Effective and Stable International Agreement on Solar Geoengineering

Abatement of greenhouse gases emissions is indispensable, but it is expensive and slow in reducing climate change risks. Therefore, solar geoengineering (SG), commonly referred to as “fast, cheap, and imperfect,” is increasingly discussed as a potential instrument of global climate policy portfolio. Because direct costs of SG deployment are low and local benefits are potentially large, the incentives for unilateral deployment are high. At the same time, SG deployment is associated with risks, and in the absence of international coordination, SG poses a real threat to global communities. We develop a formal game-theoretical framework to capture a range of current speculations about the potential design of international agreement on SG in a more formally precise way; and we propose a scheme of “adaptation transfers” and demonstrate that it can be used to reach an efficient agreement that is stable. These transfers would be directed from countries less vulnerable to climate change to those most vulnerable.
1. Introduction

Solar geoengineering (SG; also known as solar radiation management) is a proposed approach to reduce climate change damage costs by counteracting greenhouse-gas radiative forcing. Such negative radiative forcing can be produced by reflecting some fraction of a sunlight back to space. There are several ways to do so, ranging from painting roofs in white to installing mirrors in space. Yet, there is one technology that is the most promising - stratospheric aerosol geoengineering (National Academies of Sciences, Engineering, and Medicine, 2021). The idea behind it is to increase the amount of reflective aerosols in the stratosphere, thereby inducing global cooling (Crutzen, 2006). The naturally occurring analog is injection of sulfur into the stratosphere following volcanic eruptions that does, as observations indicate, cool the planet for a limited time. Stratospheric aerosol geoengineering (henceforth referred to as SG) is the focus of this paper.

In the context of ongoing climate crisis, and as a result of a delicate balance of costs and benefits, SG emerges as a potentially useful climate policy instrument. It acts fast and its direct deployment costs are estimated to be low. This potential, however, comes with a significant caveat: SG does not simply reverse GHG-driven climate change and its impacts, which are necessarily global, depend on the choice of magnitude and distribution of SG radiative forcing (RF). The problem of level of SG is also referred to as “global thermostat problem”. International coordination is what makes the difference between the success and failure in leveraging SG. Accordingly, Martin Weitzman describes SG as a public gob, which can be public good or bad - dependent on who and how uses it (Weitzman, 2015).

In the present study, we explore whether and how international agreement on SG that is effective and stable can be reached. To this end, Weitzman (2015) demonstrates that when all states have a capacity to act, the country that prefers the lowest temperature determines the temperature in Nash Equilibrium. This is ineffective outcome. If countries can counter-geoengineer, this would lead to yet another ineffective outcome – “climate clash” (Heyen et al., 2019). Effective outcome can be achieved via a voting rule, as proposed by Weitzman (2015), but in this case the question of stability remains open. Assuming countries outside of coalition cannot deploy SG and all countries within coalition gain from membership, there is an incentive to create smallest exclusive membership coalition with large enough power to deploy and sustain SG (Ricke et al., 2013). Similarly, Lloyd & Oppenheimer (2014) argue for a restricted membership, since
smaller coalitions have a better effectiveness. In both papers, there is no clear proposal on the deterrence of non-members from deployment. Parson (2014) suggests exclusive membership, where the ability to deploy SG is linked to emissions cuts. This may solve the mitigation deterrence problem, but requires the unseen level of commitment from countries.

In the present paper, we offer a systematic analysis of stability and effectiveness of international agreement on SG. For this, we develop a formal game-theoretical framework that captures a range of current speculations about the potential design of international agreement on SG. With this, modelling results and discussions claims from existing studies fall as alternative instances under alternative parameterizations/assumptions about the structure of the game.

We demonstrate that “adaptation transfers” may be used to facilitate stability of effective international agreement on SG. Overall, parametrization space for reaching an efficient and stable international agreement on SG is rather expansive. We also demonstrate that moratorium on SG deployment can be an optimal solution only if collateral damage shared by countries, is substantial.

2. The Game-theoretic Modelling Framework

We consider a two-stage coalition formation game. In the first stage, countries decide whether to join a SG deployment coalition. In the second stage, a formed coalition chooses the level of SG, or issues a moratorium on its deployment. Here, we contrast two decision-making rules: simple joint losses minimization and a supermajority voting rule proposed by Weitzman (2015). To start with, we specify what constitutes a stable and effective agreement on solar geoengineering.

Stability. The classical definition states that a coalition is internally stable with transfers if the joint payoff of its members in cooperation exceeds the sum of outside options in case of a unilateral deviation from cooperation. This is reasonable for the free-rider problem, when almost every country wins from a unilateral deviation, since it may still benefit from the efforts of other members that continue to cooperate while reducing its individual abatement costs. However, in the free-driver problem implies that not all coalition members have incentives to deviate: countries with low SG preferences do not benefit from a deviation, because they cannot set the desired level of SG, while high preference countries can inject extra aerosol to reach their optimal level. In this case, the classical concept of stability is less relevant because the focus should be on potential drivers that undermine cooperation. Here, we propose a new concept of stability, which differs
from the classical concept used in the literature on international environmental agreements (for example, Fuentes-Albero & Rubio (2010), McGinty (2006), Diamantoudi & Sartzetakis (2006) [others tba]).

Effectiveness. An agreement on SG is referred to as effective if it can reach a socially optimal SG deployment level.

Before we go into model specifications, here we state our default assumptions.

SG deployment capability. We assume that any country is capable of SG deployment, whether unilaterally or as part of a coalition. This is in contrast to previous game-theoretic studies, where countries outside of coalition assumed to cannot/will not deploy SG (e.g. in Ricke et al. (2013)). Instead, we explore the economic instruments that influence the choices that countries make and which can be used to help foster stability of agreement.

Symmetric preferences over SG level. In our specification, deviations from one’s preferred SG level induce same disutility independent of whether the deviation is positive or negative. This is a simplifying assumption needed to allow for analytical derivations. For several reasons, losses from over-doing would exceed losses from under-doing. In the Results Section we discuss for the case when the preferences are asymmetric.

Negligible deployment costs. In line with existing estimates, we assume that direct SG deployment costs are negligible.

2.1 Solar Geoengineering

We consider set N of heterogeneous countries, \( i=1,2\ldots n \), that experience climate impacts to a different extent. The amount of SG deployed by an individual country \( i \) is denoted by \( q_i \geq 0 \). Countries choose the deployment level by minimizing their individual loss function \( L_i \). Following Heyen (2016), we postulate the quadratic loss function:

\[
L_i(Q_N) = b(0.5 Q_N^2 - k_i Q_N) \tag{1}
\]

where an individual loss depends on the total level of SG deployment \( Q_N = \sum_{i=1}^{n} q_i \).

This is because reflecting particles injected into the stratosphere are transported around the world. A decreasing part of the loss function reflects the compensation for temperature rise, where
countries benefit from extra SG deployment. An increasing part represents SG-related local damages. Parameter $b > 0$ defines the shape of the loss function and is set the same for all regions. All individual differences are captured in the parameter $k_i$, which identifies countries’ preferred levels of SG deployment. This level is determined by countries’ anticipated vulnerability to climate change and local risks associated with SG deployment. For example, countries with a high average local temperature that suffer considerably from climate change would prefer to actively deploy SG to prevent temperature rise, in other words, they would have high $k_i$ values. Countries with low $k_i$, on the other hand, consider climate change as not such a severe threat compared to the risks associated with SG deployment. Equilibrium total SG deployment level would lead to underdoing of SG for some countries and overdoing for others.

3. Results

The default specification can be solved analytically, therefore the results do not depend on a specific distribution of countries’ preferences. In the case of global cooperation, the equilibrium is a socially optimal level of SG deployment, which is determined by the average of optimal levels of all countries: $\sum_{i \in N} k_i \equiv \bar{k}_N$. In this case, total losses will always be negative: $-\frac{b \bar{k}_N^2}{2} < 0$, i.e. the society always benefits from SG deployment (note that in this specification we assume that the abatement level is fixed). However, all countries whose optimal SG deployment level is above the average have incentives to deviate from cooperation: they can unilaterally inject the aerosol into the stratosphere to raise the overall level to match their preferences. This problem is a free-driving problem, and we call the countries that have such incentives potential drivers.

In the non-cooperative equilibrium, the SG level is determined by the preferences of the country with the highest optimal SG level, denoted by $k_h$. This level for all other countries is too high, so such equilibrium potentially leads to considerable losses. The total losses in no cooperation are positive if $k_h > 2 \bar{k}_N/n$, in other words, when the highest optimal level of SG is more than twice as big as the average of preferred levels of all countries. In this case, total damages associated with SG deployment exceed the benefits of preventing temperature rise and climate change. We believe that society needs to develop economic and political instruments to avoid such situation before SG technology is ready to be applied.

Different policy instruments to sustain cooperation are discussed in the literature, for example, sanctions, trade bans, and even military intervention, but in this section, we examine the
positive incentives such as international transfers. We show that for the default specification of the loss function, it is always possible to implement a system of transfers towards potential drivers to sustain a socially optimal level of SG.

**International transfer system.**

For simplicity, we divide countries into “drivers” and “non-drivers”, and compare the benefits of drivers in case of their deviation with the corresponding losses of non-drivers. We call “drivers” countries whose optimal level of SG exceeds the level set by the coalition, while others are “non-drivers”. We show that for the default specification, the sum of the additional losses of non-drivers from the deviation of any driver exceeds its additional gain from deviation, which is summarized by Lemma 1.

**Lemma 1:** in the default specification, the amount of extra damage to non-drivers (set ND) from deviation of any driver \((j \in DR)\) exceeds its gain from the deviation:

\[
\forall j \in DR: \sum_{i \in ND} (L_i(N \setminus j) - L_i(N)) > L_j(N) - L_j(N \setminus j)
\]

Proof is presented in the appendix A1.

Thus, the gains of non-drivers can be reallocated towards the potential driver to compensate for its gain from the deviation. However, if there is more than one potential driver, the gain should be compensated for all. For the default specification, we show that the total gain of all non-drivers from cooperation (versus no cooperation) exceeds the sum of the gains of all drivers from their unilateral deviation from cooperation, which is represented by Lemma 2.

**Lemma 2:** in the default specification, the total gain of non-drivers from cooperation (compared to no cooperation) exceeds the sum of the gains of all drivers from unilateral deviation from the cooperation:
\[
\sum_{i \in ND} (L_i(\emptyset) - L_i(N)) > \sum_{j \in DR} (L_j(N) - L_i(N \setminus j))
\]

This result implies that the gains of non-drivers can be redistributed towards potential drivers such that global cooperation would be the optimal strategy for all countries. This optimistic result suggests that an effective and stable agreement on the SG is possible and can be sustained with a system of transfers between countries. Transfers should go from countries with low preferences for SG to countries with high preferences, in other words, from countries less exposed to the negative effects of climate change to the most vulnerable countries. The literature shows that the vulnerable countries are mostly poor countries located in low latitudes and lacking the resources for adaptation. Thus, the proposed transfers are directed from rich to poor countries, which also does not contradict the politically implementable system. In reality, such transfer scheme can be carried out through an international adaptation fund. High-income countries with relatively low climate risks would contribute to this fund, and it can be used by more vulnerable low-income countries.

Below we explore the implications of simplifying assumptions we’ve made for the model results.

**Fixed costs associated with SG.**

Here we assume that the use of SG technology is associated with some fixed costs (FC). For example, this could be the risk of a sudden stop of aerosol injection into the stratosphere at some point in the future, or overuse of this technology. Society takes these risks at the very beginning, when it enters this “game”. These ideas are widely discussed in the literature comparing international treaties on SG with treaties on nuclear weapons testing or the deployment of weapons in space (Bunn et al. (2016), Wilson (2021)).

To take into account the risk associated with the use of SG technology, we introduce the \( FC \) parameter, which is the same for all countries: \( FC(Q_N) = 0 \) if \( Q_N = 0 \), \( FC(Q_N) = FC \) if \( Q_N > 0 \). The loss function for this specification is presented by expression (2).

\[
L_i^{fc}(Q_N) = b(0.5 Q_N^2 - k_i Q_N) + FC(Q_N) \tag{2}
\]
In this case, a moratorium on SG may become an optimal outcome. This happens if the fixed damage exceeds the gain from SG deployment. In non-cooperative equilibrium, the SG level is determined by the preferences of the country with the highest optimal SG level; this can also be a moratorium if the fixed damage exceeds the threshold value $0.5bk_i^2$. Note that the threshold value for introducing a moratorium in no cooperation is higher than in the global cooperation. Thus, when the fixed damage is between these two threshold values, in no cooperation the SG level is $k_h$, the highest preference, while the socially optimal solution is a moratorium. In this case, the total losses of countries can be tremendous.

Asymmetric preferences.

Now we relax an assumption about symmetric losses and consider that the risks from overdoing of SG exceed the losses from underdoing. We follow the approach of Martin Weitzman, who compared the risks of SG overdoing with a type I error, and the loss from underdoing with a type II error. Weitzman (2015) proposed an asymmetric piecewise linear loss function, while here we analyze a convex loss function. To capture the additional losses from the overuse of SG technology, we introduce a quadratic term $0.5d(\max(0, Q_N - k_i))^2$, which is positive only if the equilibrium level of SG deployment exceeds the optimal for a given country. Parameter $d>0$ determines the degree of asymmetry; the new loss function is represented by expression (3).

$$L_i^{\text{asy}m}(Q_N) = b(0.5Q_N^2 - k_iQ_N) + 0.5d(\max(0, Q_N - k_i))^2$$

This specification is solved numerically using the GAMS software as it is impossible to get an analytical solution. Figure 1 illustrates the preferences; two graphs compare the loss functions for three countries in the base specification (left) and specification with asymmetric preferences (right). In this case, the increasing part of the loss function is steeper than the decreasing one, reflecting the more significant losses from excessive SG deployment than insufficient SG use.
Figure 1. Illustration of preferences in case of three countries. The left graph depicts the default specification, the right graph shows the asymmetric preferences specification.

Figure 1 shows the equilibrium level of SG deployment in cooperation and no cooperation, demonstrating that the socially optimal level of SG in case of asymmetric preferences is lower than in the default specification, while the non-cooperative solution is the same. This implies that the assumption about preferences' asymmetry leads to a greater need for cooperation. Moreover, the losses of non-drivers in the case of free-driving are more significant compared to the default specification, strengthening the result about the sufficiency of "adaptation transfers" to ensure the stability of the global agreement on the SG.

Non-negligible deployment costs.

While current estimates suggest that direct deployment costs are negligible relative to climate damages, here we relax this assumption to be in line with Heyen (2016) and Heyen, Horton, Moreno-Cruz (2019). We assume a convex cost function $C_i(q_i) = 0.5 c q_i^2$, where $q_i$ is the individual level of SG deployment and $c>0$ is the slope of the marginal cost function. The deployment costs are included in the loss function presented in equation (1).

When considering the non-negligible direct costs of SG deployment, both the free-rider and the free-driver incentives arise in the game. We can say that the free-rider incentive prevails if the socially optimal level of SG deployment is higher than non-cooperative equilibrium. This happens when the mean of the optimal values of potential drivers is low enough compared to the global mean of preferred SG levels, taking into account costs and benefits parameters and the number of potential drivers. This relationship is formulated by expression (6).
\[
\frac{k_N}{n} > \frac{c/n + bn}{cn/DR + bn} \quad k_{DR} \Rightarrow Q_N^{coop} > Q_N^{nocoop}
\]

Otherwise, the free-driving incentive prevails, and then the result of the sufficiency of "adaptation transfers" to stabilize global cooperation remains.

**Counter-Solar Geoengineering.**

The literature shows that the ability of countries to counteract SG deployment of others can improve SG coordination (Heyen 2019). Counter-SG (hereafter - CSG) can be of two types: "neutralizing" CSG may be the acceleration of coagulation and hence the atmospheric deposition of aerosol particles. Another type, "countervailing" CSG, involves reversing the effects of SG particles, for example, by releasing warming agents or specially engineered solid particles to counter the change in radiative forcing caused by SG. Both forms of CSG are possible, but neither currently exists. Following Heyen (2019), we introduce CSG as a negative contribution to the overall level of SG deployment. Analyzing CSG makes sense only if the direct costs of the SG deployment are not negligible. Thus, the loss function has the same form as in the previous specification, while the use of the SG is not limited to positive values: \( q_i \in R \).

Allowing for CSG reduces the equilibrium SG level in no cooperation (see table A3.2 in the appendix) since non-drivers can actively oppose the SG deployment by drivers and reduce its level. Moreover, the level of SG in cooperation exceeds the level of SG in no cooperation, which is a classical public good underprovision problem. Thus, CSG may confront the free-driving incentives, but such a balance is not effective since it leads to a waste of resources. Moreover, CSG can be hazardous to human life and health. Heyen (2019) calls this equilibrium a "climate clash." We believe that it is necessary to develop an effective and safe mechanism to stabilize the global agreement, for example, "adaptation transfers," which have shown their effectiveness in the specification with non-negligible deployment costs.

**4. Discussion and Conclusions**

The widening gap between desired climate and climate action makes understanding the international cooperation on solar geoengineering even more important. We contribute to the literature by offering a game-theoretic modelling framework that helps to structure a discussion about the effective and stable international agreements.
Our results are optimistic: we demonstrate that “adaptation transfers” can be used to reach an effective agreement that is stable. These transfers would be directed from countries less vulnerable to climate change to those most vulnerable. While used here as an instrument for stabilization of an agreement, such transfers already exist, within the so-called Green Climate Fund. It is worth noting, though, that adaptation capacity is limited and therefore mitigation is an ultimate solution. Unlike adaptation, mitigation acts to decrease the preferred amount of SG for each country. The stronger is mitigation, the lower are adaptation transfers, relative to its business-as-usual levels.

Another commonly discussed issue is to establish moratorium on SG. Our results indicate that moratorium is result of an effective and stable agreement only when fixed costs associated with SG are substantial. At the same time, if the fixed costs are high enough for the society as a whole to prefer a moratorium, but lower than the benefit from SG deployment for some countries, then the transfers may not be enough to maintain the moratorium regime. Currently, SG technology is not used, it means that the expected risks from the start of deployment are large enough to outweigh the potential benefits even for the most vulnerable to climate change countries. However, over time, potential negative consequences of climate change would increase, and ongoing SG research would reduce the uncertainty associated with this technology. Thus, at some point, the benefits from SG deployment could outweigh the associated risks for some countries. We should arrive at this point with an effective mechanism to control the deployment of solar geoengineering technology.

Our derivations are subject to simplifying model assumptions. As such, preferences over SG should be asymmetric since over-doing is worse than underdoing. In this case, our results are reinforced. We also look at the case with non-negligible deployment costs, even though it goes against current estimates. In line with Heyen et al. (2019), we find that here two incentives -- to free-drive and to free-ride -- co-exist. Our model is static and does not capture the dynamic structure of coalition formation (Heyen & Lehtomaa, 2021). With no dynamics, our model does not capture the crucial difference between mitigation and SG in reducing climate damages - the ability of SG to act relatively fast. Last but not least, current scientific evidence on regional SG impacts is fragmented, especially when it comes to alternative specifications of SG deployment strategy.
References


Appendix A

A1 Proof of Lemma 1

∀ \( j \in DR \): \( \sum_{i \in ND} (L_i(N \setminus j) - L_i(N)) > L_j(N) - L_j(N \setminus j) \)

\[
\frac{b}{2} \left( k_j - \frac{k_N}{n} \right) \sum_{i \in ND} \left( k_j + \frac{k_N}{n} - 2k_i \right) > \frac{b}{2} \left( k_j - \frac{k_N}{n} \right)^2; \\
(ND - 1)k_j + (ND + 1) \frac{k_N}{n} > 2k_{ND};
\]

\[
(ND - 1)k_j + (ND + 1) \frac{k_N}{n} > 2ND \frac{k_N}{n} > 2k_{ND} - \text{by definition.}
\]

A2 Proof of Lemma 2

\[
\sum_{i \in ND} (L_i(\emptyset) - L_i(N)) > \sum_{j \in DR} (L_j(N) - L_i(N \setminus j))
\]

\[
\frac{b}{2} \left( k_h - \frac{k_N}{n} \right) \sum_{i \in ND} \left( k_h + \frac{k_N}{n} - 2k_i \right) > \frac{b}{2} \sum_{j \in DR} \left( k_j - \frac{k_N}{n} \right)^2; \\
\frac{b}{2} \sum_{j \in DR} \left( k_j - \frac{k_N}{n} \right)^2;
\]

\[
\sum_{i \in ND} \left( k_h + \frac{k_N}{n} - 2k_i \right) > \sum_{j \in DR} \left( k_j - \frac{k_N}{n} \right); \\

NDk_h + ND \frac{k_N}{n} - 2k_{ND} > k_{DR} - DR \frac{k_N}{n};
\]

\[
NDk_h > k_{ND} - \text{by definition.}
\]
### A3 Analytical solution

Table A3.1 - Equilibrium level of SG deployment

<table>
<thead>
<tr>
<th>Specification</th>
<th>No cooperation</th>
<th>Global cooperation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>$k_h$</td>
<td>$\sum_{i \in N} k_i / n \equiv k_N / n$</td>
</tr>
<tr>
<td>Fixed costs associated with SG</td>
<td>$k_h$, if $0.5b k_h^2 &gt; FD$&lt;br&gt;$0$, if $0.5b k_h^2 &lt; FD$</td>
<td>$k_N / n$, if $0.5b (k_N / n)^2 &gt; FD$&lt;br&gt;$0$, if $0.5b (k_N / n)^2 &lt; FD$</td>
</tr>
<tr>
<td>Different size of countries</td>
<td>$k_h$</td>
<td>$\sum_{i \in N} l_i k_i$</td>
</tr>
<tr>
<td>Non-negligible deployment costs</td>
<td>$b k_{DR} / (c + b \cdot DR)$</td>
<td>$nb k_N / (c + bn^2)$</td>
</tr>
<tr>
<td>Counter-SG available</td>
<td>$b k_N / (c + bn)$</td>
<td>$nb k_N / (c + bn^2)$</td>
</tr>
</tbody>
</table>

Table A3.2 - Total losses in equilibrium

<table>
<thead>
<tr>
<th>Specification</th>
<th>No cooperation</th>
<th>Global cooperation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>$bn k_h \left( \frac{k_h}{2} - \frac{k_N}{n} \right)$</td>
<td>$\frac{b k_N^2}{2 n}$</td>
</tr>
<tr>
<td>Fixed costs associated with SG</td>
<td>$bn k_h \left( \frac{k_h}{2} - \frac{k_N}{n} \right) + n \cdot FD$, if $0.5b k_h^2 &gt; FD$&lt;br&gt;$0$, if $0.5b k_h^2 &lt; FD$</td>
<td>$n \cdot FD - \frac{b k_N^2}{2 n}$, if $0.5b (k_N / n)^2 &gt; FD$&lt;br&gt;$0$, if $0.5b (k_N / n)^2 &lt; FD$</td>
</tr>
<tr>
<td>Different size of countries</td>
<td>$bn k_h \left( \frac{k_h}{2} - \sum_{i \in N} l_i k_i \right)$</td>
<td>$-\frac{b}{2} \left( \sum_{i \in N} l_i k_i \right)^2$</td>
</tr>
</tbody>
</table>

Table A3.3 - Adaptation transfers system for stability of the global cooperation

<table>
<thead>
<tr>
<th>Specification</th>
<th>Necessary amount of transfers</th>
<th>Condition to sustain the global cooperation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>$\frac{b}{2} \sum_{j \in DR} \left( k_j - \frac{k_N}{n} \right)^2$</td>
<td>Always</td>
</tr>
<tr>
<td>Fixed costs associated with SG</td>
<td>$\frac{b}{2} \sum_{j \in DR} \left( k_j - \frac{k_N}{n} \right)^2$, if $0.5b (k_N / n)^2 &gt; FD$&lt;br&gt;$0$, if $0.5b k_h^2 &lt; FD$</td>
<td>$0.5b (k_N / n)^2 &gt; FD$ or&lt;br&gt;$0.5b k_h^2 &lt; FD$</td>
</tr>
<tr>
<td>Different size of countries</td>
<td>$\frac{b}{2} \sum_{j \in D_R} l_j \left( k_j - \sum_{i \in N} l_i k_i \right)^2$</td>
<td>$\sum_{i \in N_D} l_i \geq \sum_{j \in D_R} l_j$</td>
</tr>
</tbody>
</table>