lett 15:094063
Weitzman ML (2015) A voting architecture for the governance of free-driver externalities, with application to geoengineering. Scand J Econ 117(4):1049–1068