International Climate Agreements under the Threat of Solar Geoengineering

David M. McEvoy, Matthew McGinty, Todd L. Cherry, Stephan Kroll

Abstract: The possibility of overshooting global emissions targets has triggered a debate about the role of solar geoengineering (SGE)—using technologies to reflect solar radiation away from Earth—in managing climate change. One major concern is that SGE technologies are relatively cheap and could potentially be deployed by a single country (the "free driver"). We develop a model to analyze how opportunities to deploy SGE impact global abatement and the effectiveness of international environmental agreements (IEAs). We show that noncooperative abatement may increase or decrease under the threat of SGE, depending on how damaging the free driver's level of deployment is to others. When free-driver externalities are significant, other countries have additional incentives to abate—called anti-driver incentives—to reduce the free driver's deployment. We also show that compared to a world without SGE opportunities, stable IEAs can be large (small) if anti-driver incentives are relatively strong (weak).

JEL Codes: C7, D7, F5, H4

Keywords: solar geoengineering, solar radiation management, international environmental agreements, self-enforcing agreements, global public goods

AS GLOBAL EMISSIONS OF GREENHOUSE GASES (GHGS) keep increasing, there is concern that international efforts focused on mitigation will not be sufficient to avoid excessive warming. The possibility of falling short of abatement targets has triggered a public debate about the role that climate interventions—like solar geoengineering—may play in managing global warming. Solar geoengineering (SGE)—also called solar radiation management—describes the process of cooling the planet by reflecting solar

David M. McEvoy (corresponding author) is in the Department of Economics, Appalachian State University, 3102 Peacock Hall, Boone, NC 28608 (mcevoydm@appstate.edu). Matthew McGinty is in the Department of Economics, University of Wisconsin–Milwaukee, Milwaukee,

Dataverse data: https://doi.org/10.7910/DVN/JFGJC4

Received November 29, 2022; Accepted October 4, 2023; Published online May 1, 2024.

Journal of the Association of Environmental and Resource Economists, volume 11, number 4, July 2024.

https://doi.org/10.1086/728140

radiation away from Earth. There are many issues with researching and deploying such technologies, and one of the major concerns is international governance. SGE technologies are relatively cheap and could potentially be deployed by a single country that effectively controls the global climate. Moreover, the availability of SGE opportunities may alter countries' incentives to cooperate on emissions reductions. In this study, we turn to game theory to help inform us about how SGE opportunities may impact emissions abatement efforts and the overall effectiveness of international environmental agreements (IEAs) on climate change.

The term SGE describes a portfolio of climate intervention technologies that intentionally reflect sunlight away from Earth (Keith et al. 2010). While SGE can take different forms (e.g., marine cloud brightening, space-based mirrors) most attention is focused on stratospheric aerosol scattering (Keith 2000; Crutzen 2006; NRC 2015; Smith and Wagner 2018; Wagner 2021), which is the process of injecting aerosols into the upper atmosphere. This form of SGE is estimated to be the least expensive and fastest method of reducing global mean temperatures (Keith et al. 2010). SGE has received growing attention in academic communities (Aldy et al. 2021), policy debates (Keith 2021; Biermann et al. 2022), and recent reports from the intergovernmental panel on climate change (IPCC 2022).

SGE technologies hold the potential to reduce some of the harmful effects of climate change that result from a build-up of greenhouse gas emissions in the atmosphere. The technologies also introduce difficult challenges. Stratospheric aerosol scattering is estimated to be inexpensive enough so that a single country could unilaterally deploy an SGE program (Barrett 2008). This means that a single actor has the potential to determine the global average temperature. Wagner and Weitzman (2012) introduced the term "free driver" to describe this possibility. The "free" part is an exaggeration of how cheap the technology is, and "driver" captures the idea that a single actor can drive the global temperature. Since countries will likely have different ideal temperatures, too much or too little SGE could be costly. For this reason SGE is often considered a "good or bad" (GoB), depending on the level of deployment (Weitzman 2015; Abatayo et al. 2020; Wagner 2021).

WI 53201. Todd L. Cherry is in the Department of Economics, University of Wyoming, Laramie, WY 82071. Stephan Kroll is in the Department of Agricultural and Resource Economics, Colorado State University, Fort Collins, CO 80523. We thank Resources for the Future and the LAD Climate Fund for providing financial support for this research. This research is based on work supported by the National Science Foundation under grant no. 2033855. We thank Joe Aldy, Billy Pizer, and Stephen Holland for their detailed feedback and suggestions, and we thank all of the participants who provided comments during the two RFF Solar Geoengineering Research Workshops in May and September 2022. Finally, we thank the editors and reviewers for their helpful comments and suggestions.

Another potential challenge with SGE is that it could impact incentives to mitigate GHGs (Reynolds 2019). On some level, emissions abatement and SGE are policy substitutes since both activities can reduce damages from climate change. However, the cooling effects of SGE are temporary and quickly dissipate if deployment stops, and they cannot completely offset damages from a buildup of GHGs. For example, SGE would not address (and could exacerbate) problems of ocean acidification associated with GHGs (Williamson and Turley 2012). Since SGE is a new and emerging technology, it may also impose costly and unintended side effects. Given the uncertainty about the effectiveness of these technologies, the most prominent concern is that the availability of SGE may reduce countries' incentives to mitigate GHGs, thereby causing an increase in the stock of GHGs. This concern is often referred to as "moral hazard" but is more accurately defined as "crowding out" incentives to mitigate emissions (Wagner 2021, 118).

Our overarching objective is to better understand how the availability of SGE technologies could impact incentives for countries to mitigate GHG emissions and the overall effectiveness of an international agreement on climate change. The model we develop begins with familiar functional forms from the game-theoretic literature on agreements that govern global greenhouse gas emissions (e.g., Barrett 1994; Finus and McGinty 2019), and we add the availability of SGE to the decision space. The characterization of SGE in the model follows the current understanding in the literature of how the technologies would impact individual and collective welfare. Most important, the deployment of SGE by any country will impact temperatures for all countries, and countries are assumed to be heterogeneous in their ideal level of SGE (e.g., Ricke et al. 2013; Weitzman 2015; Heyen et al. 2019).

Other studies have developed game-theoretic models to analyze strategic interactions among countries in the context of SGE. One branch of this literature explores decisions to deploy SGE without including mitigation decisions, and therefore these studies are unable to explore interactions between the two types of investments. Ricke et al. (2013) introduce a model in which SGE can be deployed to decrease damages from impending climate change. Countries have heterogeneous and exogenously determined preferred levels of SGE (modeled as a GoB), and they explore the effectiveness of coalitions that set SGE levels to maximize joint payoffs of the members. The coalitions can intentionally exclude others from joining, and only members of a coalition can determine how SGE is deployed. They find that large coalitions can be sustained in this setting, but their approach avoids the free-driver problem and associated governance challenges by assuming that nonmembers cannot deploy SGE.

Weitzman (2015) also explores a model in which SGE is deployed to minimize climate damages, and countries have exogenously determined preferred levels. Weitzman shows that the country with the highest preferred level (or lowest preferred temperature) will act as the free driver to set the global temperature. His model considers a voting architecture that, under certain conditions, can lead to efficient deployment of SGE.

Abatayo et al. (2020) test the free-driver hypothesis using a simple model and set of experiments in which countries, as in Weitzman (2015), have exogenously determined preferred levels of SGE. Abatayo et al. find evidence in support of the free-driver hypothesis and find both inefficiencies in SGE and counter-SGE investments. Heyen et al. (2019) also model a world with heterogeneous preferences for SGE and the underlying free-driver problem. They show that when countries have the chance to invest in counter-SGE (to offset SGE deployment by those countries with higher preferences), groups of countries can be motivated to cooperate toward a more efficient solution. It is important to reiterate that all of these studies (Ricke et al. 2013; Weitzman 2015; Heyen et al. 2019; Abatayo et al. 2020) focus on SGE decisions without considering abatement opportunities.

Another branch of the literature directly explores the link between emissions abatement and SGE deployment. Moreno-Cruz (2015) propose a two-country model with both mitigation and SGE decisions, where the two investments are imperfect policy substitutes. They also introduce costly side effects from unintended consequences from SGE deployment. Both SGE and mitigation are modeled as global public goods without heterogeneity in preferred levels (i.e., SGE is not modeled as a GoB), but countries can differ by the potential damages they suffer. They show that the availability of SGE can cause aggregate abatement levels to decrease (crowding out) or increase (crowding in) depending on how similar the countries are in terms of the damages they suffer. Cherry et al. (2022) explore the moral hazard conjecture using a laboratory experiment with both mitigation and SGE decisions. On average, they find that the threat of SGE increases mitigation efforts (i.e., crowding in). Note that Moreno-Cruz (2015) and Cherry et al. (2022) do not explore agreements that govern mitigation or SGE decisions.

Millard-Ball (2012) is perhaps the study with overarching research questions closest to ours. The study develops a model of global mitigation under the threat of unilateral SGE deployment and uses it to explore the stability and effectiveness of IEAs that govern mitigation. Like Moreno-Cruz (2015), abatement and SGE are modeled as imperfect policy substitutes. The Millard-Ball (2012) approach to modeling SGE differs significantly from other studies, and the modeling choices have important implications when interpreting the results. First, homogeneous countries are modeled with identical preferences (i.e., SGE is not a GoB). Second, deploying SGE is a binary decision and provides net benefits to the deploying country and imposes external costs to all other countries. That is, Millard-Ball takes a unique approach and models SGE as a private good with negative externalities. Third, if more than one country decides to deploy SGE, then one randomly selected country is assumed to succeed in deployment. Since all countries want to deploy SGE and all countries are identical, each has a $1/N$ chance of being the lucky deployer. They show that if the external damages are high enough, the threat of SGE deployment can cause all countries to join a cooperative agreement on mitigation in equilibrium.

Our study is the first to explore international agreements on emissions when SGE is modeled as a good or bad (GoB) and governance is complicated by the threat of a free driver (following Weitzman 2015; Wagner 2021). Our approach in this study is to compare worlds with and without the availability of SGE technologies. We start with a standard model of global emissions abatement from the IEA literature (e.g., Barrett 1994, 2003; Finus and McGinty 2019) and build in features of SGE. We derive and compare noncooperative and socially optimal levels with and without opportunities for SGE. Then we explore the formation of international agreements under the threat of SGE deployment.

We model an IEA with three stages. In the first stage (participation stage), countries decide independently and simultaneously whether or not to join the agreement. In the second stage (abatement stage), the agreement members choose abatement levels to maximize collective payoffs while nonmembers choose abatement independently. In the third stage (solar geoengineering stage), both IEA members and nonmembers make SGE decisions simultaneously and independently.

Our study provides new and useful insights that contribute to the current policy debate regarding the use of SGE in the portfolio of options to address climate change (e.g., Aldy et al. 2021). One of the main contributions of the study is the formal analysis of the relationship between emissions abatement and countries' preferred levels of SGE. We show that noncooperative abatement levels may increase or decrease under the threat of SGE, and it depends on how damaging the free driver's level of SGE deployment is on the other countries. The size of the damages from SGE, in turn, depends on the distribution of benefits from SGE. When potential free-driver damages are high, other countries have an additional incentive to abate—which we call an "anti-driver incentive"—to reduce the free driver's deployment. Indeed, it is possible that anti-driver incentives are sufficiently strong to increase abatement enough to prevent geoengineering deployment without an agreement. Our results help us better understand the nuance of the "moral hazard" debate that is centered on the conjecture that SGE opportunities will undermine abatement (e.g., Reynolds 2019; Wagner 2021). It is possible that the threat of a menacing free driver can lead to more cooperation. This finding may speak to how policymakers allocate their investments/efforts to manage climate change.

Consistent with much of the established literature on IEAs (e.g., Barrett 1994, 1999; McGinty 2007; McEvoy and Stranlund 2009), we find that IEAs under the threat of SGE lead to only marginal improvements in efficiency. In the special case of homogeneous benefits, we show that SGE opportunities do not alter the well-known result in the IEA literature that the largest stable coalitions consist of three members (e.g., Barrett 1994, 2003). With heterogeneous countries, we find that stability is challenged by two competing incentives. The free-rider incentive is the additional payoff a defecting member achieves by leaving the agreement since they lower their abatement responsibilities. On the other hand, the anti-driver incentive is the additional payoff achieved by joining an agreement and further dampening the free-driver effect through increased collective abatement. Through examples, we show that the largest stable IEA with SGE opportunities depends on the magnitude of anti-driver incentives. When anti-driver incentives are relatively low (high), stable coalitions can be relatively small (large) in comparison to a standard IEA without SGE. Ultimately, in both extremes, we find that stable agreements are unable to dramatically improve efficiency relative to the noncooperative baseline.

In section 1, we model global emissions abatement and SGE deployment. In sections 2 and 3, we derive the noncooperative and socially optimal abatement and SGE levels. Section 4 introduces the three-stage IEA and derives the stability conditions that define equilibrium agreement sizes. Toward the end of section 4 we explore two numeric examples that intentionally vary anti-driver incentives, and we include a third example using global climate change vulnerability metrics. The examples allow us to examine abatement levels, payoffs, and the size of stable IEAs in order to provide further insights. The final section offers a discussion of our main findings, the implications for policy and SGE governance, and opportunities for future research.

1. EMISSIONS ABATEMENT AND SOLAR GEOENGINEERING

Our approach is to start with a standard model of global emissions abatement from the IEA literature (Barrett 1994, 2003; Finus and McGinty 2019) and build in features of SGE. As a baseline and starting point, we first consider a world without the availability of solar geoengineering technologies and characterize noncooperative and socially optimal abatement levels.

Following Barrett (1994), we adopt a symmetric model with linear benefits and quadratic abatement costs. Emissions abatement is a pure global public good, abatement by country i is denoted as q_i and aggregate abatement is denoted as Q where $Q = \Sigma_{i \in n} q_i$ and n is the total number of countries. Global benefit is $B(Q) = bQ$, and each country has the same benefit share $1/n$, so each country's benefit is $B_i(Q) = bQ/n$. Abatement costs, denoted as C, are convex and equal to $C_i(q_i) = c(q_i)^2/2$. We assume identical abatement benefits to clearly highlight the impact of introducing the SGE option and to isolate the role of heterogeneous SGE benefits. If both abatement and SGE were heterogeneous then there would be countervailing forces that would obfuscate the role of SGE. Our approach allows us to show how countries respond to SGE as an option when they have both identical and heterogeneous SGE benefits. Indeed, we will show that heterogeneous SGE benefits result in different noncooperative abatement levels, even with identical abatement shares. Interested readers will find analytical solutions for the asymmetric abatement version of this model without SGE in Finus and McGinty (2019).

A country's payoff is $B_i(Q) - C_i(q_i)$, or

$$\pi_i(q_i, Q) = \frac{bQ}{n} - \frac{c(q_i)^2}{2}. \tag{1}$$

Country i chooses q_i to maximize equation (1). The noncooperative equilibrium abatement levels are denoted q^* and Q^* which are

$$\begin{aligned} q^* &= \frac{b}{cn}, \\ Q^* &= \frac{b}{c}. \end{aligned} \tag{2}$$

The socially optimal abatement levels maximize aggregate payoff $\Pi = \Sigma_{i \in n} \pi_i$ and are denoted q_o and Q^o:[1]

$$\begin{aligned} q^o &= \frac{b}{c}, \\ Q^o &= \frac{bn}{c}. \end{aligned} \tag{3}$$

Now we consider a world in which countries have an additional channel to manage climate change through the deployment of SGE. Deployment of SGE by any country can potentially provide benefits by reducing global temperatures. SGE, however, is an imperfect substitute for emissions abatement in the sense that it does not address the root cause of the problem and cannot entirely offset all of the damages caused by GHG emissions (e.g., does not address ocean acidification problems) (Robock 2008).

Countries are heterogeneous in their payoff-maximizing level of SGE deployment. We model the marginal benefit of abatement as homogeneous while introducing heterogeneity in benefits for SGE in order to isolate the individual effect of introducing an SGE option. However, there are multiple reasons why SGE impacts may be heterogeneous and decoupled from abatement. Benefits may be influenced by heterogeneity in the expected unintended consequences or side effects from SGE (e.g., ozone depletion) or by heterogeneity in how climate change impacts are reduced (or exacerbated) depending on geographic location. Different ethical considerations and judgments about the "right" way to management climate change could also result in heterogeneity in preferred levels. Finally, the heterogeneity in SGE benefits could capture different political implications from acts of deployment.

We assume that a country's payoff-maximizing level of SGE is a decreasing function of global emissions abatement. Intuitively, as the potential damages from climate change are reduced through emissions abatement, the less a country needs to rely on the new geoengineering technologies. We let the parameter γ_i capture the heterogeneity in

[1] The aggregate abatement and payoff differences between the noncooperative outcome and the social optimum are $Q^o - Q^* = [b(n-1)]/c] > 0$ and $\Pi^o - \Pi^* = [b^2(n-1)^2]/2cn > 0$.

benefits for SGE. Country *i*'s payoff-maximizing—or "preferred"—level of SGE is denoted as G_i^p and takes the following form:

$$G_i^p = \gamma_i(Q^o - Q), \tag{4}$$

where the term in parentheses is the abatement gap—the difference between the socially optimal abatement level and the actual abatement level. Note that if aggregate abatement $Q = Q^o$ then the abatement gap in (4) equals zero and no country benefits from positive levels of SGE. We restrict G_i^p to be nonnegative, and γ_i as a proportion bound between zero and one. If $Q = 0$, then $G_i^p = \gamma_i Q^o, \ \forall \ i \in N$, which captures that SGE is an imperfect substitute for abatement. The distribution of G_i^p in our model can be likened to the distribution of preferred levels G* in Abatayo et al. (2020) with the main difference that the preferred levels in our model are endogenous and a decreasing function of aggregate abatement.[2]

Following Weitzman (2015) and Abatayo et al. (2020), increases in SGE benefit a country up to G_i^p, but SGE levels beyond an individual country's preferred point are costly. In this way, SGE can be both a "good" or a "bad" depending on the realized level, referred to as a "GoB" in the literature (Weitzman 2015). We denote the realized level of aggregate SGE for all *n* countries as G. In our model, G is determined as a best-shot technology, which means that the aggregate level of solar geoengineering is determined by the country that chooses the highest level of deployment. Modeling SGE as a best shot follows the description of SGE technologies in Barrett (2007) and the experiments in Cherry et al. (2022).[3] A country's payoff function with emissions abatement and geoengineering opportunities is

$$\pi_i(q_i, Q, G) = \frac{bQ}{n} - \frac{c(q_i)^2}{2} + \beta_i(G), \tag{5}$$

where

$$\beta_i(G) = \begin{cases} \theta G & \text{if } G \leq G_i^p, \\ \theta G_i^p - \phi(G - G_i^p) & \text{if } G > G_i^p. \end{cases}$$

[2] Abatayo et al. (2020) do not directly consider the link between SGE benefits and emissions abatement. Likewise, Ricke et al. (2013) and Weitzman (2015) start with exogenously determined preferences for solar geoengineering without introducing mitigation.

[3] Others have modeled G using a summation technology (e.g., Moreno-Cruz 2015; Abatayo et al. 2020). Ultimately, since the country with the highest payoff-maximizing level is the only country deploying geoengineering, the choice of aggregation is inconsequential to our results.

In equation (5), $\beta_i(G)$ is country i's piecewise benefit from SGE, θ is the marginal benefit to SGE when G is less than country i's preferred level, and ϕ is the marginal loss in benefits (i.e., marginal cost) for additional deployment of SGE beyond the preferred level. Global benefits are denoted as $\beta(G) = \Sigma_{i\in N}\beta_i(G)$. For simplicity, like Barrett (2008) and Weitzman (2015) we assume that the cost of deploying SGE is extremely small,[4] and we ignore it in our model.[5]

We require a parameter restriction regarding the relative impact of marginal abatement and SGE decisions on payoffs that is consistent with SGE being an imperfect substitute for abatement. For all n countries, the maximum individual marginal benefit from SGE is no greater than the individual marginal benefit from emissions abatement. Specifically, this is

$$\frac{b}{n} \geq \theta. \tag{6}$$

Note that each country's benefit function is maximized at G_i^p. Figure 1 shows example SGE benefit functions for three countries with low (G_l^p), medium (G_m^p), and the highest (G_n^p) benefits from SGE. For levels below G_l^p all n countries increase their benefits by θ for each marginal increase in G. However, increases beyond each country's preferred level result in a marginal reduction in benefits of ϕ for those countries. This can be seen in figure 1 where the benefit function starts to decrease after peaking at the preferred level.

The benefit function from geoengineering, $\beta_i(G)$, can turn negative for sufficiently high levels of deployment beyond a country's preferred level. This is observed in figure 1 where the benefit functions for the two countries with low and medium benefits for SGE cross the horizontal axis. The condition for $\beta_i(G) < 0$ is the solution to $\beta_i(G) = \theta G_i^p - \phi(G - G_i^p) < 0$, which when substituting the expression for G_i^p from equation (4) reduces to $G > [(\phi + \theta)/\phi]\gamma_i(Q^o - Q)$. The right-hand-side of the expression is decreasing in Q and is equal to zero for all countries when $Q = Q^o$. Thus, at $Q = Q^o$, any deployment of SGE lowers all countries' payoffs and is strictly a bad.

[4] We did explore the implications of adding a fixed cost of SGE deployment to the model. Ultimately, we discovered that introducing fixed costs results in one of two outcomes. If the fixed costs are larger than the SGE benefits for the country with the highest preferred level, then $G = 0$ and we revert to a world without SGE opportunities. Alternatively, if the fixed costs are less than the SGE benefits for the country with the highest preferred level, there is no impact on SGE deployment, abatement levels, and stable coalitions.

[5] We model SGE investments as having immediate and permanent effects. In reality, SGE deployment is not permanent—the effects will dissipate over time. However, one can think of the decision to deploy in our model as a country's commitment to continuously deploy in order to maintain its preferred temperature. Indeed, the models that estimate SGE deployment costs assume repeated deployment for decades (Smith and Wagner 2018; Smith 2020).

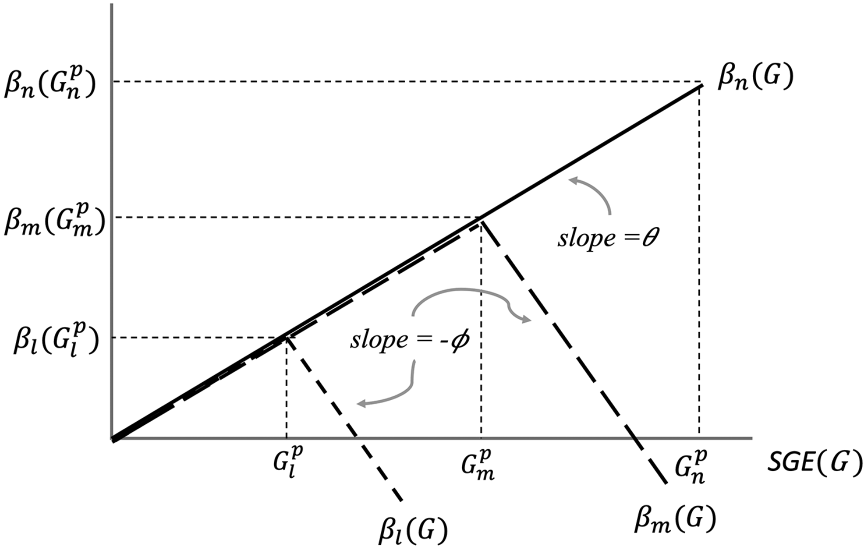

Figure 1. Benefits to SGE deployment depending on preferred levels

Let G_i^{neg} denote the critical level of SGE beyond which the benefit function turns negative for country i, which yields

$$G_i^{neg} = \left(\frac{\phi + \theta}{\phi}\right)\gamma_i(Q^o - Q). \tag{7}$$

2. NONCOOPERATIVE ABATEMENT AND SOLAR GEOENGINEERING

In the absence of an agreement, emissions abatement decisions and SGE deployment are modeled as a two-stage game. Countries independently and simultaneously decide on abatement levels in stage 1. Given aggregate emissions Q determined from stage 1, in stage 2 countries independently and simultaneously make their SGE decisions. The game is solved by backward induction, and so we begin with stage 2.

2.1. Stage 2: Solar Geoengineering

Recall that the distribution of preferred levels of SGE depends on the distribution of γ_i's for all n countries. We denote the highest γ_i as γ^{max}, and the highest preferred level as $G^{max} = \gamma^{max}(Q^o - Q)$. Since the benefit function $\beta_i(G)$ is maximized at country i's preferred level, and given that solar geoengineering is modeled as a best-shot technology, the country with G^{max} will maximize payoffs by choosing $G = G^{max}$ in stage 2.

All other countries with preferred levels less than G^{max} will anticipate that the chosen G will be above their preferred level and so they will not choose a positive G. A country with a preferred level equal to G^{max} is the "free driver" (Wagner and Weitzman

2012; Weitzman 2015) because their choice of SGE "drives" the global temperature. All countries other than the free driver are referred to as "non-free-driving countries" or simply "non-free-drivers."

Lemma 1: In stage 2, the level of SGE deployment is determined by the player(s) with the highest preferred level (i.e., the free driver), resulting in $G = G^{max}$.

While it is clear that $G = G^{max}$ after stage 2 and $\beta_i(G^{max}) > 0$ for the player(s) with the highest preferred level, it is not obvious whether the aggregate benefits from SGE deployment are positive or negative. We know that SGE levels greater than G_i^{neg} will strictly reduce a player's payoff, while levels less than G_i^{neg} increase payoffs (relative to $G = 0$). The aggregate benefit function from SGE can be expressed as

$$\beta(G^{max}) = \sum_{i \in N} \left[\theta G_i^p - \phi\left(G^{max} - G_i^p\right)\right], \tag{8}$$

then collecting G_i^p terms

$$\beta(G^{max}) = \sum_{i \in N} \left[(\theta + \phi)G_i^p - \phi G^{max}\right], \tag{9}$$

and recognizing that the last term is a constant summed n times this becomes

$$\beta(G^{max}) = -\phi n G^{max} + (\theta + \phi)\sum_{i \in N} G_i^p. \tag{10}$$

For all players other than the free driver, the benefits increase when G^{max} decreases. We can now substitute our expressions for G^{max} and G_i^p, and write this in terms of abatement, which is

$$\beta(G^{max}) = (Q^o - Q)\left[-n\phi\gamma^{max} + (\theta + \phi)\sum_{i \in N} \gamma_i\right].$$

The term in [.] consists only of parameters, and Q^o is a constant. The $(Q^o - Q)$ term is nonnegative when abatement is less than socially optimal, and if the term in [.] is negative then SGE at G^{max} results in negative aggregate benefits. Let $\bar{\gamma}$ denote the average level of γ_i. The aggregate benefits from SGE are negative if $\bar{\gamma}/\gamma^{max} < \phi/(\theta + \phi)$ and $Q < Q^o$.

Suppose, for example, that all countries were identical. In this case $\bar{\gamma}/\gamma^{max} = 1$ and so SGE strictly increases aggregate benefits. However, suppose $\bar{\gamma} = 0.4$ and $\gamma^{max} = 0.9$, then SGE at the free-driver outcome is welfare reducing if $\theta < \phi$, since $\theta < \phi$ implies $\phi/(\theta + \phi) \in (0.5, 1)$.[6]

[6] In our study we do not constrain the relationship between θ and ϕ. Weitzman (2015), however, weighs "overdone" geoengineering (our ϕ) at three times the value for "underdone" geoengineering (our θ).

At the individual level, the benefit from SGE at the free-driver outcome is

$$\beta_i(G^{\max}) = (Q^o - Q)[(\theta + \phi)\gamma_i - \phi\gamma^{\max}],$$

and the first term is nonnegative and the second is isomorphic to G_i^{neg}. That is,

$$\beta_i(G^{\max}) < 0 \text{ iff } \frac{\gamma_i}{\gamma^{\max}} < \frac{\phi}{\theta + \phi}.$$

2.2. Stage 1: Emissions Abatement

Given that $G = G^{\max}$ in stage 2, a country's payoff from (5) in stage 1 takes the following form:

$$\pi_i(q_i, Q, G^{\max}) = \frac{bQ}{n} - \frac{c(q_i)^2}{2} + (\theta + \phi)G_i^p - \phi G^{\max}, \tag{11}$$

then in terms of abatement

$$\pi_i(q_i, Q, G^{\max}) = \frac{bQ}{n} - \frac{c(q_i)^2}{2} + (Q^o - Q)[(\theta + \phi)\gamma_i - \phi\gamma^{\max}].$$

The first-order condition for abatement is

$$\frac{b}{n} - cq_i - [(\theta + \phi)\gamma_i - \phi\gamma^{\max}] = 0, \tag{12}$$

with solution

$$q_i^*(G^{\max}) = \underbrace{\frac{b}{cn}}_{1} - \underbrace{\frac{\theta\gamma_i}{c}}_{2} + \underbrace{\frac{\phi(\gamma^{\max} - \gamma_i)}{c}}_{3}. \tag{13}$$

The first term is the dominant strategy abatement level in the world without SGE. The second term is the reduction in abatement (i.e., crowding out) from SGE being a "good," where γ_i is effectively i's degree of substitutability between abatement and SGE. The closer γ_i is to one, the closer the substitutability and the greater the reduction in abatement. The parameter θ is the marginal benefit from a unit of SGE so $\theta\gamma_i$ is the effective forgone benefit from SGE for a unit of abatement. The third term captures what we refer to as the anti-driver incentive, because increasing abatement reduces the harm caused by the free driver choosing SGE beyond a country's preferred level (the "bad" part of the GoB). Recall, the marginal cost of SGE in the "bad" region is ϕ.

Note that abatement cannot be strictly dominated by SGE in our model. For abatement to be strictly dominated the condition is $q_i^*(G^{\max}) \leq 0$, which reduces to $(b/n) - \theta\gamma_i + \phi(\gamma^{\max} - \gamma_i) \leq 0$. Given $b/n \geq \theta$, $\gamma_i \in (0, 1)$ for all $i \in N$, and $\gamma_i \leq \gamma^{\max}$, abatement is strictly positive and will never be completely crowded out by SGE.

Proposition 1: The availability of geoengineering technologies causes the free driver to reduce abatement and causes non-free-driving countries to increase abatement when $\gamma_i/\gamma^{\max} < \phi/(\theta + \phi)$.

Proof: A free driver has $\gamma_i = \gamma^{\max}$ and equation (13) reduces to

$$q_i^*(\gamma^{\max}, G^{\max}) = \frac{b}{cn} - \frac{\theta\gamma^{\max}}{c},$$

which is less than the noncooperative abatement level in a world without geoengineering technologies (from eq. [2]). For all countries with preferred SGE levels less than $G^{\max}$, from equation (13) SGE opportunities will increase abatement when

$$(\theta + \phi)\gamma_i - \phi\gamma^{\max} < 0, \tag{14}$$

which, when rearranged, is

$$\frac{\gamma_i}{\gamma^{\max}} < \frac{\phi}{\theta + \phi}. \tag{15}$$

QED

Proposition 1 and its proof inform us that non-free-driving countries will increase abatement from SGE if the anti-driver incentives are strong enough. The anti-driver incentives are stronger when the gap between SGE preferred levels ($\gamma^{\max} - \gamma_i$) increases and/or the marginal damage imposed by the free driver (ϕ) increases. The intuition is that countries can reduce some of their losses from SGE by decreasing the level of $G^{\max}$, which is reduced by an increase in abatement. Note that the closer ϕ gets to zero, or the closer γ_i is to $\gamma^{\max}$, the weaker the anti-driver incentives and SGE can lead to a decrease in non-free-driver abatement (i.e., crowding out). In these cases, countries earn a positive net-benefit from SGE and are able to save costs by decreasing abatement. Non-free-driver abatement levels, however, remain strictly positive.

This result helps us better understand the “moral hazard” debate, which is centered on the conjecture that SGE opportunities will undermine abatement efforts (e.g., Reynolds 2019; Wagner 2021). We show that this is a possibility, but the opposite outcome can also occur—it depends on the relative magnitude of the anti-driver and crowding-out incentives.

We can aggregate $q_i^*(G^{\max})$ to get $Q^*(G^{\max})$ using (13),

$$Q^*(G^{\max}) = \sum_{i\in N} q_i^*(G^{\max})$$

$$Q^*(G^{\max}) = \sum_{i\in N}\left[\frac{b}{cn} - \frac{(\theta+\phi)\gamma_i - \phi\gamma^{\max}}{c}\right]$$

$$Q^*(G^{\max}) = \frac{b}{c} - \left(\frac{1}{c}\right)\left[-n\phi\gamma^{\max} + (\theta+\phi)\sum_{i\in N}\gamma_i\right]. \tag{16}$$

Then using the average level $\bar{\gamma}$, the sum $\Sigma_{i\in N}\gamma_i = n\bar{\gamma}$, and Nash equilibrium abatement without SGE in (2) $Q^* = b/c$, this becomes

$$Q^*(G^{\max}) = Q^* - \left(\frac{n}{c}\right)[-\phi\gamma^{\max} + (\theta + \phi)\bar{\gamma}], \tag{17}$$

which leads to the following proposition:

Proposition 2: The availability of geoengineering technologies reduces aggregate abatement at the free-driver outcome when $\bar{\gamma}/\gamma^{\max} > \phi/(\theta + \phi)$.

Proof: When $\bar{\gamma}/\gamma^{\max} > \phi/(\theta + \phi)$, $Q^*(G^{\max}) < Q^*$. QED

For example, if all countries have identical preferred levels of SGE, then $\bar{\gamma}/\gamma^{\max} = 1 > \phi/(\theta + \phi) \in (0, 1)$, so the availability of SGE always reduces abatement. Note that proposition 2 relies on the same condition that defines whether the aggregate benefits from SGE are positive or negative. Thus, when the aggregate benefits from SGE at the free-driver level are positive (negative), aggregate abatement decreases (increases).

To complete this section, we consider whether the anti-driver incentives could be strong enough to cause Nash abatement levels to increase up to the optimal level of Q^o. In other words, could SGE be so harmful that non-free-drivers find it optimal to increase abatement enough to prevent deployment.

Proposition 3: If anti-driver incentives are sufficiently strong, then it is possible noncooperative abatement equals Q^o, which prevents SGE deployment.

Proof: In equation (16) we have

$$Q^*(G^{\max}) = \frac{b}{c} - \frac{1}{c}\left[-n\phi\gamma^{\max} + (\theta + \phi)\sum_{i\in N}\gamma_i\right],$$

which is equal to Q^o when

$$Q^*(G^{\max}) = \frac{b}{c} - \frac{1}{c}\left[-n\phi\gamma^{\max} + (\theta + \phi)\sum_{i\in N}\gamma_i\right] = Q^o = \frac{bn}{c},$$

which reduces to

$$\gamma^{\max} = \frac{b(n-1)}{n\phi} + \frac{(\theta + \phi)}{n\phi}\sum_{i\in N}\gamma_i.$$

The average of the distribution of γ_i is $\bar{\gamma} = \Sigma_{i\in N}\gamma_i/n$, and this yields the following condition required for $Q^*(G^{\max}) = Q^o$:

$$\gamma^{\max} = \frac{b(n-1)}{n\phi} + \frac{(\theta+\phi)n\bar{\gamma}}{n\phi}. \tag{18}$$

QED

It is easy to see that the value of $\gamma^{\max}$ in the proof of proposition 3 is decreasing in ϕ, and increasing in b, n, θ, and $\bar{\gamma}$. Therefore it is possible that anti-driver incentives are sufficiently strong that countries increase abatement so that $Q^*(G^{\max}) = Q^o$ and $G^{\max} = 0$. Our study is the first to show this possibility and suggests that SGE has the potential to mitigate the underlying collective action problem in abatement, even in the absence of the international agreement.

3. OPTIMAL ABATEMENT AND SOLAR GEOENGINEERING

The socially optimal level of abatement and SGE is the solution to choosing Q and G to maximize

$$\Pi \equiv \sum_{i\in N} \pi_i(q_i, Q, G) = \sum_{i\in N}\left[\frac{bQ}{n} - \frac{c(q_i)^2}{2}\right] + \beta(G). \tag{19}$$

In stage 2, the optimal level of SGE maximizes the sum of social benefits, $\beta(G^{so}) = \Sigma_{i\in N}[\theta G_i^p - \phi(G^{so} - G_i^p)]$, where the superscript *so* indicates socially optimal levels given SGE opportunities. For any positive abatement gap after stage 1 $(Q^o - Q)$, the socially optimal level of SGE is determined by the distribution of γ_i. Without any loss of generality, order countries from lowest to highest γ_i, such that $\gamma^{\min} \equiv \gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_n \equiv \gamma^{\max}$. This implies $G^{\min} \equiv G_1^p \leq G_2^p \leq \ldots \leq G_n^p \equiv G^{\max}$. For all $\gamma_i < \gamma_{i+1}$ SGE is either a social marginal good or a social marginal bad on all units of G between G_i^p and G_{i+1}^p, given the piecewise linear SGE benefit function for each country (eq. [5]). Thus, the socially optimal level of SGE requires identifying the particular country's G_i^p that maximizes the sum of the benefits.

3.1. Optimal G

Let m be the number of countries that would prefer a lower level of SGE, $G_i^p < G$. For a given number of countries n, if the sum of the social marginal benefits $(n-m)\theta$ from SGE exceeds the sum of the social marginal cost $m\phi$, then increasing m will increase net benefits. We can operationalize this with the following rule: starting at $m = 1$,

$$\text{if } (n-m)\theta > m\phi \text{ then increase } m,$$

$$\text{if } (n-m)\theta \leq m\phi \text{ then stop.}$$

Thus, the socially optimal stopping rule m^o is the minimum value of m such that $(n - m)\theta \leq m\phi$. To illustrate, suppose $\phi = 0$, such that there is no penalty for SGE levels greater than a country's preferred level. In that case $m^o = n$, which means that it is optimal to increase G up until the highest preferred level. As another example, if $\theta = \phi$, then $m^o = n/2$ and the stopping rule results in the median preferred level of SGE.

From the stopping rule we know the optimal level of SGE is

$$G^{so} = G^p_{m^o} = \gamma_{m^o}(Q^o - Q)\,. \tag{20}$$

Note that at m^o, SGE is a marginal good for $n - m^o + 1$ countries with preferred levels higher than or equal to G^{so} and a marginal bad for $m^o - 1$ countries with lower preferred levels. We can then write the socially optimal sum of SGE benefits as

$$\beta(G^{so}) = \sum_{i=1}^{m^o-1}\left[\theta G^p_i - \phi\left(G^{so} - G^p_i\right)\right] + \sum_{i=m^o}^{n}\theta G^{so},$$

$$\beta(G^{so}) = \sum_{i=1}^{m^o-1}(Q^o - Q)[\theta\gamma_i - \phi(\gamma_{m^o} - \gamma_i)] + \sum_{i=m^o}^{n}(Q^o - Q)\theta\gamma_{m^o},$$

which simplifies to[7]

$$\beta(G^{so}) = (Q^o - Q)\left[\gamma_{m^o}[\theta(n - m^o) - \phi m^o] + (\theta + \phi)\sum_{i=1}^{m^o}\gamma_i\right]. \tag{21}$$

3.2. Optimal Q

Given the expression for $\beta(G^{so})$, we can rewrite the payoff function in (19) as

$$\Pi \equiv \sum_{i\in N}\pi_i(q_i, Q, G^{so}) = \sum_{i\in N}\left[\frac{bQ}{n} - \frac{c(q_i)^2}{2}\right] + \beta(G^{so}).$$

Then using (21) this becomes

$$\Pi = \sum_{i\in N}\left[\frac{bQ}{n} - \frac{c(q_i)^2}{2}\right] + (Q^o - Q)\left[\gamma_{m^o}[\theta(n - m^o) - \phi m^o] + (\theta + \phi)\sum_{i=1}^{m^o}\gamma_i\right]. \tag{22}$$

We can rewrite (22) recognizing that $q_i = Q/n$ in the cost-minimizing solution, and solve for Q, which yields the socially optimal level of abatement in stage 1:

$$Q^{so} = \frac{bn}{c} + \frac{n}{c}\left[\gamma_{m^o}[\phi m^o - \theta(n - m^o)] - (\theta + \phi)\sum_{i=1}^{m^o}\gamma_i\right].$$

Then writing this in terms of the social optimum without SGE, Q^o, from (3) we get

[7] In the special case where the stopping rule holds with equality, $\theta(n - m^o) = \phi m^o$, the socially optimal sum of SGE benefits is $\beta(G^{so}) = (Q^o - Q)(\theta + \phi)\Sigma_{i=1}^{m^o}\gamma_i$ and $m^o = n[\theta/(\theta + \phi)]$.

$$Q^{so} = Q^{o} + \frac{n}{c}\left[\gamma_{m^o}[\phi m^o - \theta(n - m^o)] - (\theta + \phi)\sum_{i=1}^{m^o}\gamma_i\right]. \tag{23}$$

This leads to our next proposition.

Proposition 4: The socially optimal level of emissions abatement is reduced by the availability of solar geoengineering technologies if $m^o \geq 1$.

Proof: $Q^{so} < Q^o$ if the term in brackets in equation (23) is negative. When the stopping rule holds with equality $\theta(n - m^o) = \phi m^o$, the term in brackets reduces to $-(\theta + \phi)\Sigma_{i=1}^{m^o}\gamma_i$, which is negative for any positive value of $(\theta + \phi)$, and γ_i. When the stopping rule does not hold with equality, the term in the brackets is $\gamma_{m^o}[\phi(m^o - 1) - \theta(n - m^o + 1)] - (\theta + \phi)\Sigma_{i=1}^{m^o-1}\gamma_i$. For $m = m^o - 1$ the stopping rule implies $\theta(n - m) > \phi m$, therefore all terms in brackets are negative and $Q^{so} < Q^o$. QED

The first term of the socially optimal level of abatement in (23) is the familiar optimal abatement level without opportunities for SGE in (3). To illustrate, suppose $\gamma_i = 0$ for all n countries, then $Q^{so} = Q^o = nb/c$. If we consider another extreme case in which all countries are identical (i.e., $\gamma_i = \gamma$), then (23) reduces to $Q^{so} = (nb/c) - (\theta n^2\gamma/c)$. It is immediately clear that if all countries have the same preferred level, then the optimal level of SGE is not a "bad" for any country, and abatement levels drop relative to a world without SGE opportunities.

At first glance, proposition 4 may appear at odds with proposition 3. Proposition 4 tells us that optimal abatement decreases due to SGE, while proposition 3 tells us that noncooperative abatement can increase due to SGE. The important distinction between the two is the threat of the free driver. In the noncooperative Nash equilibrium the free driver is going to unilaterally choose deployment, and it is the threat of that deployment that can cause an increase in aggregate abatement (i.e., proposition 3). On the other hand, the optimal levels of abatement and SGE are not chosen by the free driver but by maximizing collective welfare. Since SGE is a GoB it is a "good" for some positive level of deployment for all countries, and therefore SGE will cause the social optimum to rely on positive SGE deployment (lower than the free driver's deployment) and less abatement compared to a world without SGE.

4. INTERNATIONAL ENVIRONMENTAL AGREEMENTS

International environmental agreements are institutions designed to help move countries closer to the social optimum. As a point of departure and to provide a baseline for comparison, we first summarize the stability conditions and equilibrium of IEA sizes in a world without SGE opportunities. The fundamental results from abatement models with linear benefits and quadratic costs have been previously published in Barrett (1994, 2003) and Finus and McGinty (2019).

4.1. IEAs in a World without Solar Geoengineering

The IEA is a two-stage game. In stage 1, countries independently and simultaneously decide whether to join an agreement. In stage 2, the members of the agreement choose abatement levels to maximize their joint payoffs. Meanwhile, nonmembers choose their noncooperative abatement levels. Equilibrium agreement sizes are determined by two stability conditions. An IEA is internally stable if no country can earn higher payoffs by leaving the agreement. An IEA is externally stable if no country can earn higher payoffs by joining the agreement.

As demonstrated by Barrett (1994, 2003), given the linear-quadratic payoff function, the largest stable coalition size without SGE opportunities is $s = 3$ members for any values for n, b, or c when countries have identical abatement benefits.

4.2. IEAs in a World with Solar Geoengineering

We now consider agreements to manage global emissions abatement when countries have the opportunity to deploy SGE. The sequence of decision-making follows existing theoretical models and experiments that include both abatement and solar geoengineering (Moreno-Cruz 2015; Cherry et al. 2022); that is, countries make emissions abatement decisions first and then decide on SGE deployment. The IEA now has three stages. In stage 1 (participation), countries decide independently and simultaneously whether to join the IEA. In stage 2 (abatement), the IEA members choose abatement levels to maximize their joint (i.e., collective) payoffs. Meanwhile, nonmembers choose their abatement levels unilaterally (noncooperatively). Finally, in stage 3 (solar geoengineering), countries choose their SGE levels. We consider the simplest agreement: it only governs abatement, which implies that all countries are free to choose their optimal SGE level in stage 3, independent from whether they joined the agreement or not. In section 5, we discuss alternative institutions that govern both abatement and SGE.

4.2.1. Stage 3: Solar Geoengineering

Let the members of an agreement be a coalition S and the set of nonmember countries outside an agreement be denoted T with $S \cup T = N$. The number of agreement members is denoted as $|S| = s$ and indexed by $i = 1, 2, \ldots, s$. The number of nonmembers is $|T| = n - s$ and indexed by $j = 1, 2, \ldots, (n - s)$. Given coalition S forms, denote total abatement by members and nonmembers as $Q(S) = \Sigma_{i=1}^{s} q_i^s + \Sigma_{j=s+1}^{n} q_j^t$.

In an agreement that only governs abatement, the country (member or nonmember) with the highest preferred level will act as the free driver and choose $G(S) = G^{\max}(S) = \gamma^{\max}(Q^o - Q(S))$ in stage 3. Since $G(S)$ is decreasing in total abatement, the exact level of SGE is a function of the abatement decisions made by the s members and $n - s$ nonmembers in stage 2.

4.2.2. Stage 2: Abatement

In stage 2, the $n - s$ free-riding nonmembers each choose their noncooperative Nash abatement levels $q_i^*(G^{\max})$ from equation (13), which we now label as $q_j^t(G^{\max})$ and include here.

$$q_j^t(G^{\max}) = \frac{b}{cn} - \frac{\theta\gamma_j}{c} + \frac{\phi\left(\gamma^{\max} - \gamma_j\right)}{c}.$$

Meanwhile, the s coalition members choose q_i^s to maximize joint payoffs. The problem is

$$\max_{q_i^s}\sum_{i\in S}\pi_i^s = \sum_{i\in S}\frac{bQ}{n} - \frac{c(q_i^s)^2}{2} + \theta G_i^p - \phi\left(G^{\max} - G_i^p\right), \tag{24}$$

when simplifying and expressing in terms of abatement

$$\max_{q_i^s}\sum_{i\in S}\pi_i^s = \sum_{i\in S}\frac{bQ}{n} - \frac{c(q_i^s)^2}{2} + (Q^o - Q)[(\theta + \phi)\gamma_i - \phi\gamma^{\max}]. \tag{25}$$

The first-order condition is

$$\frac{bs}{n} - cq_i^s - \sum_{i=1}^{s}[(\theta + \phi)\gamma_i - \phi\gamma^{\max}] = 0, \tag{26}$$

and when solving

$$q_i^s(G^{\max}) = \frac{bs}{nc} - \frac{1}{c}\left[(\theta + \phi)\sum_{i=1}^{s}\gamma_i - s\phi\gamma^{\max}\right]. \tag{27}$$

Total emissions abatement in stage 2 is therefore

$$Q(S, G^{\max}) = \sum_{i\in S} q_i^s(G^{\max}) + \sum_{j\in T} q_j^t(G^{\max}). \tag{28}$$

4.2.3. Stage 1: Participation

In the first stage all countries decide independently and simultaneously whether to join the IEA. An equilibrium IEA is one that is both internally and externally stable. Internal stability requires that no member could increase their payoff by leaving the agreement (i.e., $IS_i(S) = \pi^s(S) - \pi^t(S\backslash i) \geq 0$), and external stability requires that no nonmember could increase their payoff by joining (i.e., $ES_i(S) = \pi^t(S) - \pi^s(S \cup i) \geq 0$).

Proposition 5: When countries have identical benefits from solar geoengineering deployment (i.e., $\gamma_i = \gamma$), the unique stable coalition size is $s^* = 3$.

Proof: If countries have identical benefits from SGE, $\gamma_i = \gamma$, signatory, nonsignatory, and aggregate abatement can be expressed as

$$q^t = \frac{b}{cn} - \frac{(\theta+\phi)\gamma - \phi\gamma}{c} = \frac{b - n\theta\gamma}{cn},$$

$$q^s = \frac{bs}{cn} - \frac{1}{c}\left[(\theta+\phi)\sum_{i=1}^{s}\gamma - s\phi\gamma\right] = \frac{bs - ns\theta\gamma}{cn} = \frac{s(b - n\theta\gamma)}{cn}, \quad (29)$$

$$Q(s) = sq^s + (n-s)q^t = \frac{(s^2 - s + n)(b - n\theta\gamma)}{cn},$$

where all subscripts are dropped for clarity. In this special case, the internal stability (IS) condition can be expressed as

$$\mathrm{IS}(s) = \left(\frac{b - n\theta\gamma}{n}\right)[Q(s) - Q(s-1)] - \frac{c\left[(q^s)^2 - (q^t)^2\right]}{2}. \quad (30)$$

Using $Q(s)$ from (29) and solving $Q(s-1) = \{[s^2 - 3s + 2 + n](b - n\theta\gamma)\}/cn$, the first term in (30) reduces to $[2(s-1)(b - n\theta\gamma)^2]/cn^2$. Given the parameter restriction in (6), $b/n \geq \theta$ stating that the individual marginal benefit from SGE cannot exceed that of abatement and that $\gamma \in (0, 1)$, the second term in (30) can be reduced to $-\{[(b - n\theta\gamma)^2(s^2 - 1)]/2cn^2\}$. Combining both terms yields

$$\mathrm{IS}(s) = \frac{(b - n\theta\gamma)^2[4(s-1) - (s^2 - 1)]}{2cn^2},$$
$$\mathrm{IS}(s) = \frac{(b - n\theta\gamma)^2(-s^2 + 4s - 3)}{2cn^2} \geq 0 \Leftrightarrow s \leq s^* = 3. \quad (31)$$

QED

From previous studies, we know that with linear-quadratic payoffs and $\gamma = 0$, the largest stable IEA size is $s = 3$. Proposition 5 illustrates that this result holds when $\gamma > 0$ as well. The availability of solar geoengineering opportunities does not alter the size of stable IEAs when countries have the same preferred levels.

When countries have heterogeneous benefits from SGE, the coalition sizes that satisfy the stability conditions will depend on the specific distribution of γ's. The stability conditions are much more complicated in this case. The internal stability condition is

$$\mathrm{IS}_i(S) = \left(\frac{b - n[(\theta+\phi)\gamma_i - \phi\gamma^{\max}]}{n}\right)[Q(S) - Q(S\backslash i)] - \frac{c\left[(q_i^s)^2 - (q_i^t)^2\right]}{2}. \quad (32)$$

Note that the IS condition is written in terms of the coalition structure S rather than simply the number of agreement members s (as in [30]). The is because the individual and aggregate abatement levels of the coalition will depend on which particular countries are members. The complexity of the IS condition in (32) does not allow for an analytical solution in the same way that the symmetric case led to proposition 5. That is, we are not able to explicitly characterize the largest stable agreement size when countries have heterogeneous preferred levels of SGE.

4.3. Anti-driver Incentives and Coalition Stability

The stability of IEAs under the threat of SGE is challenged by two competing incentives. The first incentive, which is a familiar free-rider incentive, is the additional payoff a defecting member achieves by leaving the agreement since they may lower their abatement responsibilities. The other incentive is the anti-driver incentive, which is the additional payoff achieved by the collective increase in abatement as a member of the coalition; this is further dampening the free-driver externality.

Anti-driver incentives increase with the marginal damage of SGE beyond preferred levels (ϕ) and with the gap between the free driver's preference and a nondriver i's preference ($\gamma^{\max} - \gamma_i$). At one extreme, anti-driver incentives turn to zero if all countries are identical and/or the marginal damage from too much SGE is zero. At the other extreme, anti-driver incentives can be so strong that they cause noncooperative abatement levels to increase enough to prevent SGE from being deployed in a noncooperative Nash equilibrium (proposition 3).

In the remainder of this section, we use numeric examples (and include supporting analytical results) to explore the relationship between coalition stability and anti-driver incentives. The first example is a "crowding-out" case in which anti-driver incentives are relatively low, the largest stable coalition is less than three countries, and SGE options cause a reduction in abatement compared to a world without these technologies. The second example is a case in which the anti-driver incentives are relatively high, the largest stable coalition consists of $n - 1$ countries, and the threat of SGE triggers a substantial increase in abatement.

4.3.1. Example 1: Relatively Low Anti-driver Incentives

We consider a simple case in which $n = 5$, which is a collection of countries small enough to easily test stability conditions but sufficiently large to illustrate the interaction of abatement, stability, and anti-driver incentives. We later discuss the impact of increasing n. Let $\gamma_1 = 0.2$, $\gamma_2 = 0.4$, $\gamma_3 = 0.5$, $\gamma_4 = 0.6$, and $\gamma_5 = \gamma^{\max} = 0.8$. Note that $\bar{\gamma} = 0.5$, and let $b = c = 1$ and $\phi = \theta = 0.1$.

Table 1 contains abatement levels and payoffs for the noncooperative Nash equilibrium, the full participation IEA ($s = n$), and the social optimum for each of the five countries. The Nash equilibrium abatement and payoffs are equivalent to those under trivial IEAs of size $s = 1$. Note that the noncooperative Nash level of SGE (i.e., the free driver's payoff-maximizing level) is $G^{\max}(Q^*) = 3.28$. In the Nash equilibrium, country 1 takes on the highest level of abatement and earns the lowest payoff. The intuition is that country 1 has the lowest preferred level of SGE (thus the strongest

Table 1. Abatement and Payoffs Given Relatively Weak Anti-driver Incentives

		Nash Equilibrium		Full Participation IEA		Social Optimum	
Country i	γ_i	q_i^*	π_i^*	q_i^f	π_i^f	q_i^{so}	π_i^{so}
1	.2	.24	−.013	.90	.475	.83	.477
2	.4	.20	.160	.90	.495	.83	.511
3	.5	.18	.246	.90	.505	.83	.528
4	.6	.16	.331	.90	.515	.83	.528
5	.8	.12	.501	.90	.535	.83	.528
		$Q^* = .90$	$\Pi^* = 1.225$	$Q^f = 4.50$	$\Pi^f = 2.525$	$Q^{so} = 4.15$	$\Pi^{so} = 2.572$

anti-driver incentive), and increasing abatement is a way to protect itself from some of the damages imposed by excessive SGE deployment from the free driver. On the other hand, the free driver, country 5, takes on the lowest abatement and earns the highest payoff since SGE is deployed at their preferred level. In comparison to a world without SGE, aggregate abatement is lower with SGE (0.90 vs. 1.00) and aggregate payoff is higher with SGE (1.225 vs. 0.90). In particular, countries 3, 4, and 5 abate less and have a higher payoff in the Nash equilibrium with SGE, while countries 1 and 2 abate more (country 1) or the same (country 2) and have lower payoffs.

The superscript f denotes values in an IEA with full participation (i.e., the grand coalition). The IEA only governs emissions abatement, and therefore SGE is again determined by the free driver's preferred level, which is $G^{\max}(Q^f) = 0.40$. The agreement members choose abatement levels to maximize collective payoffs, requiring that abatement (and abatement cost) is the same for all players and hence cost-effective. The IEA with full participation results in a Pareto improvement for all countries. Note that country 1, having the lowest preferred level of SGE, has the most to gain from full cooperation relative to the noncooperative outcome. However, country 1 still earns the lowest payoff of all n countries. Aggregate abatement under full participation is lower with SGE opportunities compared to without (4.50 vs. 5.0), and aggregate payoffs are higher (2.53 vs. 2.50). For those with preferred SGE levels less than the free driver, the individual gain from an agreement with full participation is increasing in the free driver's preferred level of SGE.

The superscript *so* in table 1 denotes socially optimal levels of abatement and payoffs. In the social optimum, SGE deployment is determined by equation (20). Since $\theta = \phi$ in this example, the optimal level of G is the median preferred level (country 3's level, $G = 0.425$). Note that abatement is lower and SGE is higher in the social optimum compared to an IEA with full participation. The intuition is that with the grand coalition, the agreement members fully internalize the benefits of abatement, which includes internalizing the indirect benefit of reducing the damages imposed by the free driver. Compare this with the social optimum, in which SGE is not determined by the free driver but by the stopping rule and equation (20). In short, the threat of free-driver deployment motivates the IEA members to collectively increase abatement beyond the socially optimal level.

For a coalition to be an equilibrium IEA, it must be the case that no member could be better off by defecting and leaving the agreement. Ultimately, we are able to show that the largest stable IEA given our distribution of γ's and chosen parameter values is $s = 2$, although not all coalitions of two members are stable. Table 2 contains abatement and payoffs for all stable IEAs in our example. Note that stable coalitions with members that have lower preferred levels of SGE will abate more (i.e., stronger anti-driver incentives) and achieve higher aggregate payoffs compared to stable coalitions consisting of members with higher preferred SGE levels. In all cases, stable IEAs are able to only marginally improve upon the noncooperative outcome. This can be confirmed by comparing values from table 2 with the noncooperative values from table 1. Also table 2 shows that it is possible for the free driver to be part of a stable IEA. The intuition is that the free driver's benefit from the collective abatement achieved under an IEA is greater than the additional cost of abatement relative to the noncooperative level. The stable coalition that includes the free driver generates the lowest aggregate abatement and the highest SGE deployment compared to all other stable IEAs.

4.3.2. *Example 2: Relatively High Anti-driver Incentives*

Now consider a situation in which the anti-driver incentives are relatively high. Recall, proposition 3 provides a condition in which anti-driver incentives from the threat of SGE are sufficiently powerful that Nash equilibrium abatement results in zero deployment (i.e., $Q^* = Q^o$). In this example we choose parameter values for ϕ, γ_{max}, and $\bar{\gamma}$ for which proposition 3 holds. We show that in this case, a coalition size of $n - 1$ is technically internally and externally stable, but aggregate abatement under the grand coalition is equivalent to the Nash equilibrium.

We keep $n = 5$, $b = c = 1$, and $\theta = 0.10$ and impose stronger anti-driver incentives. In particular, the parameter value for ϕ is set relatively high at $\phi = 1.475$, and the distribution of γ_i is $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.1$ and $\gamma_5 = \gamma^{\max} = 0.8$ (a big gap between the free driver's preferred level of SGE and all other countries' preferred levels). Table 3 shows abatement levels and payoffs for the noncooperative Nash equilibrium, the full participation IEA ($s = n$), and the social optimum for each of the five countries. Note that the noncooperative Nash level of SGE (i.e., the free driver's payoff-maximizing level) is $G^{\max}(Q^*) = 0$. In the Nash equilibrium, the free driver takes on the lowest level of abatement (0.12) and earns the highest payoff (0.993). The other $n - 1$ countries take on higher levels of abatement (1.22) and earn lower profits (0.254).

In an IEA with full participation, $G^{\max} = 0$, and all n members abating the same level of emissions leads to aggregate welfare gains relative to the Nash equilibrium due

Table 2. Member and Nonmember Abatement and Payoffs for All Stable IEAs

Coalition	Member Abatement	Member Payoffs	Nonmember Abatement	Nonmember Payoffs	Q	$G^{\max}$	Π
1,2	$q_i^s = .44$	$\pi_1 = .025; \pi_2 = .171$	$q_3 = .18; q_4 = .16; q_5 = .12$	$\pi_3 = .325; \pi_4 = .402; \pi_5 = .554$	1.34	2.93	1.49
1,3	$q_i^s = .43$	$\pi_1 = .029; \pi_3 = .249$	$q_2 = .20; q_4 = .16; q_5 = .12$	$\pi_2 = .244; \pi_4 = .398; \pi_5 = .551$	1.32	2.94	1.47
2,3	$q_i^s = .38$	$\pi_1 = .184; \pi_3 = .258$	$q_2 = .24; q_4 = .16; q_5 = .12$	$\pi_2 = .078; \pi_4 = .392; \pi_5 = .546$	1.28	2.98	1.46
2,4	$q_i^s = .36$	$\pi_2 = .187; \pi_4 = .337$	$q_1 = .24; q_4 = .18; q_5 =. 12$	$\pi_1 = .074; \pi_4 = .311; \pi_5 = .544$	1.26	2.99	1.45
3,4	$q_i^s = .34$	$\pi_3 = .265; \pi_4 = .341$	$q_1 = .24; q_2 = .20; q_5 = .12$	$\pi_2 = .069; \pi_4 = .228; \pi_5 = .542$	1.24	3.01	1.44
4,5	$q_i^s = .28$	$\pi_4 = .350; \pi_5 = .502$	$q_1 = .24; q_2 = .20; q_3 = .18$	$\pi_2 = .054; \pi_4 = .216; \pi_5 = .296$	1.18	3.06	1.42

Table 3. Abatement and Payoffs Given Relatively Strong Anti-driver Incentives

		Nash Equilibrium		Full Participation IEA		Social Optimum	
Country i	γ_i	q_i^*	π_i^*	q_i^f	π_i^f	q_i^{so}	π_i^{so}
1	.1	1.22	.254	1.0	.50	.95	.501
2	.1	1.22	.254	1.0	.50	.95	.501
3	.1	1.22	.254	1.0	.50	.95	.501
4	.1	1.22	.254	1.0	.50	.95	.501
5	.8	.12	.993	1.0	.50	.95	.501
	$\bar{\gamma} = .20$	$Q^* = 5.0$	$\Pi^* = 2.01$	$Q^f = 5.00$	$\Pi^f = 2.50$	$Q^{so} = 4.75$	$\Pi^{so} = 2.51$

to cost-effective emissions control. Aggregate payoffs are lower in the Nash equilibrium because without cooperation the free driver chooses a low abatement level, forcing the other $n - 1$ countries to increase theirs to protect themselves from SGE damages. With a full participation IEA, each country has the same payoff, unlike table 1, because in this case there is no deployment of SGE.

In the social optimum, however, $G > 0$. The intuition is that some positive level of G is universally beneficial given that it is no higher than the lowest preferred level. To locate the socially optimal level we use the stopping rule in section 3, which yields $G^{so} = 0.025$. Compared to the full participation IEA, the social optimum is characterized by positive deployment of SGE, lower aggregate abatement, and higher payoffs.

An IEA with full cooperation, however, is not internally stable. The free driver has an incentive to leave and free-ride off the abatement of the other $n - 1$ countries. The free driver earns a payoff of 0.5 in the full cooperative outcome compared to 0.993 when it remains outside of an agreement of size $n - 1$. This result is not specific to our numeric example and can be defined more generally in the following proposition:

Proposition 6: A free-driver country will always defect from the grand coalition if $Q = Q^o$ holds after it leaves the coalition.

Proof: A free driver will defect if the abatement cost inside the agreement at the grand coalition is greater than the free-driver abatement cost outside the agreement; that is if

$$\frac{c(q^s)^2}{2} > \frac{c(q^t)^2}{2},$$

where $q^s = q^o = b/c$ and $q^t = (b/cn) - (\theta\gamma^{\max}/c)$. The inequality reduces to

$$b(n - 1) + n\theta\gamma^{\max} > 0,$$

which always holds. Thus, the free driver will never want to be in a grand coalition that solves the abatement problem ($Q = Q^o = bn/c$) if the anti-driver effect is strong enough so that $Q = Q^o = bn/c$ with $n - 1$ signatories. QED

The intuition for proposition 6 is straightforward. If $Q = Q^o$ when $s = n - 1$, then SGE will not be deployed. Thus, the only difference from being inside or outside the coalition for the free driver is the increase in abatement cost it incurs as a member of a coalition size $s > 1$. For any distribution of γ_i, IEA members internalize the benefits of abatement, which includes the indirect impact of reducing damages from SGE. Since the free driver cannot gain from this collective anti-driver effect, it will always defect from the grand coalition.

Given our parameter choices in the current section, a coalition of $s = n - 1 = 4$ that excludes the free driver is technically both internally and externally stable. None of the four IEA members are better off leaving the coalition (they receive the same payoff of 0.254 both in and outside the coalition) and the free driver has no incentive to join. Individual abatement levels and payoffs in the coalition of $s = 4$ with the free driver as the only nonmember are identical to those under the Nash equilibrium from table 3. That a large stable coalition is possible, however, does not suggest it is likely. In our numeric example with strong anti-driver incentives, the marginal damage from SGE beyond the preferred level (ϕ) is over 15 times the size of the marginal benefit of SGE below the preferred level (θ). While there is some support for assuming that the marginal damages exceed the marginal benefits for SGE (e.g., Weitzman 2015), these levels appear extreme.

Through our two numeric examples with heterogeneity in SGE benefits, we have demonstrated that the largest stable IEA can differ compared to a world without SGE opportunities. The same qualitative results can be demonstrated when n is larger. Stable coalitions can be as small as two countries when anti-driver incentives are low, and as large as $n - 1$ countries when anti-driver incentives are high. While the precise size of the largest stable IEA depends on parameter values, it is important to recognize that the same fundamental tensions remain between countries with different SGE benefits, no matter how those benefits are distributed. Mainly, the free-driving country strictly benefits from deployment levels that match their preferred level, and in an agreement that only governs abatement, there is no way to directly stop the free driver from deploying excessive SGE. The best the other countries can do is limit SGE damages indirectly by increasing their collective abatement. This is why coalitions consisting of members with relatively low preferred levels of SGE will always abate more than equal size coalitions with members that have relatively high preferred levels of SGE.

4.4. An Example Using Regional Vulnerability to Climate Change Damages

For a final illustrative example of IEA formation under the threat of SGE, we choose parameter values for γ_i's that are based on climate change damage vulnerability estimates. The metrics we use are derived from the Notre Dame Global Adaptation

Initiative (ND-GAIN).[8] The ND-GAIN vulnerability index is a country-level metric based on how much exposure countries will have to climate change, how dependent countries are to sectors impacted by climate change, and their current capacity to adapt. The vulnerability index is a relative ranking truncated between 0 and 1 with higher values indicating greater vulnerability. We divided the globe into the five United Nations regional groups, which include Western European and Other States, Eastern European States, Latin American and Caribbean States, Asia-Pacific States, and African States.[9] We use the ND-GAIN metric as a country's γ value, which captures the heterogeneity in preferred levels for SGE deployment in our model. The underlying assumption is that the more vulnerable a country is to climate damages, the higher its preferred level of SGE deployment. The γ_i value for each region ($n = 5$) is determined by the single most vulnerable party in its group.[10] The African States make up the group most vulnerable to climate change ($\gamma = 0.678$) and act as the free driver in all scenarios. The Western European and Other States and the Eastern European States are the least vulnerable ($\gamma = 0.355$ and $\gamma = 0.413$, respectively), and the Latin American and Caribbean States and Asia-Pacific States are in between ($\gamma = 0.514$ and $\gamma = 0.616$, respectively).

Our example explores three different combinations of ϕ and θ. Following Weitzman (2015), we consider the case where the marginal damages from excess deployment are three times the marginal benefits from SGE when below the preferred level (i.e., $\phi = 3 \times \theta$). The other two combinations vary the relative damages from excessive SGE relative to this baseline; one in which the marginal damage is equal to the marginal benefit from SGE to explore weaker anti-driver incentives ($\phi = \theta$), and one in which marginal damages are nine times the marginal benefits to explore stronger anti-driver incentives (i.e., $\phi = 9 \times \theta$). For the remaining parameters, we use the same values from previous examples ($b = c = 1$ and $\theta = 0.1$).

Table 4 contains results for all stable IEAs based on regional vulnerability. The table includes the specific agreement members, individual and aggregate abatement, free-driver deployment of SGE (i.e., $G^{\max}$), and total payoffs. The final column is an efficiency metric following McGinty (2007). It is the proportion of the payoff gap (i.e., difference between aggregate payoffs in the social optimum and Nash equilibrium) that is closed by the coalition.

From table 4 it is immediately clear that the largest stable agreement consists of two regions for all ϕ and θ combinations. The most efficient agreements in all three scenarios have Western European and Other states and Eastern European states as members. Those agreements are most efficient because the least vulnerable members take on relatively

[8] See https://gain.nd.edu/our-work/country-index/rankings/.

[9] See https://www.un.org/dgacm/en/content/regional-groups.

[10] Using the average in each group instead of the maximum does not change the relative rankings.

Table 4. Stable Agreements Based on the Five United Nations Regional Groups

Agreement Members	Member Abatement	Nonmember Abatement	Q	$G^{\max}$	Π	Efficiency Gain (%)
		$\phi = .1 ; \theta = .1$				
Western European and Other States; Eastern European States	.38	$q_{lac} = .17; q_{as} = .14; q_{afs} = .13$	1.21	2.57	1.70	18.8
Western European and Other States; Latin American and Caribbean States	.36	$q_{ee} = .19; q_{as} = .15; q_{afs} = .13$	1.19	2.59	1.69	17.9
Western European and Other States; Asia-Pacific States	.34	$q_{ee} = .19; q_{lac} = .17; q_{afs} = .13$	1.17	2.60	1.68	17.0
Eastern European States; Latin American and Caribbean States	.35	$q_{we} = .20; q_{as} = .14; q_{afs} = .13$	1.17	2.59	1.69	17.9
Eastern European States; Asia-Pacific States	.33	$q_{we} = .20; q_{lac} = .17; q_{afs} = .13$	1.15	2.61	1.68	17.0
Eastern European States; African States	.32	$q_{we} = .20; q_{lac} = .17; q_{as} = .14$	1.14	2.62	1.68	17.0
Latin American and Caribbean States; Asia-Pacific States	.31	$q_{we} = .20; q_{ee} = .19; q_{as} = .14$	1.13	2.63	1.67	16.1
Latin American and Caribbean States; African States	.30	$q_{we} = .20; q_{ee} = .19; q_{as} = .14$	1.12	2.63	1.67	16.1
Asia-Pacific States; African States	.28	$q_{we} = .20; q_{ee} = .19; q_{lac} = .17$	1.10	2.64	1.66	15.2
		$\phi = .3; \theta = .1$				
Western European and Other States; Eastern European States	.50	$q_{lac} = .20; q_{as} = .16; q_{afs} = .13$	1.49	2.38	1.24	18.7
Western European and Other States; Latin American and Caribbean States	.46	$q_{ee} = .24; q_{as} = .16; q_{afs} = .13$	1.45	2.41	1.23	18.1
Eastern European States; Latin American and Caribbean States	.44	$q_{we} = .26; q_{as} = .16; q_{afs} = .13$	1.42	2.43	1.23	18.1
Latin American and Caribbean States; Asia-Pacific States	.35	$q_{we} = .26; q_{ee} = .24; q_{afs} = .13$	1.34	2.48	1.19	15.6
Asia-Pacific States; African States	.29	$q_{we} = .26; q_{ee} = .24; q_{afs} = .20$	1.28	2.53	1.16	13.8
		$\phi = .9; \theta = .1$				
Western European and Other States; Eastern European States	.85	$q_{lac} = .30; q_{as} = .19; q_{afs} = .13$	2.33	1.81	.26	24.8
Eastern European States; Latin American and Caribbean States	.69	$q_{we} = .45; q_{as} = .19; q_{afs} = .13$	2.17	1.92	.21	23.1

Note. The subscripts *we*, *ee*, *lac*, *as*, and *afs* denote Western European and Other States, Eastern European States, Latin American and Caribbean States, Asia-Pacific States, and African States, respectively.

high abatement levels, which in turn decreases the free driver's deployment of SGE and reduces the negative externalities. Related to this, agreement members take on more abatement when excessive SGE is more damaging (when ϕ increases relative to θ).

In summary, our example using real-world climate vulnerability metrics demonstrates that stable coalitions are expected to be small and will likely achieve only modest efficiency gains relative to the noncooperative baseline. The example provides insight into the collective response of low-vulnerability regions to the threat of SGE deployment by a highly vulnerable free driver. Of course there are several limitations. Consistent with our theoretical model, the exercise only explores heterogeneity in preferred levels of SGE while leaving all other parameters constant. Although outside the scope of this study, future research could advance this study by predicting stable IEAs using a fully heterogeneous and calibrated simulation.

5. POLICY IMPLICATIONS AND CONCLUSION

The emergence of SGE technologies presents more social complexities than technical complexities. Indeed, the discussions and debates largely focus on policies that govern SGE research, development, and deployment (Aldy et al. 2021; Biermann et al. 2022; Wieners et al. 2023). Underlying these issues are the implications of SGE for strategic responses to emissions abatement and international institutions of governance. Our game-theoretic analysis provides insights on both strategic responses to SGE and governance of SGE.

We show that SGE opportunities can lead to more or less emissions abatement depending on how costly the free driver's level of SGE is for the other countries (triggering anti-driver incentives). When anti-driver incentives are relatively strong, the threat of SGE can cause an increase in aggregate emissions abatement. On the other hand, when anti-driver incentives are relatively weak, the threat of SGE can cause countries to decrease abatement. This result contributes to our understanding of the "moral hazard" argument, which is centered on the conjecture that SGE opportunities will undermine abatement efforts (e.g., Reynolds 2019; Wagner 2021). We show that this is a possibility, but the opposite outcome can also occur. It depends on the distribution of SGE benefits. Indeed, if the anti-driver incentives are strong enough, noncooperative abatement levels can be high enough to prevent SGE deployment in the absence of an international environmental agreement.

We also demonstrate that the social optimum with SGE opportunities requires lower abatement levels compared to the social optimum without SGE, and optimal SGE deployment is positive. The intuition for this is that SGE is a "good" before it turns "bad" for all countries, and since the social optimum is the solution to joint maximization of both abatement and SGE, there is some level of substitution between both investments. The important point is that inefficiencies with SGE deployment are caused by the free driver, the country that deploys SGE beyond the socially optimal level.

The IEA we consider is the simplest institution, one that only governs emissions abatement, and all countries are free to choose their SGE deployment. When countries have homogeneous benefits from SGE, the largest stable IEAs consist of three members, which means that SGE opportunities do not alter the main conclusion from the existing literature using the same underlying global public-goods model (Barrett 2003; Finus 2008; Finus and McGinty 2019). With heterogeneous benefits, stability is challenged by two competing incentives. The familiar free-rider incentive is the additional payoff a defecting member achieves by leaving the agreement since they lower their abatement responsibilities. The anti-driver incentive is the additional payoff achieved by joining an agreement and further dampening the free-driver externality through increased collective abatement. Ultimately, we find that the largest stable IEA depends, in part, on the relative magnitude of anti-driver incentives; it can be smaller or larger than an equilibrium IEA without SGE. While we find that the threat of a menacing free driver can lead to more abatement, agreements are unable to dramatically improve upon the noncooperative outcome, a result consistent with findings in the established game-theoretic literature on IEAs.

The approach we take in this study follows others that assume the cost of SGE is relatively small compared to abatement (Barrett 2008; Weitzman 2015), and therefore we neglect it in the model. The low-cost assumption is based on the best estimates from the engineering and atmospheric science literature (Keith et al. 2010; Smith and Wagner 2018; Smith 2020). One may wonder why, if the low-cost assumption is justified, we are yet to observe large-scale SGE deployment/free driving in the real world. While the basic science of injecting aerosols into the atmosphere to reflect a fraction of solar radiation away from earth is understood, more research is needed prior to pursuing SGE to better understand the implications and unintended consequences (Aldy et al. 2021). Even if a country or collective of countries decide to pursue SGE, it would likely take years to develop the program (Smith 2020). Our model looks ahead at the strategic implications in a world prepared to deploy SGE.

The results from our theoretical analysis provide insights on how the threat of SGE deployment can alter the strategic landscape of international climate negotiations. The conventional wisdom is that the availability of SGE will crowd out efforts to reduce GHG emissions and hence undermine efforts to achieve a meaningful climate agreement. Although uncertainties exist (Harding et al. 2020), the threat of a free driver likely will emerge among the countries that are most harmed and threatened by climate change, as they will prefer the most aggressive action to cool the planet (Weitzman 2015). Those include vulnerable countries in the global South and small island developing states (SIDS) that emit less and currently have a weaker position in climate negotiations focused only on abatement. Our model shows that the emergence of SGE can shift the balance of power by providing leverage to those nations most vulnerable to climate change (free driver). Other nations (anti-drivers), particularly those in the global North, have a strategic incentive to increase abatement efforts to defuse the threat of SGE deployment. Thus, with SGE, climate negotiations may lead to greater aggregate abatement with the distribution of efforts shifting away from the free-driver state or collective.

Our results also indicate that for some positive level of SGE deployment, international climate targets can be realized with less abatement because SGE opportunities reduce the gap between noncooperative and socially optimal abatement. Since agreements may require less collective abatement with SGE, the threat of SGE deployment could improve the prospects for successful international cooperation on emissions. Previous work has shown that it is easier to get high participation in agreements when the abatement gap is smaller (Barrett 1994). In addition, the relationship between SGE and emissions also implies a lower social cost of carbon since SGE deployment reduces climate damages (which is consistent with Heutel et al. [2018] and Acemoglu and Rafey [2023]). However, the GoB nature of SGE may mean that a nation with a very low target level of SGE would have a much higher national cost of carbon to reduce own emissions and hence reduce SGE deployment. Regardless, it is crucial that future research address the current lack of empirical estimates regarding optimal SGE levels to accurately determine the free-driver and anti-driver incentives in an increasingly warmer world.

What is made clear in our analysis is that when an IEA only governs emission abatement (e.g., Paris Agreement), in most cases the free driver will deploy SGE at a level that is inefficiently high. An obvious extension to our analysis is considering different types of IEAs, particularly institutions that govern both abatement and SGE. One possibility, which is motivated by the ongoing debate about the acceptability of SGE technologies (e.g., Biermann et al. 2022), is a non-use agreement, in which members to an IEA agree not to deploy SGE (Heyen et al. [2019] refer to this a "moratorium agreement"). Of course this type of agreement can only exacerbate the already strong free-rider incentives we observe in the abatement-only agreement. Another possible agreement structure inspired by the established IEA literature is a collective maximization agreement in which countries jointly determine abatement and SGE levels to maximize the coalition's payoffs. Or a close variant to this, a no-harm agreement in which SGE levels are set to match the lowest preferred level of the members (i.e., no member suffers any losses from too much SGE).

Previous studies have demonstrated the important role financial transfers play at maintaining stability and increasing the effectiveness of IEAs among heterogeneous countries (McGinty 2007). Given the heterogeneity in SGE benefits, one could consider how similar transfer rules could be designed to better align the would-be free driver's preferred level with the social optimum (e.g., Ghidoni et al. 2023). The general idea would be that if IEA members earn a significant surplus from jointly cooperating, it may be possible for the coalition to compensate a potential free driver so that they are weakly better off reducing their SGE deployment to a more efficient level. The challenge with this scheme is that, even if it is possible to pay the country with the highest preferred level to reduce their SGE deployment, then the country with the second-highest preferred level is the new potential free driver. While a formal analysis is needed, it is easy to see how an agreement with transfers could unravel because of the fluidity of the potential free-driving country.

Our theoretical analysis is a first step in exploring international agreements when solar geoengineering is modeled as a good or bad (GoB) and governance is complicated by the threat of a free driver (following Weitzman 2015; Wagner 2021). Of course, our study has limitations. One limitation is that we only consider one functional form for emissions abatement and SGE benefits/costs. For that reason we cannot generalize our results to other settings, including an often-explored case in which both the benefits and costs to abatement are quadratic (e.g., Barrett 1994). Another limitation is that we avoid altogether the potential risks from SGE technologies. SGE technologies are new and untested, which introduces uncertainty about costly side effects, and therefore extending our analysis to include these features is a necessary next step. There is also the related political risk of deploying SGE and becoming a pariah state, which could impact decision-making. Finally, our model does not allow countries to invest in counter-SGE, which may help moderate the free-driver problem but at an efficiency loss. Despite these limitations and future opportunities, our research provides new and important insights into the global governance challenge of managing climate change under the threat of solar geoengineering.

REFERENCES

Abatayo, Anna L., Valentina Bosetti, Marco Casari, Riccardo Ghidoni, and Massimo Tavoni. 2020. Solar geoengineering may lead to excessive cooling and high strategic uncertainty. *Proceedings of the National Academy of Sciences* 117 (24): 13393–98.

Acemoglu, Daron, and Will Rafey. 2023. Mirage on the horizon: Geoengineering and carbon taxation without commitment. *Journal of Public Economics* 219:104802.

Aldy, Joseph, Tyler Felgenhauer, William A. Pizer, Massimo Tavoni, Mariia Belaia, Mark E. Borsuk, Arunabha Ghosh, et al. 2021. Social science research to inform solar geoengineering: What are the benefits and drawbacks, and for whom? *Science* 374 (6569): 815–18.

Barrett, Scott. 1994. Self-enforcing international environmental agreements. *Oxford Economic Papers* 46:878–94.

———. 1999. A theory of full international cooperation. *Journal of Theoretical Politics* 11 (4): 519–41.

———. 2003. *Environment and statecraft: The strategy of environmental treaty-making*. Oxford: Oxford University Press.

———. 2007. *Why cooperate? The incentive to supply global public goods*. Oxford: Oxford University Press.

———. 2008. The incredible economics of geoengineering. *Environmental and Resource Economics* 39 (1): 45–54.

Biermann, Frank, Jeroen Oomen, Aarti Gupta, Saleem H. Ali, Ken Conca, Maarten A. Hajer, Prakash Kashwan, et al. 2022. Solar geoenegineering: The case for an international non-use agreement. *WIREs Climate Change* 13 (3): 754.

Cherry, Todd L., Stephan Kroll, David M. McEvoy, David Campoverde, and Juan B. Moreno-Cruz. 2022. Climate cooperation in the shadow of solar geoengineering: An experimental investigation of the moral hazard conjecture. *Environmental Politics* 32 (2): 362–70.

Crutzen, Paul J. 2006. Albedo enhancement by stratospheric sulfur injections: A contribution to resolve a policy dilemma? *Climatic Change* 77 (3–4): 211.

Finus, Michael. 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, Michael, and Matthew McGinty. 2019. The anti-paradox of cooperation: Diversity may pay! *Journal of Economic Behavior and Organization* 157:541–59.

Ghidoni, Riccardo, Anna L. Abatayo, Valentina Bosetti, Marco Casari, and Massimo Tavoni. 2023. Governing climate geoengineering: Side-payments are not enough. *Journal of the Association of Environmental and Resource Economists* 10 (5): 1149–77.

Harding, Anthony R., Katharine Ricke, Daniel Heyen, Douglas G. MacMartin, and Juan Moreno-Cruz. 2020. Climate econometric models indicate solar geoengineering would reduce inter-country income inequality. *Nature Communications* 11:227.

Heutel, Garth, Juan B. Moreno-Cruz, and Soheil Shayegh. 2018. Solar geoengineering, uncertainty, and the price of carbon. *Journal of Environmental Economics and Management* 87:24–41.

Heyen, Daniel, Joshua Horton, and Juan Moreno-Cruz. 2019. Strategic implications of counter-geoengineering: Clash or cooperation. *Journal of Environmental Economics and Management* 95:153–77.

IPCC (Intergovernmental Panel on Climate Change). 2022. Working Group III contribution to the sixth assessment report.

Keith, David W. 2000. Geoengineering the climate: History and prospect. *Annual Review of Energy and the Environment* 25 (1): 245–84.

———. 2021. Toward constructive disagreement about geoengineering. *Science* 374 (6569): 812–15.

Keith, David W., Edward Parson, and M. Granger Morgan. 2010. Research on global sun block needed now. *Nature* 463 (7280): 426–27.

McEvoy, David M., and John K. Stranlund. 2009. Self-enforcing international environmental agreements with costly monitoring for compliance. *Environmental and Resource Economics* 42 (4): 491–508.

McGinty, Matthew. 2007. International environmental agreements among asymmetric nations. *Oxford Economic Papers* 59 (1): 45–62.

Millard-Ball, Adam. 2012. The Tuvalu syndrome. *Climatic Change* 110 (3): 1047–66.

Moreno-Cruz, Juan B. 2015. Mitigation and the geoengineering threat. *Resource and Energy Economics* 41:248–63.

NRC (National Research Council). 2015. *Climate intervention: Reflecting sunlight to cool Earth*. Washington, DC: National Academies Press.

Reynolds, Jesse L. 2019. *The governance of solar geoengineering: Managing climate change in the Anthropocene*. Oxford: Oxford University Press.

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

Robock, Alan. 2008. 20 reasons why geoengineering may be a bad idea. *Bulletin of the Atomic Scientists* 64 (2): 14–18.

Smith, Wake. 2020. The cost of stratospheric aerosol injection through 2100. *Environmental Research Letters* 15 (11): 114004.

Smith, Wake, and Gernot Wagner. 2018. Stratospheric aerosol injection tactics and costs in the first 15 years of deployment. *Environmental Research Letters* 13 (12): 124001.

Wagner, Gernot. 2021. *Geoengineering: The gamble*. Cambridge: Polity Press.

Wagner, Gernot, and Martin Weitzman. 2012. Playing God. *Foreign Policy*, October 24.

Weitzman, Martin. 2015. A voting architecture for the governance of free-driver externalities. *Scandinavian Journal of Economics* 117:1049–68.

Wieners, Claudia E., Ben P. Hofbauer, Iris E. de Vries, Matthias Honegger, Daniele Visioni, Hermann W. J. Russchenberg, and Tyler Felgenhauer. 2023. Solar radiation modification is risky, but so is rejecting it: A call for balanced research. *Oxford Open Climate Change* 3 (1).

Williamson, Phillip, and Carol Turley. 2012. Ocean acidification in a geoengineering context. *Philosophical Transactions of the Royal Society A* 370:4317–42.