CESifo WORKING PAPERS

7180
2018

July 2018

Strategic Implications of Counter-Geoengineering: Clash or Cooperation?

Daniel Heyen, Joshua Horton, Juan Moreno-Cruz

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editors: Clemens Fuest, Oliver Falck, Jasmin Gröschl
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7180*
Category 10: Energy and Climate Economics

Strategic Implications of Counter-Geoengineering: Clash or Cooperation

Abstract

Solar geoengineering has received increasing attention as an option to temporarily stabilize global temperatures. A key concern surrounding these technologies is that heterogeneous preferences over the optimal amount of cooling combined with low deployment costs may allow the country with the strongest incentive for cooling, the so-called free-driver, to impose a substantial externality on the rest of the world. We analyze whether the threat of counter-geoengineering technologies capable of negating the climatic effects of solar geoengineering can overcome the free-driver problem and tilt the game in favor of international cooperation. Our game-theoretical model of asymmetric countries allows for a rigorous analysis of the strategic interaction surrounding solar geoengineering and counter-geoengineering. We find that the free-driver outcome becomes unstable once counter-geoengineering is available, but not always with benign effects. The presence of counter-geoengineering leads to either a climate clash where countries engage in a non-cooperative escalation of opposing climate interventions (negative welfare effect), a moratorium treaty where countries commit to abstain from either type of climate intervention (indeterminate welfare effect), or cooperative deployment of solar geoengineering (positive welfare effect). We show that the outcome depends crucially on the degree of asymmetry in temperature preferences between countries.

JEL-Codes: Q540, H410, D620, D020, D740.

Keywords: climate intervention, solar geoengineering, counter-geoengineering, free-driver, strategic conflicts, game theory, cooperation, externality, global warming, international environmental agreements.

*Daniel Heyen**
Grantham Research Institute
London School of Economics
United Kingdom – London WC2A 2AE
d.heyen@lse.ac.uk

Joshua Horton
Belfer Center, Harvard Kennedy School
USA - Cambridge, MA 02138
joshua_horton@hks.harvard.edu

Juan Moreno-Cruz
School of Environment, Enterprise and Development, University of Waterloo
Canada - Waterloo, Ontario N2L 3G1
juan.moreno-cruz@econ.gatech.edu

*corresponding author

July 23, 2018

We gratefully acknowledge funding from the ESRC Centre for Climate Change Economics and Policy and Grantham Foundation for the Protection of the Environment and thank the organisers of the Harvard Solar Geoengineering Research Residency. The first author gratefully acknowledges support from the German Research Foundation (DFG), grant HE 7551/1. This paper has benefited substantially from discussions and comments at the following events: seminars at FEEM Milan and the Institute for Science, Innovation and Technology Oxford; the EAERE 2017 conference in Athens, Climate Engineering Conference CEC 17 in Berlin, EENR 2018 workshop in Orléans, WCERE 2018 conference in Gothenburg and the GRI workshop at LSE.

1 Introduction

One option for addressing climate change that is gaining increased attention is Solar Geoengineering (SG), also known as Solar Radiation Management (SRM) (National Research Council 2015). SG aims at (partially) compensating the global warming caused by increased atmospheric levels of greenhouse gases by either releasing cooling particles in the stratosphere (stratospheric aerosol injection) or modifying marine cloud reflectivity (marine cloud brightening). While an optimally designed and implemented SG scheme appears to have the potential to reduce global temperature damages (Moreno-Cruz et al. 2012; Keith and MacMartin 2015; National Research Council 2015), there are concerns that SG's potential benefits are reduced and possibly even reversed in a decentralized world of international 'anarchy'. A key fear is that presumably low deployment costs (McClellan et al. 2012) together with asymmetric preferences over the optimal global temperature change (Heyen et al. 2015) may result in unilateral SG deployment that harms the rest of the world (Horton 2011; National Research Council 2015; Pasztor et al. 2017). This has been termed the "free-driver" problem (Weitzman 2015).[1]

Against this backdrop of potentially welfare deteriorating strategic incentives surrounding a potentially beneficial technology, a recent paper (Parker et al. 2018) explores the idea of counter-geoengineering (CG), or a set of technologies that would give countries threatened by or subject to the free-driver's whims a tool for quickly negating what they regard as harmful SG. In principle, CG could be either 'neutralizing' (i.e. neutralizing the cooling SG particles, for instance by injecting a base to counteract the sulphate aerosols most commonly considered for SG) or 'countervailing' (i.e. the release of a warming agent such as difluoromethane to reverse the effects of SG particles). The reason why the availability of such CG capabilities might prove beneficial is obviously not because further global warming is globally desirable; rather, the very availability of CG might deter the free-driver from unilateral SG deployment and instead promote international cooperation on climate interventions. If CG has this potential to steer climate technology use to overall beneficial levels, then there is a case for countries to invest in CG today as a deterrent to future unilateral SG use.

The present paper provides a first rigorous analysis of the strategic effects of introducing SG and CG into an otherwise standard model of climate economics. We regard SG and CG as two separate and contrasting forms of *climate intervention*. With this understanding, we model intervening in the climate (via either SG or CG) as a public good game: the operational costs of any climate intervention are borne only by the deploying country, whereas the resulting global temperature change affects all countries.

[1] The 'free-driver' terminology emphasizes two things: first, the public good nature of interventions in the global climate, i.e. non-excludability and non-rivalness; and second, the potential for a single actor to get in the 'driver seat' (due to low deployment costs) and shape the global climate as she wishes (as contrasted with the well-known 'free-rider' problem). To emphasize heterogeneous preferences, Weitzman (2015) also refers to SG as a 'public gob', that is, a public good or bad, depending on circumstances.

The latter is captured by a non-monotonic benefit function that exhibits an optimal level of global temperature change, and a key assumption is that countries disagree about the optimal temperature change and therefore their preferred climate intervention. We give countries two distinct options for cooperation. The first is a *deployment treaty* where countries jointly decide on the climate intervention that maximizes the coalition's overall payoff. The second option, which constitutes one of the novel contributions of the present paper, is a *moratorium treaty*. In a moratorium, an idea often raised in the geoengineering debate (Parker 2014; Victor 2008), countries commit themselves to abstain from any form of climate intervention. As usual, we assume each country individually determines its willingness to cooperate by comparing payoffs under alternative treaties to the non-cooperative outcome. We study how CG affects the incentives to cooperate by analyzing the game first when only SG is available and hence climate intervention is restricted to cooling; and second when CG is also available and countries are able to cool or warm.

Despite the simplicity of the setting we produce a rich set of findings. In the absence of CG, if countries are sufficiently different in their temperature preferences, the non-cooperative outcome is the free-driver equilibrium. If countries are similar in their temperature preferences, then the non-cooperative outcome is the usual free-rider equilibrium. In both cases, cooperation incentives are overall weak: The moratorium treaty is never supported by both countries and therefore unstable, and the deployment treaty is only stable for a relatively small set of parameter constellations. The effect of introducing CG is to render the free-driver equilibrium unstable: those who regard the free-driver's cooling as excessive now have a tool to counteract it, and they use it. Absent the opportunity to cooperate, this results in a 'climate clash', an escalation of cooling by SG and warming by CG that typically has no winners and is overall sharply detrimental. If cooperation is an option, however, this bleak outlook of CG in a non-cooperative world may encourage countries to work together. In particular, the free-driver, typically unwilling to cooperate in the absence of CG, may be ready to compromise on climate interventions. Yet cooperation is not assured, and the outcome might still be a destructive climate clash. And even if cooperation does occur, it might take the form of a moratorium, which is worse than the free-driver outcome in overall welfare terms when climate damages are sufficiently high. The outcome depends crucially on the degree of asymmetry in temperature preferences between countries.

Our paper is related to two strands of literature. The first is the quickly emerging literature on geoengineering (Klepper and Rickels 2014; National Research Council 2015) that generally emphasizes the importance of international cooperation and governance (e.g. Horton 2011; Barrett et al. 2014). Particularly relevant in this context is Weitzman (2015), who identifies the free-driver problem as a significant challenge for the governance of geoengineering, and Emmerling and Tavoni (2017), who quantify free-driving in an integrated assessment model with game-theoretic interactions. Emmerling and

Tavoni (2017) interpret free-driving as over-provision relative to the cooperative (global first-best) solution; free-driving in this sense can also occur in symmetric settings as they explicitly account for the side effects of SG. Our interpretation differs and is closer to Weitzman (2015) in allowing for heterogeneous preferences over climate interventions. We extend Weitzman (2015) and Emmerling and Tavoni (2017) by including the possibility of CG. The first paper that has put CG center-stage is Parker et al. (2018); we extend their initial findings by deriving our results using a calibrated model (including climate impacts and operational costs) and a richer game-theoretic interaction, including the possibility of cooperation. Other papers related to our work consider the strategic implications of SG for mitigation (Moreno-Cruz 2015; Emmerling et al. 2016; Manoussi and Xepapadeas 2015; Urpelainen 2012; Millard-Ball 2012) and R&D incentives (Heyen 2016).

Our paper also contributes more broadly to the environmental economics literature on public goods, externalities and cooperation, see for instance Barrett (1994) and Finus (2008). The subtle and important role of heterogeneity in strategic environmental settings has been emphasized by Barrett (2001) and McGinty (2007). There are three innovations of our paper in this context. First, following Weitzman (2015) we allow for the over-provision of a public good by modelling non-monotonic benefit functions with *heterogeneous optimal levels*, a feature not present in other asymmetric public good settings; in contrast to Weitzman (2015), however, we include deployment costs (giving rise to much richer findings) and situate this discussion in a standard public-good setting with a smooth benefit function. The second innovation of our paper is to consider CG, which essentially allows agents to make 'negative' contributions to a public good, an aspect that may be of interest in future research beyond geoengineering. Finally, the third contribution of our paper to the environmental economics literature on public goods and cooperation is to introduce a moratorium treaty where agents agree to abstain from contributions to the public good altogether. This option, which has not received attention in the literature (unsurprising in light of the focus on symmetric settings), may be of general interest for the analysis of strategic interaction of asymmetric agents, in particular when the option of side-payments is not available.

We proceed as follows. Sec. 2 presents the model components in detail, with a focus on the case of two countries. Sec. 3 analyzes the deployment stage, in particular the non-cooperative outcomes both with and without CG; these non-cooperative outcomes are the reference points for countries when choosing whether to cooperate, discussed in Sec. 4. Sec. 5 calibrates the model and Sec. 6 covers the general case of n countries. Sec. 7 concludes.

2 The Model

In our model two asymmetric countries decide on climate intervention levels, i.e. changes to global temperatures using either SG or CG. The general case with n countries is covered in section 6. Because changes to global temperatures affect every country, we model climate intervention as a public good provision game.

2.1 Timing of Events

Figure 1 gives a graphical representation of the two stages of the game. In the first stage the two countries can cooperate by forming a climate intervention treaty. The two available options are a *moratorium treaty*, in which the countries commit themselves to deploy neither SG nor CG, and a *deployment treaty*, in which the countries within the coalition commit themselves to choose technology levels so as to maximize the coalition's sum of payoffs. By definition, the deployment treaty implements the climate intervention that maximizes global welfare. If neither treaty comes into effect, countries in the second period choose their climate intervention levels *non-cooperatively*. In order to assess the game-changing potential of CG we contrast two cases. First, the 'SG only' case when CG is not available and hence climate interventions are restricted to cooling. We then compare this with the 'CG available' case in which countries have the option to increase or decrease global mean temperatures. The non-cooperative outcome depends on whether CG is available or not, and this will in turn have implications for the attractiveness of the treaties.

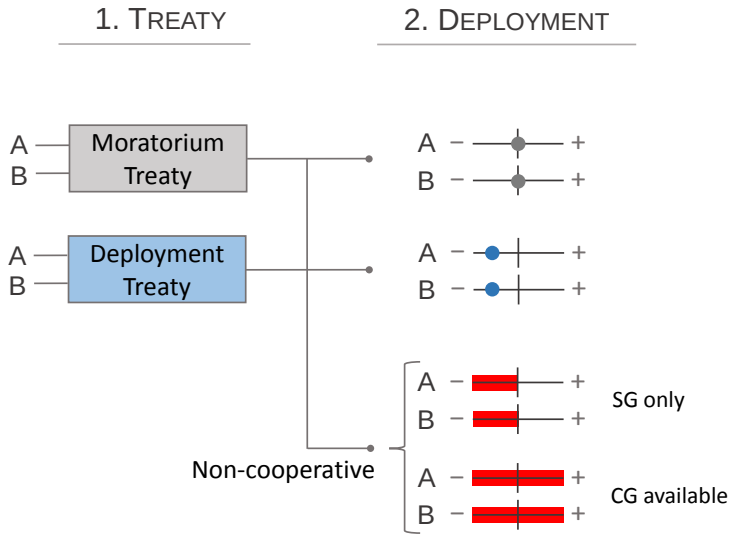


Figure 1: Timing of the game, illustrated for two countries labelled A and B. In the first stage the countries decide whether to cooperate via a moratorium treaty (countries commit to deploy neither SG nor CG) or a deployment treaty (countries commit to implement the coalition's optimal climate intervention). If neither treaty comes into effect, countries choose their technology levels (non-positive in the 'SG only' case, any level in the 'CG available' case) non-cooperatively.

2.2 Definitions and Assumptions

Climate intervention levels $g_i \in \mathbb{R}$, $i = A, B$, are measured in terms of the resulting temperature change. The global average temperature under climate change T_0 – the *status quo* temperature countries face when making their climate intervention choice – is normalized to zero, $T_0 = 0$.[2] Hence, the *change in global average temperature* T due to climate intervention is

$$T = g_A + g_B. \tag{1}$$

We assume that costs and benefits are quadratic (Barrett 1994; McGinty 2007; Finus and Rübbelke 2013; Diamantoudi and Sartzetakis 2006; Heyen 2016).[3] The costs are

$$C(g_i) = \frac{c}{2} g_i^2 \quad , \; i = A, B \tag{2}$$

with $c > 0$.[4] We assume for simplicity that SG and CG have the same country-independent cost structure. The climate benefits are

$$B_i(T) = -\frac{b}{2} (T_i - T)^2 \quad , \; i = A, B \tag{3}$$

with $b > 0$.[5] In contrast to private operational costs, the benefit function reflects the public good nature of the climate intervention: Benefits depend on the global average temperature T and hence on the climate intervention levels of both countries. The benefits are highest at $T = T_i$ which justifies calling T_i *country i's preferred temperature.* For a country that suffers from climate change, which is the typical situation, $T_i < 0$.

Country i's *payoff* under the climate intervention profile $g = (g_i)_{i=A,B}$ is

$$\pi_i(g) = B_i(T) - C(g_i) \; . \tag{4}$$

[2]This temperature includes the positive effects of any previous mitigation efforts. We do not model mitigation explicitly. The reason is that we are interested in the strategic interaction surrounding SG and CG that can be expected to unfold on a fairly short timescale: climate interventions would have an almost immediate temperature response effect, whereas the effects of mitigation need much longer to materialize.

[3]The calibration in Sec. 5 justifies this assumption.

[4]For the analysis within a public good framework it is crucial to focus on those costs that are borne by each country individually. In the context of a climate intervention these are the direct operational costs of modifying the global climate. Indirect costs that are climate-related are captured within the non-monotonic benefit function B. Indirect costs not related to climate indicators, e.g. health impacts from sulfur particles, are not incorporated in this simple model; including them would likely only strengthen our results as they add another source of external effects. Also see the discussion in section 7.

[5]We make in this paper the simplifying assumption that countries assess climate outcomes solely in terms of temperature levels. This is a strong assumption as precipitation and presumably other climate indicators will be relevant as well. We consider the restriction to temperatures productive in the context of the present paper because we are interested in the strategic implications when countries disagree about what an optimal climate intervention would be. And given current knowledge it is plausible that different regions form different preferences over climate interventions based on regional temperature outcomes, whereas precipitation outcomes are still prone to significant uncertainties. Also see the discussion in section 7.

A central component of the model is to allow for different T_i and hence heterogeneous preferences over the optimal amount of climate intervention.[6] Without loss of generality let $T_A \leq T_B$. Accordingly, from now on A is the country that favours relatively strong deployment of SG, whereas country B prefers moderate cooling, if any. We define the *mean optimal temperature change* $\bar{T} = \frac{T_A+T_B}{2}$ and write

$$T_A = \bar{T} - \Delta \quad , \quad T_B = \bar{T} + \Delta, \quad \text{where} \quad \Delta = \frac{T_B - T_A}{2}. \tag{5}$$

We refer to Δ as the *asymmetry parameter* which equals the standard deviation of the optimal temperature changes T_A and T_B. For $\Delta = 0$, both countries agree on how much the climate ought to change; the higher Δ, the higher the disagreement between the two countries in terms of how to set the global thermostat.[7] One of the advantages of this definition of Δ is that it can easily be extended to the general n country case that we discuss in section 6.

Our analysis requires that we make an assumption regarding the overall desirability of some amount of SG. That is, at the time countries consider a climate intervention through SG (or CG), past efforts at mitigation and 'negative emissions' such as bioenergy with carbon capture and storage (BECCS) have proved insufficient to curb temperatures.

Assumption 1. *The world without any climate intervention is on average too warm,* $\bar{T} < 0$.

In particular, $T_A < 0$. We do not impose assumptions on T_B, so country B might prefer a warmer climate, $T_B > 0$.

2.3 The Decision to Enter a Treaty

We model a climate intervention treaty in line with the literature on international environmental agreements (e.g. Barrett 1994, 2001; Finus 2008). Instead of joint decisions on emission abatement levels, countries in a coalition jointly decide on *climate intervention levels*. In the first type of treaty, the **deployment treaty**, countries choose the amount of SG that maximizes the coalition's total payoff, i.e. the sum of payoffs

[6] It is worth emphasizing that our model's approach to capture heterogeneity in terms of different optimal *levels* of a public good is novel. With the exception of Heyen (2016), the typical focus in the literature has been to assume the same optimal level of the public good but different slopes of the marginal benefit function (e.g. McGinty 2007).

[7] A simple illustrative example provides evidence that countries may prefer different global average temperatures. Assume country A and country B to have pre-industrial temperatures of $16°C$ and $10°C$, respectively. Further assume that climate change increases temperatures in both countries by $3°C$. The climate impact literature suggests that growth rates are maximal for a certain *universal*, i.e. country-independent, temperature; Burke et al. (2015) finds growth rates to follow a quadratic inverted U shape with a maximum at $13°C$. If country A and country B both regard $13°C$ as their optimal temperature, then we have in our notation $T_A = -6°C$ and $T_B = 0°C$, resulting in $\Delta = 3°C$. Such a universal optimal temperature, even if countries' preferences are only partially determined by it, provides a strong argument for heterogeneous preferences over climate intervention in a world of heterogeneous baseline temperatures.

across its members. One of the innovations of our paper is to allow for a second type of treaty, the **moratorium treaty**. Here, the countries commit themselves to abstain from climate interventions altogether, $g_i = 0$. One reason to consider this additional type of treaty is the importance of a moratorium in the geoengineering debate (Victor 2008; Parker 2014); furthermore, the aspect of winners and losers is particularly pronounced in the present paper and a moratorium treaty – by definition less appealing than a deployment treaty in terms of the sum of payoffs – might possibly be attractive due to its distributional implications.

In this context it is important to note that we do not include side payments (also known as transfers) in our model. The importance of side payments in increasing the attractiveness of cooperation has often been noted, especially for asymmetric countries (McGinty 2007; Barrett 2001).[8] Yet we often observe that international treaties designed to overcome domestic interests face strong opposition and that side payments in particular are often seen as politically unacceptable (Gampfer et al. 2014; Diederich and Goeschl 2017). This suggests that studying incentives for cooperation that do not rely on transfers is an important benchmark. The deployment treaty and moratorium treaty are two specific, yet salient, forms of cooperation in the absence of transfers.

We model a country's choice regarding treaty participation as the submission of a ranked ordering of the country's preference over the three possible outcomes, i.e. the non-cooperative outcome, the deployment treaty and the moratorium treaty. Stability of a treaty is defined relative to the non-cooperative outcome. With only two countries, the condition for a coalition to be *stable* reduces to *internal stability*, i.e. whether both countries want to be a member of the coalition compared to the non-cooperative Nash solution. Furthermore, with only two countries it does not make a difference whether the coalition is modelled as an *open membership* game, in which a country can enter a coalition without the other members' invitation, or an *exclusive club*, where access to a coalition is conditional on the members' consent Ricke et al. (2013). See section 6 for a treatment of the case with n countries. Note that coalitions that are stable in the 'SG only' scenario need not be stable under the 'CG available' case, and vice versa.

As we will show below, the deployment and moratorium treaty can both be stable at the same time. While equilibrium selection is not a focus of our paper, we aim to make the analysis in the $n = 2$ case as easy to follow as possible and hence make the following tie-breaking assumption.

Assumption 2 (Tie-breaking rule)**.** *If both treaties are stable, i.e. if both countries are willing to enter either of the two treaties, then the one most preferred by both countries comes into effect if there is such a clear ordering. If countries disagree on the preferred order, we assume that the moratorium treaty comes into effect.*

[8] Indeed, in the absence of negotiation and transaction costs, it is well known that transfer schemes exist to ensure that the socially optimal configuration makes each party better off (Coase 1960).

The rationale for this tie-breaking rule is that the status quo of non-deployment may be a focal point, for instance because an error of geoengineering 'commission' is assumed to be worse than an error of geoengineering 'omission' (Weitzman 2015). We will see below that equilibrium selection has a significant impact on the analysis.

We proceed with the equilibrium analysis. We first discuss the non-cooperative equilibria, the fallback option when none of the treaties comes into effect. The relative attractiveness of the non-cooperative case, in turn, determines countries' willingness to enter the moratorium and/or deployment treaty.

3 Optimal Deployment and Non-cooperative Equilibria

We solve the equilibrium via backward induction and thus begin our description with the climate intervention deployment stage. The countries simultaneously choose $g_i \in \mathbb{R}$, $i = A, B$. In the 'SG only' case, deployment is restricted to cooling, $g_i \leq 0$. When CG is available, any temperature level $g_i \in \mathbb{R}$ is feasible.[9]

3.1 Global Optimum

We denote by $(g_i^{**})_{i=A,B}$ the socially optimal configuration that maximizes global welfare $\pi(g) = \pi_A(g) + \pi_B(g)$. The solution to this problem following standard procedure is

$$g_i^{**} = \frac{2b}{4b+c}\bar{T} \qquad , \; i = A, B \; . \tag{6}$$

It is efficient that both countries deploy the same amount due to the homogeneous cost structure. Owing to $\bar{T} < 0$ (Assumption 1), the socially optimal deployment scheme features SG deployment by both countries. Whether CG is available or not has, therefore, no implications for the socially optimal deployment profile.

3.2 Non-cooperative equilibria

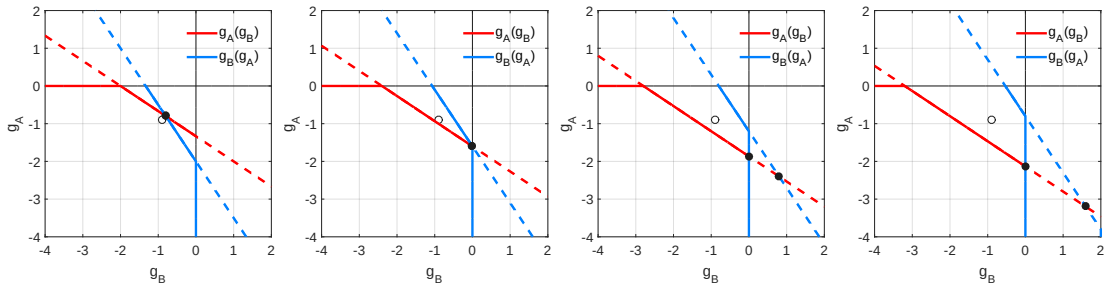
The first step in determining the non-cooperative Nash equilibria is to calculate the best response functions. The conceptually simplest case is when CG is available and hence $g_i \in \mathbb{R}$ unrestricted. In this case, the best response of country i to the other country's climate intervention level g_{-i} is characterized by the first-order condition $d\pi_i(g_i; g_{-i})/dg_i = 0$. In the 'SG only' case, we also need to check whether the non-

[9] The absence of an upper limit on the level of CG corresponds to 'countervailing' CG (Parker et al. 2018), e.g. the release of a potent GHG. The maximal amount of 'neutralizing' CG, in contrast, would be a function of the deployed SG level. We find that CG levels are smaller than SG levels, see below, so that in the context of the present paper it is inconsequential whether we understand CG as countervailing or neutralizing. Also note that we assume SG and CG are deployed simultaneously. Other modelling assumptions, for instance that CG can only be deployed after SG has initiated, are possible and should be explored in future research.

positive constraint binds. We get the best response function

$$g_i(g_{-i}) = \begin{cases} \min\left\{\frac{b}{b+c}\left(T_i - g_{-i}\right), 0\right\} & \text{SG only} \\ \frac{b}{b+c}\left(T_i - g_{-i}\right) & \text{CG available} \end{cases} \tag{7}$$

Figure 2 shows how the best response functions depend on the asymmetry Δ.



(a) $\Delta = 0$ K (b) $\Delta = 0.4$ K (c) $\Delta = 0.8$ K (d) $\Delta = 1.2$ K

Figure 2: Best response functions (country A in red, country B in blue) for different asymmetry levels Δ. In all plots $b = 2\,\$\text{K}^{-2}$, $c = 1\,\$\text{K}^{-2}$, $\bar{T} = -2\,\text{K}$. The solid lines and dashed lines show the best response functions without CG and with CG, respectively. The unfilled circle indicates the socially optimal benchmark (g_A^{**}, g_B^{**}). Under these parameter settings we get an asymmetry threshold of $\bar{\Delta} = 0.4\,\text{K}$, see (8). For $\Delta > \bar{\Delta}$ the equilibrium outcome, indicated by a filled black circle, depends on whether CG is available or not.

We now summarize non-cooperative equilibria in the 'SG only' and 'CG available' scenarios and hence determine the game-changing effect of CG in the absence of cooperation possibilities. We define the *asymmetry threshold*

$$\bar{\Delta} := -\frac{c}{2b+c}\bar{T}\ . \tag{8}$$

The asymmetry threshold plays an important role in the following discussion, as it helps explain which equilibria obtain under different conditions.

Proposition 1 (Game-changing potential of CG. Non-cooperative equilibria)**.** *There is a unique Nash equilibrium and the outcome depends on parameter settings and whether CG is available:*

(i) The 'SG only' case. For low levels of asymmetry, $\Delta < \bar{\Delta}$, there is a ***free-rider equilibrium*** *in which both countries engage in SG,*

$$g_A^* = \frac{b}{2b+c}\bar{T} - \frac{b}{c}\Delta \ < 0 \quad , \quad g_B^* = \frac{b}{2b+c}\bar{T} + \frac{b}{c}\Delta \ < 0\ . \tag{9}$$

For high levels of asymmetry, $\Delta \geq \bar{\Delta}$, we call the unique equilibrium a ***free-driver equilibrium*** *as only country A deploys SG,*

$$g_A^* = \frac{b}{b+c}T_A \quad , \quad g_B^* = 0\ . \tag{10}$$

(ii) *The 'CG available' case. For low levels of asymmetry, $\Delta < \bar{\Delta}$, there is no incentive to deploy CG. The unique equilibrium is therefore the free-rider outcome (9).*

For high levels of asymmetry, $\Delta \geq \bar{\Delta}$, we obtain a ***climate clash equilibrium*** *in which country A cools and, simultaneously, country B warms,*

$$g_A^* = \frac{b}{2b+c}\bar{T} - \frac{b}{c}\Delta \ < 0 \quad , \quad g_B^* = \frac{b}{2b+c}\bar{T} + \frac{b}{c}\Delta \ \geq 0 \ . \tag{11}$$

(iii) *The transformation from free-driver to climate clash is always detrimental for country A and detrimental for country B iff $b/c > \frac{1+\sqrt{5}}{2}$. Overall, the transformation is unambiguously detrimental.*

Proof. See appendix A. □

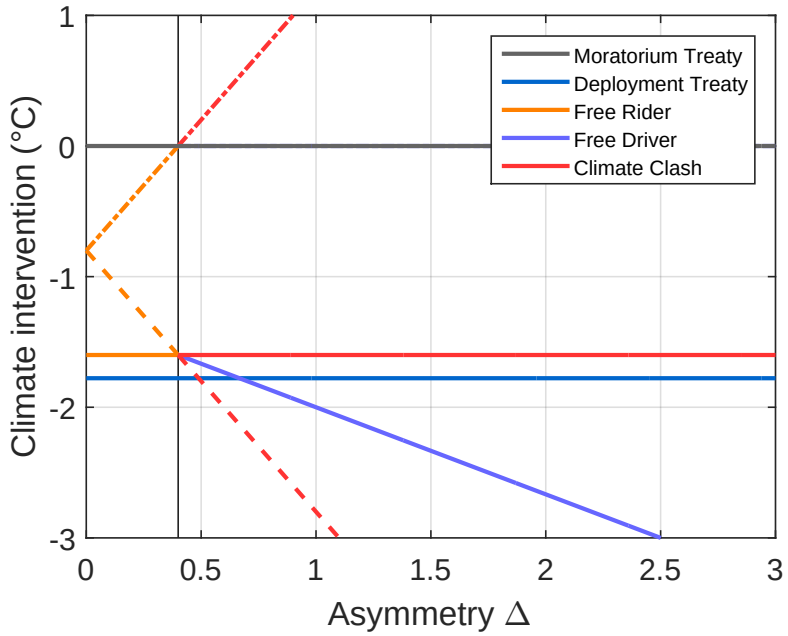


Figure 3: Climate intervention levels of the non-cooperative equilibria as a function of the asymmetry Δ. The parameter settings are as in Figure 2, i.e. $b = 2\,\$\mathrm{K}^{-2}$, $c = 1\,\$\mathrm{K}^{-2}$ and $\bar{T} = -2\,\mathrm{K}$. The vertical line is at the asymmetry threshold $\bar{\Delta}$. The dashed and dot-dashed lines represent the deployment by country A and country B, respectively, while the solid lines show net levels. For comparison we include the total climate intervention levels under the moratorium treaty (zero) and the deployment treaty (the total level is twice the amount in (6)).

Figure 3 shows climate intervention levels for both countries under the non-cooperative equilibria as a function of the asymmetry level Δ. For comparison we include the total SG level under the moratorium treaty (i.e. zero) and the deployment treaty. The free-driver SG level (solid purple line) depends on country A's optimal temperature change T_A but not on T_B and hence the cooling intensifies as the asymmetry level Δ increases. The total temperature change in the climate clash (solid red line) matches the free-rider

level (solid orange line) and is independent of the asymmetry level Δ, but is the result of ever diverging SG and CG levels (dashed and dot-dashed red lines) by country A and country B respectively.

The free-rider equilibrium is a well-known outcome in the literature; in particular, the symmetric case $\Delta = 0$ is of this type. The more interesting outcome in the 'SG only' case is the *free-driver* equilibrium. The terminology is from (Weitzman 2015) who develops the concept of over-provision of a public good in a setting without deployment costs and with a specific kinked utility function. Our definition coincides with the one in Heyen (2016). The defining characteristic of the free-driver equilibrium is that cooling is excessive from country B's perspective, $T \leq T_B$, and country A is essentially in control of the global thermostat. This excessive cooling does not necessarily imply that country B is worse off relative to a world without any climate intervention. Importantly for our analysis, the free-driver equilibrium becomes unstable once CG is available. The dominated country B now has a tool to counter the over-provision of the public good, and due to zero marginal costs (at the point of non-deployment), country B uses this tool. The best response of country A, in turn, is to increase her SG efforts. The only reason why the SG and CG levels are bounded in this escalation equilibrium is convex deployment costs.

This section has demonstrated that CG renders the free-rider equilibrium unstable, replacing it with a climate clash. This game-changing effect is overall detrimental as countries waste significant resources on SG and CG in an escalation of individually rational, yet overall harmful climate interventions. But might CG lead to a different outcome if cooperation is possible? The next section is dedicated to this question.

4 Incentives for Cooperation

This section analyzes the incentives to cooperate on climate intervention via either a deployment treaty or a moratorium treaty. We begin with the 'SG only' case in 4.1 and cover the 'CG available' case in section 4.2. All findings are illustrated in Figure 4.

4.1 Cooperation incentives when only SG is available

The non-cooperative deployment equilibria derived in the previous section (cf. Proposition 1) are the appropriate reference points when countries are deciding whether they are willing to cooperate by entering a moratorium or deployment treaty. We start with the low asymmetry case where non-cooperation would result in the free-rider outcome.

Proposition 2 (Cooperation incentives in the 'SG only' case. Low asymmetry, $\Delta < \bar{\Delta}$)**.** *Country A prefers the deployment treaty over the free-rider equilibrium irrespective of the level of asymmetry Δ. Country B however prefers the deployment treaty only when*

$0 \leq \Delta < \Delta_{\text{Max}}^{\text{FreeRider}}$*, which is therefore the region where the deployment treaty comes into effect. Both countries prefer the non-cooperative free-rider to the moratorium treaty.*

Proof. The algebraic expression for $\Delta_{\text{Max}}^{\text{FreeRider}} \leq \bar{\Delta}$ and derivations are in Appendix A. □

That neither country finds the moratorium treaty attractive is intuitive as both countries engage in SG in the non-cooperative equilibrium, indicating that they find SG valuable even under these non-cooperative conditions; to completely abstain from SG in a moratorium treaty then must be unattractive. The reason why country A prefers the deployment treaty to the non-cooperative free-rider outcome is cost-sharing. The disadvantage from having to compromise with country B on SG deployment levels is, due to the relatively aligned preferences in low asymmetry settings, small compared to the gain from splitting deployment costs. Country B opposes the deployment treaty for asymmetry levels above $\Delta_{\text{Max}}^{\text{FreeRider}}$ since the final temperature outcome in the non-cooperative free-rider equilibrium is close to country B's optimal level T_B (matching this level exactly at $\Delta = \bar{\Delta}$) and country A shoulders the main part of deployment cost. In other words, country B is significantly free-riding on country A's SG deployment. We will see below that country B's opposition to the deployment treaty also extends into the free-driver and climate clash region.

We move on to the case of high asymmetry, where the non-cooperative outcome would be the free-driver equilibrium.

Proposition 3 (Cooperation incentives in the 'SG only' case. High asymmetry, $\Delta \geq \bar{\Delta}$)**.**

(i) Country A prefers the free-driver equilibrium to the moratorium treaty throughout and prefers the deployment treaty over the free-driver equilibrium if $\Delta < \Delta_{\text{Max}}^{\text{SG}}$.

(ii) Country B opts for the moratorium treaty when $\Delta > \Delta_{\text{Morat}}^{\text{SG}}$ *and prefers the deployment treaty over the free-driver equilibrium if* $\Delta > \Delta_{\text{Min}}^{\text{SG}}$*. It is* $\bar{\Delta} < \Delta_{\text{Min}}^{\text{SG}} < \Delta_{\text{Max}}^{\text{SG}}$.

Therefore, the deployment treaty is stable for $\Delta_{\text{Min}}^{\text{SG}} < \Delta < \Delta_{\text{Max}}^{\text{SG}}$*, whereas the moratorium is never stable.*

Proof. The algebraic expressions for all relevant levels of the asymmetry parameter Δ and other derivations are in Appendix A. □

Here we see for the first time the appeal of a world without any climate intervention. Country B is willing to enter the moratorium treaty if the disadvantage from being dominated by the free-driver is sufficiently high. However, it is intuitive that the moratorium is not appealing to country A, the free-driver. Therefore, there are no circumstances under which the moratorium treaty can be expected to materialize. But the deployment treaty has better chances to form. If the asymmetry exceeds $\Delta_{\text{Min}}^{\text{SG}}$, the free-driver outcome is too harmful for country B which is hence willing to enter the deployment

treaty.[10] Country A is also willing to enter the deployment treaty, yet under almost inverse conditions. Specifically, for relatively moderate asymmetry levels, $\Delta < \Delta^{SG}_{Max}$, the sharing of deployment costs is attractive enough to justify the compromise in temperature levels. For asymmetry levels higher than Δ^{SG}_{Max}, however, the gap between the temperature compromise implicit in the deployment treaty on the one hand and what country A would like to implement on the other hand is too wide. But $\Delta^{SG}_{Min} < \Delta^{SG}_{Max}$, and so there do exist constellations where countries, faced with a looming free-driver outcome, decide to cooperate.

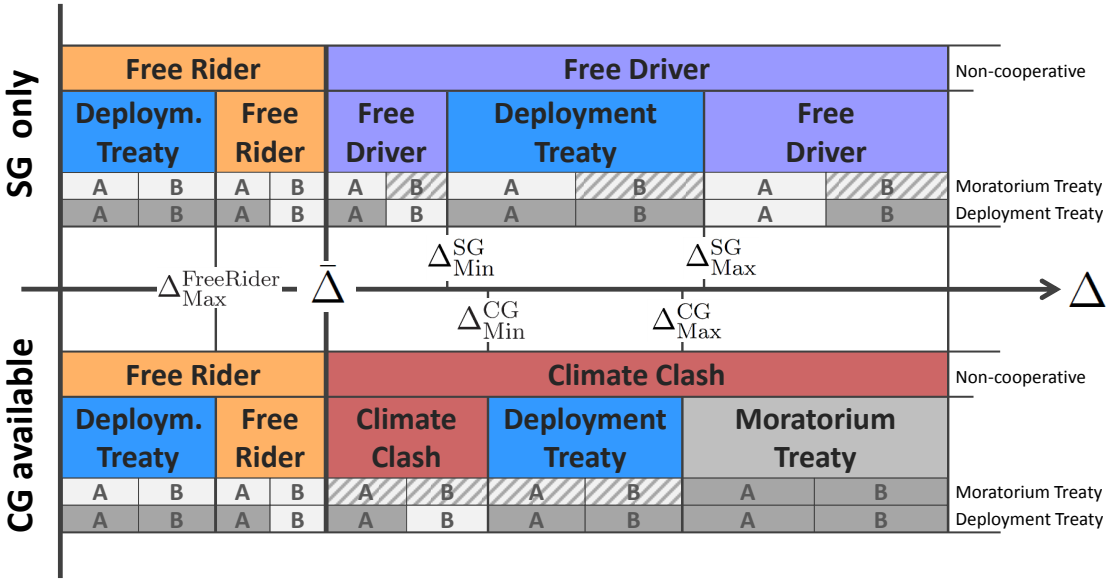


Figure 4: Schematic representation of all equilibria as a function of the asymmetry parameter Δ. The upper part shows equilibria in the 'SG only' case, the lower part in the 'CG available' case. The boxes below the equilibrium label indicate for both treaties whether countries A and B are willing to join (dark fill) or not (light fill), respectively. A hatched fill indicates that a country's decision whether to join or not is parameter-dependent but inconsequential for the final outcome. When both treaties are stable (i.e. both treaties are attractive for both countries) and countries disagree about which of the two they prefer, then our tie-breaking rule in Assumption 2 resolves the disagreement. Note that the relative size of the treaty equilibria with and without CG depends on parameter values. See section 5 for a calibration and sensitivity analysis.

4.2 Cooperation incentives with CG

If asymmetry is low, $\Delta \leq \bar{\Delta}$, cooperation incentives are not changed by the availability of CG as countries have no incentives to deploy CG anyway. We hence focus on the high-asymmetry case where non-cooperation would result in the climate clash.

Proposition 4 (Cooperation incentives in the 'CG available' case. High asymmetry, $\Delta \geq \bar{\Delta}$)**.**

[10]The intuition why country B still prefers the free-driver outcome over the deployment treaty for moderate asymmetry levels, $\Delta < \Delta^{SG}_{Min}$, is the same as in Proposition 2. The free-driver equilibrium involves no deployment costs for country B, and final temperature changes T, while excessive, are still relatively close to its optimal level T_B.

(i) Country A unambiguously prefers the deployment treaty over both the climate clash and the moratorium treaty, and prefers the moratorium over the climate clash iff $\Delta > \Delta^{\mathrm{CG,A}}_{\mathrm{Morat}}$.

(ii) Country B prefers the deployment treaty over the climate clash iff $\Delta > \Delta^{\mathrm{CG}}_{\mathrm{Min}} > \bar{\Delta}$*, prefers the moratorium over the climate clash iff* $\Delta > \Delta^{\mathrm{CG,B}}_{\mathrm{Morat}}$*, and prefers the moratorium over the deployment treaty iff* $\Delta > \Delta^{B}_{\mathrm{Morat,Treaty}}$*. It is* $\Delta^{\mathrm{CG,B}}_{\mathrm{Morat}} < \Delta^{\mathrm{CG,A}}_{\mathrm{Morat}}$*, while the size of* $\Delta^{B}_{\mathrm{Morat,Treaty}}$ *relative to other asymmetry levels depends on parameter settings.*

Therefore, the deployment treaty is stable for $\Delta > \Delta^{\mathrm{CG}}_{\mathrm{Min}}$ *and the moratorium treaty is stable for* $\Delta > \Delta^{\mathrm{CG,A}}_{\mathrm{Morat}}$*. Under the tie-breaking Assumption 2, the separating level between deployment treaty and moratorium treaty is* $\Delta^{\mathrm{CG}}_{\mathrm{Max}} := \max(\Delta^{\mathrm{CG,A}}_{\mathrm{Morat}}, \Delta^{B}_{\mathrm{Morat,Treaty}})$.

Proof. See appendix A. □

The moratorium treaty is stable, i.e. preferred by both countries over the climate clash, once the asymmetry exceeds $\Delta^{\mathrm{CG,A}}_{\mathrm{Morat}}$. The interest in the moratorium underlines how unattractive the climate clash is. Country B is more interested in the moratorium treaty than country A, which is expressed both by a wider opt-in region ($\Delta^{\mathrm{CG,B}}_{\mathrm{Morat}} < \Delta^{\mathrm{CG,A}}_{\mathrm{Morat}}$) and by a preference for the moratorium over the deployment treaty for levels beyond $\Delta^{B}_{\mathrm{Morat,Treaty}}$ (a preference that country A never has). This is intuitive when we recall that temperatures under climate change absent any climate intervention (the outcome under the moratorium treaty) are relatively less harmful for country B than for country A. There is a simple intuition why country A is keen to cooperate via the deployment treaty. Not only are deployment costs in the cooperative solution much lower than in the climate clash, the social optimal SG deployment level is also more ambitious and thus closer to T_A. That country B prefers the deployment treaty to the climate clash for moderate asymmetry levels Δ is similar to before: country B's deployment costs are low and the final temperature change is relatively close to B's optimum T_B.

To summarize, we find a rich set of potential outcomes that are depicted in Figure 4. Every outcome (the non-cooperatives as well as the two treaties) materializes under certain conditions, and the boundaries that separate different outcomes are non-trivial. A parameter calibration and sensitivity analysis of the equilibrium boundaries are presented in Section 5. Our findings suggest a substantial potential of CG to change the statics of the global thermostat game: The basic mechanism is to transform a free-driver equilibrium into a climate clash under non-cooperative conditions, and this transformation is always bad for the free-driver A (and often for country B as well). It is this mechanism that brings the free-driver to the negotiating table when cooperation is possible: the free-driver is now always willing to enter the global optimal deployment treaty. In order to prevent the wasteful climate clash, the free-driver is, under certain conditions, even willing to accept the otherwise very unattractive conditions of a moratorium

treaty. We will show in section 6 that this basic mechanism also shapes the general n country case.

4.3 Welfare ranking of outcomes

We have now gained a comprehensive understanding of CG's potential to change the global thermostat game. Are the changes induced by CG for the better or worse? We have partially answered this question above. Proposition 1 shows that the transformation from a free-driver outcome to a climate clash is detrimental as it decreases global welfare. On the other hand, whenever this bleak outlook induces countries to form a deployment treaty, which by definition implements the global best, then CG's game-changing effect is beneficial. What remains to be understood is how the moratorium treaty ranks in welfare terms. The following result shows that the transformation from a free-driver outcome to a moratorium treaty, cf. Proposition 4, is only beneficial for high levels of asymmetry. For completeness we also compare the moratorium treaty to the climate clash. While not important for the welfare impact induced by the presence of CG, this result sheds light on the value of having cooperation options once CG is part of the game.

Proposition 5 (Welfare of Moratorium Treaty)**.**

(i) Global welfare under the moratorium treaty is higher than in the free-driver equilibrium iff $\Delta > \Delta^{\text{Welfare}}_{\text{Morat,Driver}}$*, where* $\Delta^{\text{Welfare}}_{\text{Morat,Driver}} > \bar{\Delta}$

(ii) Global welfare under the moratorium treaty is higher than in the climate clash equilibrium iff $\Delta > \Delta^{\text{Welfare}}_{\text{Morat,Clash}}$*, where* $\Delta^{\text{Welfare}}_{\text{Morat,Clash}} > \bar{\Delta}$*. It is* $\Delta^{\text{Welfare}}_{\text{Morat,Driver}} > \Delta^{\text{Welfare}}_{\text{Morat,Clash}}$*.*

Proof. See appendix A. □

5 Calibration, sensitivity analysis, and welfare impact

In this section we first calibrate the model parameters b, c and $\bar{T}$. We then determine the sensitivity of equilibrium boundaries to changes in parameters. Finally, we discuss the welfare effect of CG in the calibrated model.

5.1 Parameter calibration

Our calibration of the benefit parameter b rests on Burke et al. (2015) who show that the relationship between (local) temperatures and growth rates follows a universal quadratic relationship. The calibration of the cost parameter c is based on data on stratospheric SG with sulfur aerosols. It combines data on operational cost per kg of load material with the non-linear relation between sulfur load and reduction in radiative forcing. Finally,

$\bar{T}$ expresses the amount of atmospheric cooling required to achieve the global optimal temperature at the point of climate intervention. This clearly depends on emissions scenarios. Appendix B provides details on the calibration that results in the following parameter values

$$b = 179.5 \text{ bn}\,\$/K^2 \quad , \quad c = 13.4 \text{ bn}\,\$/K^2 \quad , \quad \bar{T} = -2.1K. \tag{12}$$

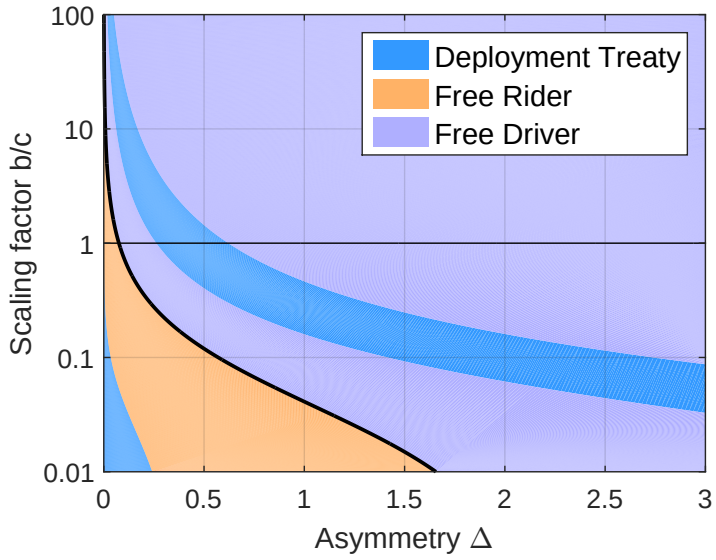
We keep asymmetry Δ as an open parameter for two reasons. First, this parameter is the hardest to calibrate as it depends on regional/country-specific preferences over climate outcomes (in contrast to $\bar{T}$ which is a measure of globally aggregated preferences). Second, this provides us with a degree of freedom to describe a variety of interactions between potentially very different agents.

5.2 Outcome boundaries and sensitivity

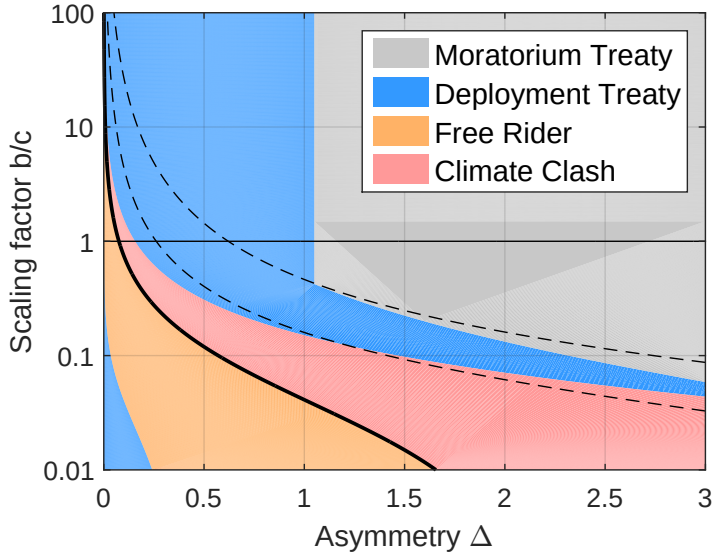
What outcome can we expect under parameters calibrated as above, and how sensitive are the equilibrium boundaries in Figure 4 to parameter settings? Analysis of the algebraic expression of the equilibrium boundaries (see appendix A) reveals that all boundaries scale linearly with $\bar{T}$ and depend only on the benefit-cost ratio b/c, not b and c separately. The horizontal line in Figure 5 shows the equilibrium boundaries for the best estimate $b/c = 179.5/13.4$, cf. (12). We check for sensitivity by scaling the benefit-cost ratio upwards and downwards by two orders of magnitude. Figure 5a and 5b depict the 'SG only' and 'CG available' cases, respectively. The solid black line represents $\bar{\Delta}$, the asymmetry threshold from (8) that separates free-rider outcomes to the left from free-driver ('SG only') and climate clash ('CG available') outcomes to the right.

The first observation is that the asymmetry threshold $\bar{\Delta}$ is very small for the calibrated parameter values.[11] This means that, in the absence of CG, even a small disagreement over the best use of SG will result in the free-driver outcome. Also note that the deployment treaty is only plausible under a fairly narrow asymmetry range. Overall, the free-driver equilibrium is the most likely outcome in the 'SG only' case. We see that, under the calibrated parameter values, CG strictly enlarges the conditions under which the deployment treaty materializes, whereas the climate clash is the predicted outcome only for a narrow range of asymmetry values. The moratorium treaty, according to our tie-breaking assumption 2, is the predicted outcome for the 'CG available' case under a wide range of asymmetry values.

[11] What are large and small values of the asymmetry parameter Δ? Consider the example sketched in footnote 7 where country A and country B have pre-industrial average temperatures of $16°C$ and $10°C$, respectively (this is not an extreme scenario as multiple regions experienced pre-industrial average temperatures beyond $20°C$). If both countries determined their preferences over climate interventions based solely on a universal optimal temperature, e.g. the $13°C$ in Burke et al. (2015), then $\Delta = 3K$. If, less extreme, both countries considered the midpoint between pre-industrial and a certain universal temperature as optimal, then $\Delta = 1.5K$. In this sense it is justified to say that the asymmetry threshold is typically very small.



(a) SG only.



(b) CG available.

Figure 5: Sensitivity analysis of the equilibrium boundaries in the $n = 2$ case. The reference benefit-cost ratio $b/c = 179.5/13.4$, cf. (12), is represented as the horizontal line. To check sensitivity we scale b/c upwards and downwards by two orders of magnitude. The solid black curve in both plots represents $\bar{\Delta}$. The dashed lines in (b) represent the deployment treaty boundaries of the 'SG only' case in (a).

In terms of sensitivity to parameter changes, we have noted above that all boundaries scale linearly with $\bar{T}$. Thus we can focus on the effect of changes in the benefit-cost ratio b/c. Figure 5 demonstrates that all our observations from above are only strengthened if b/c gets higher, for instance if operational costs of a climate intervention were significantly lower than current estimates. The free-driver is the typical outcome in the

'SG only' case, and cooperation (through either a deployment or moratorium treaty) in the 'CG available' case almost certain. The outcomes are very different if we consider lower benefit-cost ratios, for instance because climate damages are seen as relatively minor and/or climate interventions much more costly than currently expected. Then, the free-rider outcome could be plausible, in which case the presence of CG would be inconsequential. Interestingly, the plausibility of cooperation (via either deployment or moratorium treaty) decreases as b/c decreases, and the climate clash becomes increasingly plausible for high levels of asymmetry.

5.3 Calibrated welfare impacts

In this section we give a calibrated answer to the question whether the game-changing potential of CG is beneficial or detrimental. Figure 6 shows the effect of CG on global welfare.[12] As in Figure 5, the horizontal axis shows the asymmetry between country A and country B, whereas the vertical axis shows a range of benefit-cost ratios; the horizontal line represents the best estimate of b/c in (12).

There are two regions where CG does not change the game and hence leaves global welfare unchanged (indicated by white coloring). First, all asymmetry levels to the left of the asymmetry threshold $\bar{\Delta}$. Here, the non-cooperative outcome is the free-rider equilibrium, and neither country wants to deploy CG in the first place. The second region is where the deployment treaty was the outcome in the 'SG case' and remains the outcome when CG is available.

We find two reasons why CG can be beneficial, indicated by green colors in Figure 6. First, CG can transform a free-driver into a deployment treaty, and our findings suggest that this is likely for intermediate levels of asymmetry and relatively high benefit-cost ratios. The second situation in which CG increases overall welfare is when an extreme free-driver equilibrium is transformed into a moratorium treaty; in order for the technology-free world to be globally preferable, the asymmetry level must be high so that the free-driver outcome is very problematic.

If the asymmetry is not extreme, however, this transformation from free-driver to moratorium is detrimental (potentially for both countries), represented here in dark red colors. A second (as it turns out relatively rare) scenario in which CG is detrimental is when a deployment treaty (bordered by the dashed lines) in the 'SG only' case is transformed into either a climate clash or a moratorium treaty. Finally, there is a

[12] The plotted quantity in the contour plot is the welfare difference between the 'CG available' and the 'SG only' case. This difference is, for each separate parameter setting, expressed in terms of the absolute welfare under the social optimal outcome; if, for instance, global welfare under the deployment treaty is -10 units (recall that welfare levels are non-positive by assumption), then a value of -50% in the contour plot means that CG reduces global welfare by 5 units. Note that this plot necessitates reducing one degree of freedom. While the equilibrium boundaries in Figure 5 depend only on the benefit-cost ratio b/c, any welfare analysis depends on both parameters b and c separately. There are different ways to reduce one degree of freedom; here we stipulate that the welfare in the symmetric case $\Delta = 0$ under the social optimal deployment profile (g_A^{**}, g_B^{**}) is independent of the benefit-cost ratio b/c.

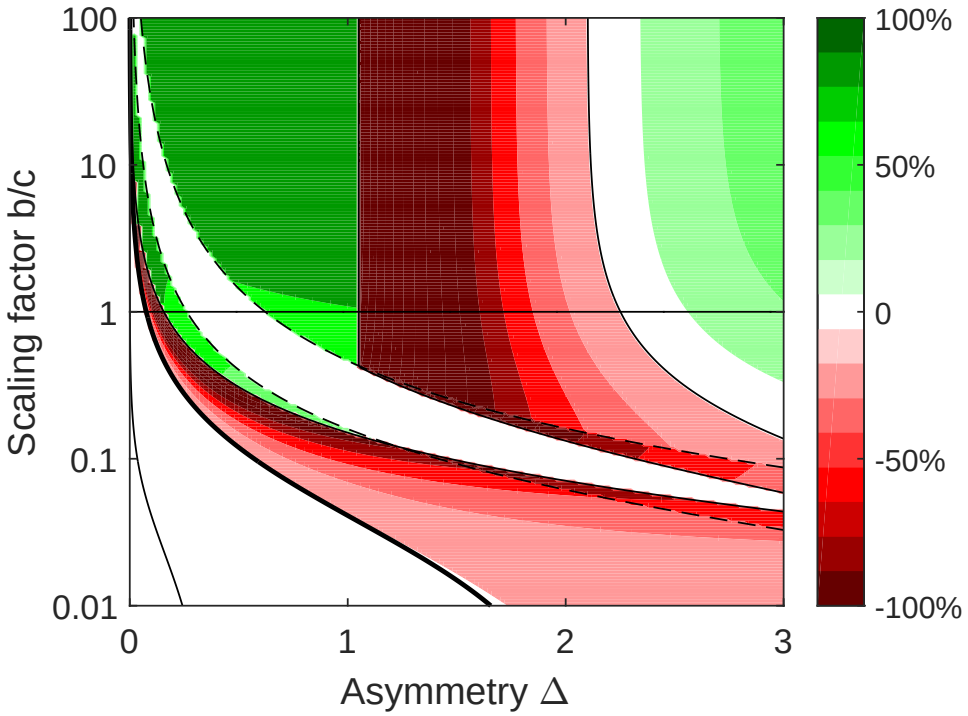


Figure 6: Welfare effect of CG. For every pair of Δ and b/c, the quantity plotted in the contour plot is the difference in welfare with and without CG, normalized by the social optimal welfare (i.e. welfare of the deployment treaty). Green and red colors indicate settings where the impact of CG is positive and negative, respectively, whereas white indicates that CG has no effect on global welfare.

third and important situation in which CG can reduce global welfare: If neither form of cooperation is attractive, then CG transforms what used to be a bad free-driver outcome into an even worse climate clash. This scenario is especially plausible for low benefit-cost ratios.

Appendix C demonstrates that the country-specific effects of CG are fairly clear: typically country A is worse off under CG, whereas country B benefits from the availability of CG. The mixed picture that we see in Figure 6 is hence the superposition of generally contrasting country-specific effects.

6 The n countries case

This section extends the setup to the general case of n countries. We derive analytical and numerical results to check the robustness of our results derived under the two-country model.

6.1 Model setup

The general structure in terms of benefit function, cost function and timeline of the model remains as before. There are now n countries, each with climate intervention level g_i, $i = 1, \ldots, n$. The *change in global average temperature* T due to climate interventions is $T = \sum_{i=1}^{n} g_i$. The *mean optimal temperature change* is $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$, where T_i is country i's preferred global average temperature change. We keep assumption 1, i.e. $\bar{T} < 0$. We write

$$T_i = \bar{T} + \Delta\delta_i \ , \tag{13}$$

where $\delta_i \leq \ldots \leq \delta_n$ with $\sum_{i=1}^{n} \delta_i = 0$. We normalize $\frac{1}{n}\sum_{i=1}^{n} \delta_i^2 = 1$, so that Δ is the standard deviation of the optimal temperature change T_i. We call Δ as above the *asymmetry parameter*. For $\Delta = 0$, all countries agree on how much the climate ought to change; increasing Δ represents growing disagreement across countries. Note that the definition for $n = 2$ in section 2 coincides with the definition given here: it is $\delta_A = -1$ and $\delta_B = 1$. We denote by $(g_i^{**})_i$ the socially optimal configuration that maximizes global welfare $\sum_{i=1}^{n} \pi_i(g)$. It is straightforward to show that

$$g_i^{**} = \frac{nb}{n^2 b + c}\bar{T} \qquad , \ i = 1, \ldots, n \ . \tag{14}$$

It is efficient for all countries to deploy the same amount of SG due to the homogeneous cost structure. Owing to $\bar{T} < 0$ (Assumption 1), the socially optimal deployment scheme features SG deployment by all countries. In particular, whether CG is available or not has no implications for the socially optimal deployment profile.

6.2 Non-cooperative equilibria

As before, the asymmetry Δ determines how many countries deploy SG in equilibrium, and the number of countries deploying SG is monotonically decreasing in Δ. The remaining countries, in any case, consider the overall temperature reduction by SG as too high and accordingly either do not deploy SG (in the 'SG only' case) or deploy CG (in the 'CG available' case). We find a set of $\Delta^{(m)}$ $(m = 0, \ldots, n)$ that is decreasing in m. Here, $\Delta^{(0)} = \infty$ and $\Delta^{(n)} = 0$. We define $\theta = b/c$, $\beta_m = \frac{m\theta}{m\theta+1}$ and the average optimal temperature change among the first m countries $\bar{T}^{(m)} = \frac{1}{m}\sum_{i=1}^{m} T_i$. With these preliminaries, we are ready for our next proposition.

Proposition 6 (Non-cooperative equilibria. General n)**.** *Let the asymmetry parameter be in the interval* $\Delta \in [\Delta^{(m)}, \Delta^{(m-1)}]$.

(i) The 'SG only' case has a unique equilibrium where the m countries with the highest preference for cooling deploy SG

$$g_i^{(m)} = \theta(T_i - \beta_m \bar{T}^{(m)}) \quad i = 1, \ldots m \tag{15}$$

and the remaining countries do not deploy, $g_i^{(m)} = 0$, $i = m+1, \ldots, n$.

(ii) When CG is available, all countries' deployment levels are given by

$$g_i^{(n)} = \theta(T_i - \beta_n \bar{T}) \quad i = 1, \ldots n \tag{16}$$

where the first m *are negative (SG deployment) and the remaining* $n - m$ *positive (CG deployment).*

(iii) The transformation induced by CG is typically detrimental, but there are exceptions to this rule.

Proof. See appendix A. □

In the case $n = 2$ we have, as required, $\Delta^{(1)} = \bar{\Delta}$ and the quantities given in Proposition 1 all coincide with (15), evaluated at $m = 2$ (free-rider and climate clash) or $m = 1$ (free-driver).

6.3 Cooperation: assumptions and results

The two forms of treaties, Moratorium Treaty and Deployment Treaty, are both modelled as *open-membership games*. Under the moratorium treaty *all* countries bind themselves to abstain from any technology deployment. For a deployment treaty, we stipulate that at most one coalition can form, and this coalition decides on the optimal deployment of SG, where the objective is maximization of the coalition's total payoff. In terms of timing we adopt the Stackelberg leadership assumption: After the coalition has made its decision, the other countries ('fringe') decide simultaneously and non-cooperatively on optimal SG (in the 'SG only' scenario) or between SG and CG deployment (in the 'CG available' scenario). Note that we allow for at most one coalition, and rule out CG as the coalition's action. The reason is simplicity and to allow a good comparison of the 'SG only' and 'CG available' case. One might defend this assumption by saying that further warming the climate through CG clearly has less international justification than SG, so international treaties on CG are less plausible. Nevertheless it would be worthwhile to explore alternative forms of cooperation in future research.

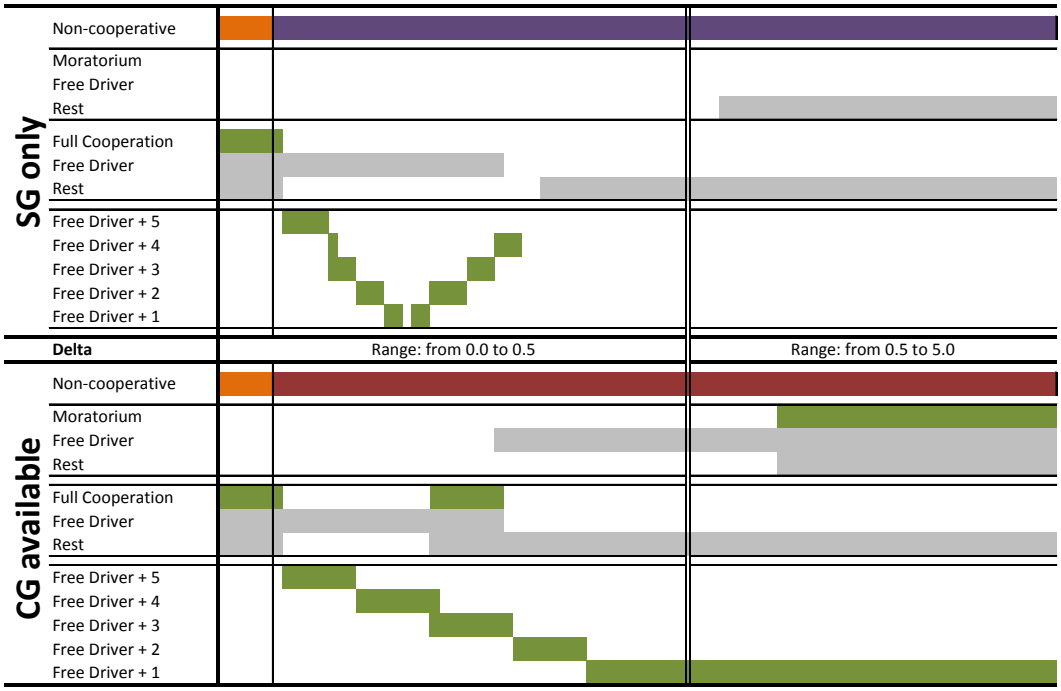
Stability of coalitions. Stability is defined relative to the non-cooperative outcome; hence, coalitions that are stable in the 'SG only' scenario need not be stable under the 'CG available' case, and vice versa. The moratorium treaty is stable if and only if *all* countries prefer the technology-free world over the non-cooperative technology deployment. A deployment treaty is stable if it is *internally* and *externally* stable. Internal stability means that every coalition member's payoff is higher or the same compared to a scenario in which he leaves the coalition. External stability of a coalition means that no fringe country can improve her outcome by joining the coalition. Both

stability concepts take the decisions of other countries as given. In other words we follow the usual simplifying approach without farsighted players (Mariotti and Xue 2003).

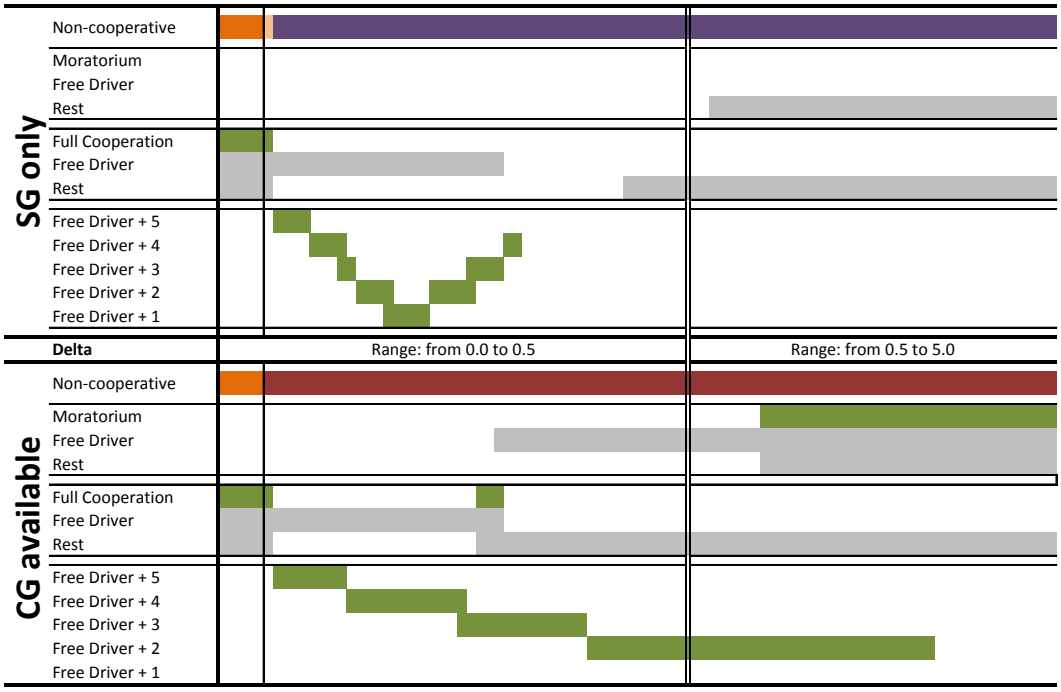
We find that there are stable coalitions that deploy no SG at all. The reason is our simplifying assumption that at most one coalition can form. Countries form a coalition in order to prevent a coalition that actually deploys SG. We regard this as an artifact of our model assumptions and accordingly disregard non-deploying stable coalitions.

Results. Figure 7 shows the results for a setting with $n = 7$ countries (we found similar results for other values of n). We focus on a setting with a clear 'free-driver' in the following sense: One country prefers temperatures lower than $\bar{T}$, all others are symmetric and prefer warmer temperatures. This is the upper part of Figure 7. In the lower part this setting has been slightly altered to test for robustness: the other countries' optimal temperatures have been randomly changed. In particular they are not perfectly symmetric anymore. In both figures, the leftmost vertical line is at $\Delta = 0$, the next separates two types of non-cooperative equilibria: low asymmetry levels where all countries deploy SG (orange), and levels where only the free-driver deploys SG (purple). The latter is transformed into a climate clash (red) when CG is available. We show two ranges (of different scales) of the asymmetry parameter Δ. The first range (left to the double vertical bar) is $[0, 0.5]$, the second (to the right of the double vertical bar) is the range $[0.5, 5]$. We show the stability of moratorium and deployment treaty outcomes in green, the latter differentiated into full cooperation deployment treaties (all countries are part of the treaty) and partial deployment treaties (only some countries participate). We find that stable coalitions take the form 'free-driver + k others'. For moratorium treaty and full cooperation deployment treaty outcomes, we show individual incentives for cooperation in grey, separated into incentives for the 'free-driver' and for the other countries bundled as 'rest' (we code 'rest' as incentivized to cooperate if *all* other countries are willing to cooperate); coalitions are stable where incentives to cooperate overlap.

Our results are in line with the $n = 2$ findings. The non-cooperative case is perfectly analogous. For low asymmetry Δ we see a free-rider equilibrium (where all countries deploy SG) that is unaffected by the presence of CG; high levels of asymmetry are characterized by a free-driver equilibrium (where only one country deploys), and this free-driver outcome is transformed into a climate clash once CG is available. The cooperation incentives are fairly similar to the $n = 2$ case. Starting with the moratorium treaty, in the 'SG only' case the free-driver is unambiguously opposed to it, while the other countries prefer the technology-free world if asymmetry (and accordingly the gap between other countries' optimal temperatures and what the free-driver implements) is high. Once CG is available, though, the moratorium treaty becomes attractive; in particular, the free-driver – even at intermediate levels of asymmetry – is willing to jointly abstain from deployment in order to prevent the costly climate clash.



(a) $(\delta_i)_{i=1,\dots,7} = [-2.450, 0.408, 0.408, 0.408, 0.408, 0.408, 0.408]$.



(b) $(\delta_i)_{i=1,\dots,7} = [-2.447, 0.3260.370, 0.408, 0.408, 0.449, 0.490]$.

Figure 7: Global thermostat game setting for $n = 7$. Upper part (a): the six countries that prefer warmer temperatures are symmetric. Lower part (b): optimal temperatures have been slightly changed to test for robustness. The δ-vector has zero mean and standard deviation of 1. The figure presents non-cooperative equilibria and stability of moratorium treaty, full deployment treaty and partial deployment treaties (here always of the form "free-driver + x other countries").

Moving on to the deployment treaties (including full and partial cooperation), in the 'SG only' case we find that the deployment treaty can only be stable for low and intermediate levels of asymmetry; for larger levels the free-driver is not willing to compromise. This is to some extent changed when CG is available. Full cooperation remains fairly unattractive for the free-driver (recall that all coalition members count equally when the coalition's deployment level is determined), but the zone where at least partial cooperation is stable is significantly extended under 'CG available'. Our robustness analysis (lower part of Figure 7) suggests however that incentives supporting partial cooperation via a deployment treaty are less robust than those favoring a moratorium treaty.

7 Conclusions

We have studied the strategic interaction over fast-acting climate interventions when countries disagree on how much to modify the climate. We have modelled this interaction as a public good game in which asymmetric countries anticipate the Nash equilibrium of the non-cooperative game and have the option to cooperatively decide on the level of climate intervention. Our main focus has been technological capabilities to quickly counter other countries' (excessive) cooling by means of counter-geoengineering (CG), and in particular the question *how* CG alters the statics of the game and under which circumstances the resulting change in outcomes can prove beneficial.

Our findings are summarized as follows. When climate intervention is restricted to cooling by means of solar geoengineering (SG), then the typical outcome is the 'free-driver' equilibrium. The free-driver, the country that suffers from climate change the most and hence wants to cool the most, may set global temperatures as it pleases; other countries may suffer damages from this excessive cooling but have no measure against it. Cooperation incentives in this case are relatively weak, first and foremost because the free-driver has little reason to compromise. The availability of CG changes this game significantly. We demonstrate that the free-driver outcome becomes unstable once dominated countries have CG at their disposal, yet the resulting Nash equilibrium is an even more harmful 'climate clash' in which countries waste significant resources in an escalation of SG and CG deployment. This destructive prospect is the very reason why – under certain circumstances – the existence of CG can significantly increase countries' willingness to cooperate. Specifically, the would-be free-driver understands that a climate clash would harm him substantially, and is hence (under a broad set of circumstances) willing to make climate intervention decisions cooperatively. This can enhance collective welfare. Crucially, however, other countries might prefer cooperation in the form of a moratorium that reduces global welfare, or even a climate clash over cooperation altogether.

From a policy perspective the central question is what difference does the existence of a CG capability make to a world where SG is contemplated. Our analysis identifies three

key factors that crucially determine the answer. First, the ratio of benefits and costs of climate intervention matters. CG tends to increase cooperation incentives for high benefit-cost ratios but may give rise to a climate clash for low benefit-cost ratios. Second, multiple cooperative agreements can be stable and it matters which of them materializes. Even in the simple, stylized $n = 2$ case, both the moratorium and deployment treaty can be stable, and which one obtains determines how CG affects aggregate welfare. Finally, a key factor for understanding CG's influence is the level of asymmetry among countries. Where asymmetry is low, the strategic interaction is essentially a free-rider equilibrium and CG makes no difference. Where it is intermediate, a climate clash may ensue. For high levels of asymmetry countries are more willing to cooperate, but our result suggests that extreme levels of asymmetry may favour a welfare-imperfect moratorium.

Given the novelty both of this topic and of our analytical approach to it, we opted to keep the modeling framework as simple as possible, and thus we see various opportunities for extending it. The first possible extension is to allow countries' preferences to depend on climate indicators beyond temperature. The most obvious candidate is precipitation, in particular as a major concern surrounding SG is its potential to alter precipitation patterns. An interesting question in this context is whether the inclusion of indicators other than temperature exacerbates the asymmetry between countries or, instead, mitigates the free-driver concern. This points toward linkage with the emerging literature on 'optimal climate states'. A second possible extension is to include indirect effects of geoengineering that are not climate-related, for instance the health effects caused by the particles used for geoengineering (e.g. acid rain and ozone loss from stratospheric SG with sulfur particles). These effects can be captured with a second (negative) 'benefit' function; here, the effects of SG and CG may depend on the sum of *absolute* SG and CG levels and thus not cancel out as in the case of climate related effects. While a thorough analysis is left for future research (building on research into potential SG and CG particles and their possible secondary effects), we can speculate on how this would change our results. It seems plausible that these additional external effects would have little effect on individual choices, but render the climate clash significantly less attractive from a global welfare perspective. In that sense our findings of a problematic climate clash can be interpreted as a lower bound.

Three more potential extensions revolve specifically around cooperation. First, several valuable robustness checks on our modelling assumptions in the general n country case could be performed: modeling coalition and fringe decisions as simultaneous, in contrast to the Stackelberg leader assumption we have adopted; modeling a coalition as an exclusive club, in contrast to our assumption of an open membership game; and allowing, in addition to the SG coalition, a second CG coalition that deploys CG. Second, a richer set of cooperation possibilities could be explored, for example, the potential of transfers to enhance cooperation, or other forms of cooperation treaties that go beyond moratorium treaty and deployment treaty. Lastly, the subject of equilibrium selection

is ripe for further research. Our results have demonstrated that multiple stable cooperation equilibria are possible, and which one obtains determines the ultimate desirability of climate interventions such as CG. Our assessment of CG would be much more positive if we were sure that, where deployment and moratorium treaties are both stable, the former materializes. In this sense it is of central interest to understand which of multiple stable treaties is more likely to emerge.

References

Barrett, S. (1994). Self-enforcing international environmental agreements, *Oxford Economic Papers* **46**: 878–894.

Barrett, S. (2001). International cooperation for sale, *European Economic Review* **45**(10): 1835–1850.

Barrett, S., Lenton, T. M., Millner, A., Tavoni, A., Carpenter, S., Anderies, J. M., Chapin III, F. S., Crpin, A.-S., Daily, G., Ehrlich, P., Folke, C., Galaz, V., Hughes, T., Kautsky, N., Lambin, E. F., Naylor, R., Nyborg, K., Polasky, S., Scheffer, M., Wilen, J., Xepapadeas, A. and de Zeeuw, A. (2014). Climate engineering reconsidered, *Nature Climate Change* **4**(7): 527–529.

Burke, M., Hsiang, S. M. and Miguel, E. (2015). Global non-linear effect of temperature on economic production, *Nature* **527**(7577): 235–239.

Coase, R. H. (1960). The Problem of Social Cost, *The Journal of Law and Economics* **3**: 1–44.

Diamantoudi, E. and Sartzetakis, E. S. (2006). Stable international environmental agreements: An analytical approach, *Journal of Public Economic Theory* **8**(2): 247–263.

Diederich, J. and Goeschl, T. (2017). Does Mitigation Begin At Home?, *Technical report*, Discussion Paper Series, University of Heidelberg, Department of Economics.

Emmerling, J. and Tavoni, M. (2017). Quantifying Non-Cooperative Climate Engineering, *Fondazione Eni Enrico Mattei Working Paper*.

Emmerling, J., Manoussi, V. and Xepapadeas, A. (2016). Climate Engineering under Deep Uncertainty and Heterogeneity.

Finus, M. (2008). Game Theoretic Research on the Design of International Environmental Agreements: Insights, Critical Remarks, and Future Challenges, *International Review of Environmental and Resource Economics* **2**(1): 29–67.

Finus, M. and Rübbelke, D. T. G. (2013). Public good provision and ancillary benefits: The case of climate agreements, *Environmental and Resource Economics* **56**(2): 211–226.

Gampfer, R., Bernauer, T. and Kachi, A. (2014). Obtaining public support for North-South climate funding: Evidence from conjoint experiments in donor countries, *Global Environmental Change* **29**: 118–126.

Heyen, D. (2016). Strategic Conflicts on the Horizon: R&D Incentives for Environmental Technologies, *Climate Change Economics* **7**(4): 1650013.

Heyen, D., Wiertz, T. and Irvine, P. J. (2015). Regional disparities in SRM impacts: the challenge of diverging preferences, *Climatic Change* **133**(4): 557–563.

Horton, J. B. (2011). Geoengineering and the myth of unilateralism: pressures and prospects for international cooperation, *Stanford Journal of Law, Science & Policy* **4**: 56–69.

Keith, D. W. and MacMartin, D. G. (2015). A temporary, moderate and responsive scenario for solar geoengineering, *Nature Climate Change*.

Klepper, G. and Rickels, W. (2014). Climate Engineering: Economic Considerations and Research Challenges, *Review of Environmental Economics and Policy*.

Manoussi, V. and Xepapadeas, A. (2015). Cooperation and Competition in Climate Change Policies: Mitigation and Climate Engineering when Countries are Asymmetric, *Environmental and Resource Economics* pp. 1–23.

Mariotti, M. and Xue, L. (2003). *Farsightedness in coalition formation*, Cheltenham: Edward Elgar.

McClellan, J., Keith, D. W. and Apt, J. (2012). Cost analysis of stratospheric albedo modification delivery systems, *Environmental Research Letters* **7**(3): 034019.

McGinty, M. (2007). International environmental agreements among asymmetric nations, *Oxford Economic Papers*.

Millard-Ball, A. (2012). The Tuvalu Syndrome, *Climatic Change* **110**(3-4): 1047–1066.

Moreno-Cruz, J. B. (2015). Mitigation and the geoengineering threat, *Resource and Energy Economics* **41**: 248–263.

Moreno-Cruz, J. B., Ricke, K. L. and Keith, D. W. (2012). A simple model to account for regional inequalities in the effectiveness of solar radiation management, *Climatic Change* **110**(3-4): 649–668.

National Research Council (2015). Climate Intervention: Reflecting Sunlight to Cool Earth, *Technical report*.

Parker, A. (2014). Governing solar geoengineering research as it leaves the laboratory, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **372**(2031): 20140173.

Parker, A., Horton, J. B. and Keith, D. W. (2018). Stopping Solar Geoengineering Through Technical Means: A Preliminary Assessment of CounterGeoengineering, *Earth's Future*.

Pasztor, J., Scharf, C. and Schmidt, K.-U. (2017). How to govern geoengineering?, *Science* **357**(6348): 231–231. 00000.

Pierce, J. R., Weisenstein, D. K., Heckendorn, P., Peter, T. and Keith, D. W. (2010). Efficient formation of stratospheric aerosol for climate engineering by emission of condensible vapor from aircraft, *Geophysical Research Letters*.

Ricke, K. L., Moreno-Cruz, J. B. and Caldeira, K. (2013). Strategic incentives for climate geoengineering coalitions to exclude broad participation, *Environmental Research Letters* **8**(1): 014021.

Shaviv, N. J. (2005). On climate response to changes in the cosmic ray flux and radiative budget, *Journal of Geophysical Research: Space Physics* **110**: A08105. 00095.

Urpelainen, J. (2012). Geoengineering and global warming: a strategic perspective, *International Environmental Agreements: Politics, Law and Economics* **12**(4): 375–389.

Victor, D. (2008). On the regulation of geoengineering, *Oxford Review of Economic Policy* **24**(2): 322–336.

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): 1049–1068.

Appendix A Proof of the Propositions

Proof of Proposition 1. To prove (i) and (ii), begin with the 'CG available' case. The best response functions lead to $g_A^* = \frac{b}{2b+c}\bar{T} - \frac{b}{c}\Delta$ and $g_B^* = \frac{b}{2b+c}\bar{T} + \frac{b}{c}\Delta$. It is straightforward to see that $g_B^* < 0$ iff $\Delta < \bar{\Delta}$. This is accordingly the free-rider region where it is inconsequential whether CG is available or not. For $\Delta \geq \bar{\Delta}$ we have a climate clash with $g_B^* \geq 0$ in the 'CG available' case. In the 'SG only' case it is necessarily $g_B^* = 0$ when $\Delta \geq \bar{\Delta}$. Country A's best response is the free-driver level $g_A^* = \frac{b}{b+c}T_A$. To show (iii) we look at the difference in payoffs

between the free-driver and climate clash. This difference is

$$\begin{cases} \frac{1}{2}\frac{b^2}{c(b+c)(b+c/2)}\left[(b^2+\frac{5}{2}bc+c^2)(\Delta-\bar{\Delta})^2+2c(b+c)(-\bar{T})(\Delta-\bar{\Delta})\right] & \text{country A} \\ \frac{1}{8}\frac{b^2(b^2-bc-c^2)(2b+c)^2}{c(b+c)^2(b+c/2)^2}(\Delta-\bar{\Delta})^2 & \text{country B} \\ \frac{b^2(c^2/2+bc+b^2)}{c(b+c)^2}(\Delta-\bar{\Delta})^2+\frac{b^2}{b+c/2}(-\bar{T})(\Delta-\bar{\Delta}) & \text{total welfare} \end{cases} \tag{17}$$

Because of $-\bar{T}>0$ and $\Delta\geq\bar{\Delta}$, these quantities are always positive for country A and total welfare, and positive for country B iff $b^2-bc-c^2>0$. The latter is equivalent with $b/c>\frac{1+\sqrt{5}}{2}$.

Proof of Proposition 2.

Moratorium Treaty. For country A,

$$\pi_A(0,0)-\pi_A\left(g_A^*,g_B^*\right)=\frac{2b^2}{c(2b+c)^2}\Big[-\bar{T}^2c(b+\tfrac{3}{4}c)\;+\;\bar{T}c(b+\tfrac{1}{2}c)\Delta\;+\;(b+\tfrac{1}{2}c)^2\Delta^2\Big]$$

This is negative at $\Delta=0$. The only positive root is at

$$\Delta_{\text{Morat}}^{\text{CG,A}}=-\frac{c+2\sqrt{bc+c^2}}{2b+c}\bar{T}\;, \tag{18}$$

which is larger than $\bar{\Delta}$. The label indicates that we make use of this quantity in the 'CG available' case. See proposition 4. For country B,

$$\pi_B(0,0)-\pi_B\left(g_A^*,g_B^*\right)=\frac{2b^2}{c(2b+c)^2}\Big[-\bar{T}^2c(b+\tfrac{3}{4}c)\;-\;\bar{T}c(b+\tfrac{1}{2}c)\Delta\;+\;(b+\tfrac{1}{2}c)^2\Delta^2\Big]$$

This is negative at $\Delta=0$. The only positive root is at

$$\Delta_{\text{Morat}}^{\text{CG,B}}=-\frac{-c+2\sqrt{bc+c^2}}{2b+c}\bar{T}\;, \tag{19}$$

which is again larger than $\bar{\Delta}$. So neither country prefers the moratorium treaty over the free-rider outcome.

Deployment Treaty. We begin with country A. It is

$$\pi_A(g_A^{**},g_B^{**})-\pi_A\left(g_A^*,g_B^*\right)=\frac{b^2}{2c(2b+c)^2(4b+c)}\Big[c^3\bar{T}^2-2c\bar{T}(8b^2+10bc+3c^2)\Delta+(16b^3+20b^2c+8bc^2+c^3)\Delta^2\Big]$$

which is positive at $\Delta=0$. Because the expression has no positive root we see that country A always prefers the treaty over the free-rider equilibrium. For country B, it is

$$\pi_B(g_A^{**},g_B^{**})-\pi_B\left(g_A^*,g_B^*\right)=\frac{b^2}{2c(2b+c)^2(4b+c)}\Big[c^3\bar{T}^2+2c\bar{T}(8b^2+10bc+3c^2)\Delta+(16b^3+20b^2c+8bc^2+c^3)\Delta^2\Big]$$

which is positive at $\Delta=0$. The unique root smaller than $\bar{\Delta}$ is

$$\Delta_{\text{Max}}^{\text{FreeRider}}:=\frac{c\left(-3c-4b+2\sqrt{4b^2+5bc+2c^2}\right)}{(4b+c)(2b+c)}\bar{T} \tag{20}$$

So country B prefers the deployment treaty over the free-rider iff $\Delta<\Delta_{\text{Max}}^{\text{FreeRider}}$.

Proof of Proposition 3. We begin with the comparison of moratorium treaty and free-driver. For country A,

$$\pi_A(0,0)-\pi_A\left(\tfrac{b}{b+c}T_A,0\right)=-\frac{b^2T_A^2}{2b+2c}\;<0\;,$$

so country A always prefers the free-driver. For country B,

$$\pi_B(0,0) - \pi_B\left(\tfrac{b}{b+c}T_A, 0\right) = \frac{3b^2\left(b+\frac{2}{3}c\right)}{2(b+c)^2}\left(\Delta - \bar{T}\right)\left(\Delta + \frac{b+2c}{3b+2c}\bar{T}\right) \ ,$$

which is positive for $\Delta > \Delta^{\text{SG}}_{\text{Morat}} := \frac{b+2c}{3b+2c}\bar{T}$.
We continue with the comparison of deployment treaty and free-driver. We begin with country A. It is

$$\pi_A(g_A^{**}, g_B^{**}) - \pi_A\left(\tfrac{b}{b+c}T_A, 0\right) = \frac{b^2}{2(b+c)(4b+c)}\Big[3c\bar{T}^2 \ - \ 6c\bar{T}\Delta \ - \ (4b+c)\Delta^2\Big]$$

which is positive at $\Delta = 0$. The unique positive root is

$$\Delta^{\text{SG}}_{\text{Max}} := -\frac{3c + 2\sqrt{3bc+3c^2}}{4b+c}\,\bar{T} \tag{21}$$

and $\Delta^{\text{SG}}_{\text{Max}} > \bar{\Delta}$. This means that country A prefers the deployment treaty to the free-driver outcome iff $\bar{\Delta} \leq \Delta < \Delta^{\text{SG}}_{\text{Max}}$. We continue with country B. It is

$$\pi_B(g_A^{**}, g_B^{**}) - \pi_B\left(\tfrac{b}{b+c}T_A, 0\right) = \frac{b^2}{2(b+c)^2(4b+c)}\Big[\bar{T}^2c(2c-b) + 2\bar{T}c(4c+7b) + (12b^2+11bc+2c^2)\Delta^2\Big] \ .$$

The root larger than $\bar{\Delta}$ is

$$\Delta^{\text{SG}}_{\text{Min}} := -\frac{7bc + 4c^2 + 2\sqrt{3b^3c + 9b^2c^2 + 9bc^3 + 3c^4}}{12b^2 + 11bc + 2c^2}\,\bar{T} \tag{22}$$

At $\Delta = \bar{\Delta}$, the above expression is

$$-\frac{2b^2(b+c)\bar{T}^2c}{(4b+c)(2b+c)^2} \ < 0$$

so that country B prefers the free-driver outcome to the deployment treaty iff $\Delta < \Delta^{\text{SG}}_{\text{Min}}$. It is $\Delta^{\text{SG}}_{\text{Max}} - \Delta^{\text{SG}}_{\text{Min}} = -2\frac{\sqrt{c(b+c)}}{12b^2+11bc+2c^2}\bar{T}\cdot\Big[\sqrt{3}(3b+2c) + \sqrt{c(b+c)} - \sqrt{3}(b+c)\Big]$, which is clearly positive. The relative size of $\Delta^{\text{SG}}_{\text{Morat}}$ on the one hand and $\Delta^{\text{SG}}_{\text{Min}}$ and $\Delta^{\text{SG}}_{\text{Max}}$ on the other hand is dependent on b and c.

Proof of Proposition 4. The algebraic expressions for climate clash and free-rider equilibrium are the same; because of that some relevant quantities have already been defined in Proposition 2.

(i) That country A prefers the moratorium treaty over the climate clash iff $\Delta > \Delta^{\text{CG,A}}_{\text{Morat}}$ has been demonstrated in the proof of Proposition 2. To see that country A always prefers the deployment treaty over the moratorium, note that

$$\pi_A(g_A^{**}, g_B^{**}) - \pi_A(0,0) = -\frac{2b^2}{4b+c}\bar{T}(2\Delta - \bar{T})$$

which is positive due to $-\bar{T} > 0$.

(ii) That country B prefers the moratorium treaty over the climate clash iff $\Delta > \Delta^{\text{CG,B}}_{\text{Morat}}$ has been demonstrated in the proof of Proposition 2. It is immediately clear that $\Delta^{\text{CG,A}}_{\text{Morat}} > \Delta^{\text{CG,B}}_{\text{Morat}}$. Comparing deployment treaty and climate clash, country B prefers the former iff Δ is larger than

$$\Delta^{\text{CG}}_{\text{Min}} := -\frac{c\left(3c + 4b + 2\sqrt{4b^2+5bc+2c^2}\right)}{(4b+c)(2b+c)}\bar{T} \ . \tag{23}$$

In terms of deployment treaty vs. moratorium we have

$$\pi_B(g_A^{**}, g_B^{**}) - \pi_B(0,0) = \frac{2b^2}{4b+c}\bar{T}(\bar{T}+2\Delta) \ .$$

This means that country B prefers the moratorium treaty to the deployment treaty iff

$$\Delta > -\frac{1}{2}\bar{T} =: \Delta^B_{\text{Morat,Treaty}} \ . \tag{24}$$

(iii) For the moratorium treaty to be stable it is necessary that both countries prefer it over the climate clash; this is equivalent with $\Delta > \Delta^{\text{CG,A}}_{\text{Morat}}$. In addition, because of assumption 2, only one of the two countries needs to prefer the moratorium over the deployment treaty. From (i) we know that country A never prefers the moratorium, from (ii) we know that country B prefers the moratorium treaty over the deployment treaty iff $\Delta > \Delta^B_{\text{Morat,Treaty}}$. Under assumption 2 the moratorium treaty hence realizes for all asymmetry levels above $\Delta^{\text{CG}}_{\text{Max}} := \max(\Delta^{\text{CG,A}}_{\text{Morat}}, \Delta^B_{\text{Morat,Treaty}})$.

Proof of Proposition 5.

(i) It is

$$\pi(0,0) - \pi\left(\tfrac{b}{b+c}T_A, 0\right) = \frac{b^2}{2(b+c)^2}(\Delta - \bar{T})\left(\Delta(2b+c) + \bar{T}(2b+3c)\right)$$

which is negative at $\Delta = 0$. The unique positive root is

$$\Delta^{\text{Welfare}}_{\text{Morat,Driver}} := -\frac{2b+3c}{2b+c}\bar{T} \tag{25}$$

and it is straightforward to show that $\Delta^{\text{Welfare}}_{\text{Morat,Driver}}$ is larger than $\bar{\Delta}$.

(ii) We have

$$\pi(0,0) - \pi(g_A^*, g_B^*) = \frac{b^2}{c(2b+c)^2}\Big[-\bar{T}^2 c(4b+3c) \ + \ (4b^2+4bc+c^2)\Delta^2\Big] \ ,$$

which is negative at $\Delta = 0$. The unique positive root is at

$$\Delta^{\text{Welfare}}_{\text{Morat,Clash}} := -\frac{\sqrt{4bc+3c^2}}{2b+c}\bar{T}$$

and it is straightforward to show that $\Delta^{\text{Welfare}}_{\text{Morat,Clash}}$ is larger than $\bar{\Delta}$ and $\Delta^{\text{Welfare}}_{\text{Morat,Driver}} > \Delta^{\text{Welfare}}_{\text{Morat,Clash}}$.

Proof of Proposition 6. We prove part (i) and (ii) together. Consider the general n country case. Define $\theta = b/c$, $\beta_m = \frac{m\theta}{m\theta+1}$ and the average optimal temperature change among the first m countries $\bar{T}^{(m)} = \frac{1}{m}\sum_{i=1}^m T_i$. The best response of country i to the other countries' geoengineering deployment level $T_{-i} = \sum_{j\neq i}^n g_j$ is characterized by the first order condition $\frac{d\pi_i(g_i;T_{-i})}{dg_i} = 0$. In the 'SG only' world it is necessary to check whether the non-positive constraint binds. We calculate the best response function

$$g_i(T_{-i}) = \begin{cases} \min\left\{\frac{b}{b+c}\left(T_i - T_{-i}\right), 0\right\} & \text{SG only} \\ \frac{b}{b+c}\left(T_i - T_{-i}\right) & \text{CG available} \end{cases} \tag{26}$$

The game consisting only of the first m countries, i.e. the m countries with the highest preferences for cooling, has the equilibrium

$$g_i^{(m)} = \theta(T_i - \beta_m \bar{T}^{(m)}) \ . \tag{27}$$

The overall temperature change in this equilibrium is $\sum_{i=1}^{m} g_i^{(m)} = \beta_m \bar{T}^{(m)}$. This is the equilibrium of the 'SG only' case if and only if country $m+1$ considers the temperature reduction as too much (and hence is unwilling to deploy more SG) and country m is willing to contribute SG (i.e. the game of the first $m-1$ countries results in a total temperature reduction that does not exceed country m's optimal reduction so that country m, due to vanishing marginal costs at the point of non-contribution, is willing to deploy SG). This is the case iff

$$\beta_{m-1} \bar{T}^{(m-1)} > T_m \geq \beta_m \bar{T}^{(m)} \ , \tag{28}$$

which is equivalent to

$$\Delta \left(\beta_{m-1} \bar{\delta}^{(m-1)} - \delta_m \right) > (1 - \beta_{m-1}) \bar{T} \quad \text{and} \quad \Delta \left(\beta_m \bar{\delta}^{(m)} - \delta_{m+1} \right) \leq (1 - \beta_m) \bar{T} \tag{29}$$

Define $\Delta^{(m)} = \frac{1-\beta_m}{min(0, \beta_m \bar{\delta}^{(m)} - \delta_{m+1})} \bar{T} \in [0, \infty]$ for $m = 1, \ldots, n-1$ and set $\Delta^{(n)} = 0$ and $\Delta^{(0)} = \infty$. It is easy to see that $\Delta^{(m)}$ decreases in m. That (27) is the equilibrium of the SG only game then is equivalent with $\Delta^{(m)} \leq \Delta < \Delta^{(m-1)}$. The equilibrium when CG is available is always characterized by (27) with $m = n$ the first m contributions being negative and the remaining $n - m$ positive. In the case $n = 2$ we have, as required, $\Delta^{(1)} = \bar{T}$ and the quantities given in Proposition 1 all coincide with (27), evaluated at $m = 2$ (free-rider and climate clash) or $m = 1$ (free-driver).

We turn to part (iii). Assume that the 'SG only' case is such that exactly m countries deploy SG, $\Delta^{(m)} \leq \Delta < \Delta^{(m-1)}$. It is straightforward to see that the availability of CG decreases welfare relative to the 'SG only' case iff

$$E := (1+\theta) \sum_{k=1}^{n} (T_k - \beta_n \bar{T})^2 - (1+\theta) \sum_{k=1}^{m} (T_k - \beta_m \bar{T}^{(m)})^2 - \sum_{k=m+1}^{n} (T_k - \beta_m \bar{T}^{(m)})^2 \ > 0 \tag{30}$$

We use (13) to write expression E as a quadratic function in Δ, $E = C_0 + C_1 \Delta + C_2 \Delta^2$. We find

$$C_0 = (1+\theta) n \bar{T}^2 (1-\beta_n)^2 - (1+\theta) m \bar{T}^2 (1-\beta_m)^2 - (n-m) \bar{T}^2 (1-\beta_m)^2 \tag{31}$$

$$C_1 = 2 \bar{T} \bar{\delta}^{(m)} (1-\beta_m)^2 \left(\tfrac{\beta_m}{1-\beta_m} n - m\theta \right) \tag{32}$$

$$C_2 = \theta \sum_{k=m+1}^{n} \delta_k^2 + \beta_m (\bar{\delta}^{(m)})^2 \big(2m\theta - \beta_m (m\theta + n) \big) \tag{33}$$

This polynomial is quite intricate and it is cumbersome to analytically determine the parameter constellations for which $E > 0$. We instead used a matlab file to get a sense of the conditions. The first observation is that the extreme free-driver setting, $\delta_1 < 0$ and $\delta_k = -\delta_i/(n-1)$, seems to have the highest potential to result in $E < 0$, i.e. exceptions to the rule of welfare-decreasing CG. In this extreme free-driver setting, we find constellations with $E < 0$ for all $n \geq 5$ (whereas an equidistant δ-profile has $E < 0$ constellations only for $n \geq 9$). Constellations with $E < 0$ seem to be characterized by high levels of asymmetry Δ and low benefit-cost ratios θ. Future research is needed to analytically determine the conditions under which CG decreases/increases welfare in the non-cooperative case.

Appendix B Calibration

For costs and benefits, we focus on benefits and expenditures in a given year.

Benefit parameter b. Let g denote the growth rate. From Burke et al. (2015) (Extended Data Figure 1, i) we read

$$\frac{dg}{dT} = -\frac{1}{100}(T - 13) \ . \tag{34}$$

This is just read from the graph and should be checked again. The quadratic component of the growth rate g is thus $-\frac{1}{200}(T-13)^2$. For the benefits in a given year $B(T) = Y_0(1+g)$, where Y_0 is the GDP at the beginning of the period. Because $B(T) = -\frac{b}{2}(T - T_i)^2$ we see that $b = \frac{1}{100}Y_0$. For our simple analysis we assume two countries each the size of the US who base their geoengineering deployment decisions on the change in temperature benefits in a single year. For the US, GDP Y_0 in 2015 was 17.95 trillion \$, and hence

$$b = 179.5 \text{ bn}\,\$/K^2 \ . \tag{35}$$

Cost parameter c. The following table reflects the best available current cost estimates on stratospheric geoengineering with sulfur. The range of stratospheric sulfur load is taken from Pierce et al. (2010). The cost estimate is on the high end of the range in National Research Council (2015), referring to McClellan et al. (2012). The effect of stratospheric load on changes in radiative forcing is read from the SO2 scenario in Figure 4 in Pierce et al. (2010). The associated change in temperatures is based on the climate sensitivity lambda=$0.54\,Km^2/W$, which corresponds to an equilibrium temperature change of 2.1 K (Shaviv 2005). We can then

Variable						Source
Sulfur load (Mt)	0	2	5	10	20	McClellan et al. (2012)
Costs (bn \$)	0	3.2	8	16	32	National Research Council (2015)
Δ RF (Wm^{-2})	0	0.9	1.8	2.8	4.1	Pierce et al. (2010)
Δ T (K)	0	0.486	0.972	1.512	2.214	Shaviv (2005)

Table 1: Available data for cost estimates of stratospheric geoengineering with sulfur.

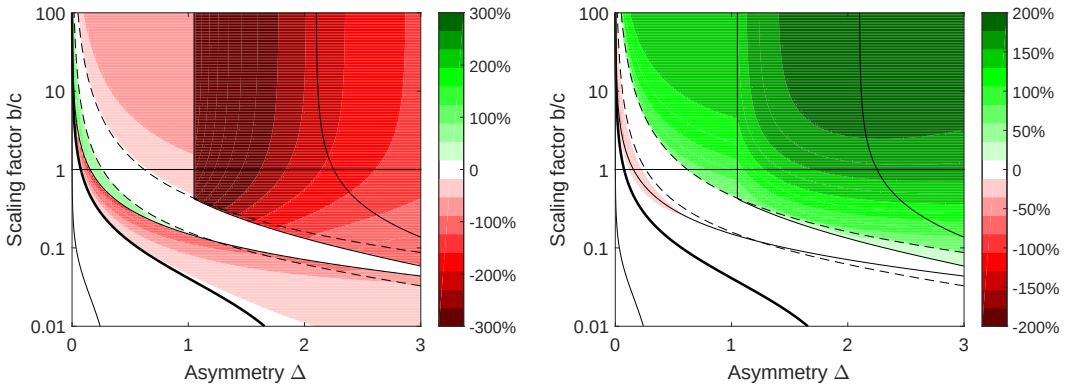
fit the model $C(T) = \frac{c}{2}(\Delta T)^2$ to the relationship between costs and temperature change. We calculate

$$c = 13.4 \text{ bn}\,\$/K^2 \ . \tag{36}$$

Temperature parameter $\bar{T}$. The parameter $\bar{T}$ expresses the amount by which average temperature exceeds the optimum at the beginning of the global thermostat game. We assume that preindustrial temperatures were on average optimal, and use for our numerical illustration Shaviv (2005) with an equilibrium temperature change of $2.1K$. This then corresponds to $\bar{T} = -2.1K$.

Appendix C Country-specific welfare change from CG

Figure 8 shows the country-specific welfare impact of CG for $n = 2$; this effectively disaggregates the aggregate effect shown in Figure 6. As before, red and green colors indicate a harmful and beneficial impact of CG, respectively. The plots suggest that country A is typically worse off under CG, while country B benefits from the availability of CG.



(a) Country A. (b) Country B.

Figure 8: The welfare impact of CG, differentiated into effects on country A and country B. As in Figure 6, the welfare differences between CG and SG are normalized by the *total* welfare under the deployment treaty. Note the country-specific scales which are different from the $[-100\%, 100\%]$ range in Figure 6.