Contesting the Climate
Security Implications of Geoengineering[*]

Muhammet Bas[†]
Aseem Mahajan[‡]

Forthcoming, *Climatic Change*

Abstract

Scientists predict higher global temperatures over this century. While this may benefit some countries, most will face varying degrees of damage. This has motivated research on *solar geoengineering*, a technology that allows countries to unilaterally and temporarily lower global temperatures. To better understand the security implications of this technology, we develop a simple theory that incorporates solar geoengineering, *countergeoengineering* to reverse its effects, and the use of military force to prevent others from modifying temperatures. We find that when countries' temperature preferences diverge, applications of geoengineering and countergeoengineering can be highly wasteful due to deployment in opposite directions. Under certain conditions, countries may prefer military interventions over peaceful ones. Cooperation that avoids costs or waste of resources can emerge in repeated settings, but difficulties in monitoring or attributing interventions make such arrangements less attractive.

Keywords: geoengineering, counter-geoengineering, conflict, climate change

[*] We thank the Weatherhead Initiative on Climate Engineering and the Solar Geoengineering Research Program of Harvard University for support. We also thank Elena McLean, Torben Mideksa, Dustin Tingley and Gernot Wagner and the participants in Harvard University's Political Economy workshop for comments on earlier versions of this paper.

[†] Associate Professor of Political Science, New York University, Abu Dhabi Saadiyat Campus, mbas@nyu.edu

[‡] Corresponding author: PhD Candidate, Department of Government, Harvard University mahajan@g.harvard.edu

Introduction

A 2014 report by the Intergovernmental Panel on Climate Change (IPCC) predicts that unabated human greenhouse gas (GHG) emissions will increase global mean surface temperatures between 1.4°C and 4.8°C over the course of the century. These temperature changes and their impacts are distributed unevenly, as some countries will face more damages than others. In some cases, the temperature effects of climate change may even produce short-term regional benefits in the form of greater economic productivity, better weather, and higher crop yields (Burke, Hsiang, and Miguel 2015b; Egan and Mullin 2016). Changing temperatures may also confer geopolitical benefits. As the Arctic melts, proximate countries recognize that this could provide greater access to fishing grounds, oil and natural gas reserves, and the strategically important Northwest Passage commercial shipping route (Gautier et al. 2009; Yumashev et al. 2017).

How do countries' divergent temperature preferences affect their climate and security policies? The advent of *geoengineering* — defined as the "deliberate large-scale manipulation of the planetary environment to counteract anthropogenic climate change" (Shepherd 2009) — has expanded the necessary scope of climate governance beyond GHG mitigation, raising new questions. Recent work in political science and economics focuses on stratospheric aerosol scattering, commonly called *solar geoengineering*[1], which cools the earth by reflecting sunlight away from the planet. Because scientists estimate that solar geoengineering will produce immediate effects at a low cost, it is seen as a more technologically viable approach than other types of geoengineering, such as bioenergy combined with carbon capture and sequestration (BECCS) (Caldeira, Bala, and Cao 2013). Concerns that a country with extreme preferences will unilaterally geoengineer have raised speculation about a *countergeoengineering* response, wherein those with warmer ideal temperature points deploy atmospheric particles designed to warm the planet to counter applications of geoengineering.

This paper revisits both the security and climate implications of divergent temperature preferences in an environment where countries can unilaterally or jointly alter global temperatures. To do so, we develop a simple model in which two countries with divergent temperature preferences (vis-à-vis the observed status quo) choose whether to geoengineer or countergeoengineer, or engage in military conflict to establish unilateral control over manipulating the climate.[2] We show that Pareto-inferior equilibria can emerge in such cases that involve waste of resources due to deployment in opposite directions or costly conflict.[3] We then specify the conditions under which countries can sustain cooperative agreements that are Pareto-superior to such equilibria, which we define as deployment arrangements that

[1] Solar geoengineering actually encompasses a class of technologies, including stratospheric aerosol scattering, seeking to break links between GHG concentrations and temperatures (Harvard Solar Geoengineering Research Program 2020). In keeping with related social science research, we refer to stratospheric aerosol scattering using the metonym "solar geoengineering" in the remainder of the paper.

[2] We do not discuss normative issues about the deployment of solar geoengineering, countergeoengineering or about the use of conflict to restrict deployment. See Parker (2014) for ongoing ethical debates on the topic.

[3] In the remainder of the paper, we will use *deployment* as a general term for geoengineering or countergeoengineering applications.

can achieve the same or more attractive temperature outcomes, at lower costs, and hence strictly preferred by the two countries. More specifically, by considering the possibility of countergeoengineering and conflict, we develop three key insights.

First, if geoengineering and countergeoengineering are sufficiently cheap and if countries' ideal temperature points are sufficiently far from one another, equilibrium deployment may be highly wasteful in achieving the temperature outcome as countries deploy in opposite directions. Surprisingly, such Pareto inferior equilibria due to waste of resources may occur even when countries' ideal points fall on the same side of the status quo. Second, if waste of resources from deployment is sufficiently severe, countries may prefer costly conflict to establish unilateral control over implementation direction and levels. Third, in some infinitely repeated settings, cooperation may emerge that avoids wasteful deployments or conflict. We demonstrate that successful cooperation depends crucially on states' patience, as well as abilities to monitor, detect, and attribute deployment.

Beyond its contributions to existing literature in political science and economics, this paper engages with current research in international relations in two ways. First, it demonstrates the interaction between free-driving models (Weitzman 2015) and formal models of conflict. In doing so, it shows how countries' decisions about geoengineering, countergeoengineering, and engaging in conflict reflect a tradeoff between the inefficiencies generated from each choice. These inefficiencies resemble those arising in a Prisoners' Dilemma interaction, and are similar to those that occur in arming models (Bas and Coe 2016; Powell 1999).[4] Second, it contributes to a growing recent body of work in political science and economics that explores the relationship between climate change, weather shocks, and conflict (Bollfrass and Shaver 2015; Buhaug et al. 2014; Burke, Hsiang, and Miguel 2015a).

Background and Related Literature

Policymakers, journalists, and social scientists often use the term *solar geoengineering* to describe the deployment of sulfate aerosols or calcium carbonate to reflect sunlight away from the planet (Crutzen 2006; Harvard Solar Geoengineering Research Program 2020).[5] While humans have not tested or deployed the technology, scientists believe that such aerosols would quickly reduce global temperatures at substantially lower costs than GHG mitigation (Mahajan, Tingley, and Wagner 2019). Because the aerosols disperse, the effects would be transient, requiring periodic replenishment to maintain given temperatures. Furthermore, the effects of deployment are non-excludable — meaning that one country's deployment would affect others — and would vary by geography. Collectively, these properties have generated interest and raised concerns among journalists, governments, international organizations, and researchers in the natural and social sciences.[6]

Compared to solar geoengineering, there is limited research on countergeoengineering, which

[4] What differentiates geoengineering is that superior military or economic power is not required for deployment. Unilateral application by one of the many actors can have global or regional implications. As we discuss later in the paper, monitoring deployment is also likely to pose additional challenges.

[5] See Vaughan and Lenton (2011) for a review of different proposed methods of deployment.

[6] See Appendix A for further description of the properties and their implications.

seeks to negate or counteract the effects of solar geoengineering. Parker, J. B. Horton, and Keith (2018) categorize countergeoengineering as *neutralizing* or *countervailing*. Neutralizing countergeoengineering describes attempts to disable others' solar geoengineering deployments, by, for instance, adding bases into the atmosphere to produce potential salts with less radiative forcing (Keith et al. 2016). In contrast, countervailing methods entail using warming agents to actively heat the environment. Actors may, for instance, increase GHG emissions or release solid particles, coated with a thin metal layer, to reflect thermal infrared into the atmosphere without reflecting inbound solar radiation (Teller, Wood, and Hyde 1997). Like solar geoengineering, most countergeoengineering proposals are inexpensive and non-excludable, but their unpredictability, speed, and transience remain subject to debate. Some scholars speculate that countergeoengineering measures would be transient, though the duration countervailing deployment could vary with the lifetimes of GHGs (Parker and Keith 2015). Here, we focus on transient countergeoengineering with similar characteristics to solar geoengineering.

Related literature

Broadly, strategic models of solar geoengineering raise and seek to address two concerns.[7] First, by virtue of solar geoengineering's low cost and non-excludability, single states or coalitions of states with extreme temperature preferences or high risk tolerance may unilaterally over-deploy solar geoengineering (Schelling 1983, 1996). This *free-driving* produces uninternalized externalities in the form of excessively low temperatures, climate destabilization, and negative side-effects (Weitzman 2015). Second, solar geoengineering may produce "moral hazard problems" by displacing mitigation, as governments and companies avoid costly and time-consuming mitigation efforts in anticipation of a fast and inexpensive alternative (Broecker 1985; Keith 2000).

Research on the regulation and governance of solar geoengineering has generated a number cooperative geoengineering governance proposals (Lloyd and Oppenheimer 2014; Ricke, Moreno-Cruz, and Caldeira 2013; Weitzman 2015). Because such proposals lack incentives and mechanisms to compel or deter participation (Barrett 2014), their enforcement is unlikely in an international and anarchic environment. In line with other research in the field of international relations, actors behave as sovereign states and are thus unable to credibly delegate power to other parties (e.g., Meirowitz et al. 2019; Waltz 1959). With this in mind, we develop a model in which countries can only sustain self-enforcing agreements.

Our focus on the international system also limits the applicability of models primarily concerned with characterizing a single actor or social planner's optimal policy. A recent example of the former is Ahlvik and Iho (2018), which describes two countervailing forces affecting how much an actor initially deploys when experimenting with solar geoengineering.[8] Illustrative of the latter is Acemoglu and Rafey (2018), which has a benevolent social planner who

[7] Further discussion about the relationship between the technology's properties and challenges in governing it can be found in Appendix A. For a more extensive review, see Reynolds (2019).

[8] The "Inquisitive Effect" encourages high deployment levels to distinguish the effects of solar geoengineering from stochastic noise, while "Flexibility Effect" encourages low deployment that can be scaled up if solar geoengineering is effective and produces few side-effects.

can impose a carbon tax on firms to induce clean technology investment but cannot commit to it.[9] While both models lend useful insights, their central findings describe a single actor's optimal behavior rather than interactions between multiple actors. These differences are also reflected in our assumptions. Solar geoengineering's relevance in Acemoglu and Rafey (2018) stems from its effects on firm production and investment decisions, and so they treat it as an exogenous technological breakthrough. In contrast, we treat it as the endogenous product of interactions between countries.

Other models in which deployment is part of a self-enforcing equilibrium outcomes often pair geoengineering with GHG mitigation, focusing on the interaction between the two. Urpelainen (2012) argues that the threat of unilateral solar geoengineering may, counter-intuitively, encourage GHG mitigation to prevent rogue countries from engaging in it in the future,[10] Moreno-Cruz (2015) find that asymmetry in two countries' sensitivities to temperature and to geoengineering and mitigation costs may produce over-mitigation vis-à-vis the social optimum, as the more cost-sensitive country seeks to disincentivize the other from geoengineering. Manoussi and Xepapadeas (2017) reach similar conclusions in a dynamic setting.

Our model makes two notable departures from existing work. First, we incorporate the possibility of countergeoengineering. To date, only Heyen, J. Horton, and Moreno-Cruz (2019) and Parker, J. B. Horton, and Keith (2018) consider countergeoengineering in strategic settings.[11] Parker, J. B. Horton, and Keith (ibid.) argue that credible threats to countergeoengineer may provide a state with the ability to veto others' unilateral decision to solar geoengineer, while Heyen, J. Horton, and Moreno-Cruz (2019) find that countries with asymmetric preferences may either engage in wasteful deployment in opposite directions or commit to a moratorium treaty in which they abstain from climate interventions.

Unlike Heyen, J. Horton, and Moreno-Cruz and Parker, J. B. Horton, and Keith, we consider how imperfect monitoring of deployments affect prospects of international cooperation. Research on the governance of solar geoengineering after deployment is sparse (Reynolds 2019), and it is unclear how countries might behave when they are unable perfectly ascertain whether others have deployed. Regardless of whether actors' strategies under imperfect attribution scenarios are based on political considerations or technical expertise (MacMartin et al. 2019), attribution analyses will involve some degree of uncertainty due to false positives and false negatives. Our model contributes to existing work by describing how strategic actors may respond to such uncertainty. We also improve upon the approach used by Heyen, J. Horton, and Moreno-Cruz and Parker, J. B. Horton, and Keith and qualify their results by studying the interaction between geoengineering and countergeoengineering in a dynamic setting with the possibility of conflict. In line with conventional assumptions from

[9] In decreasing the marginal cost of emissions, solar geoengineering also decreases the credibility of future carbon taxes, prompting firms to under-invest in clean energy technology.

[10] Urpelainen builds on Millar-Ball (2012) — which considers non-strategic countries making a binary decision about whether to engage in a mitigation treaty — to a pair of strategic countries choosing how much to geoengineer and mitigate.

[11] Both Moreno-Cruz (2010) and Weitzman (2015) refer to countergeoengineering in passing, but neither considers how countergeoengineering would alter countries' strategic decision-making with respect to solar geoengineering.

international relations, we require cooperative agreements to be self-enforcing, thus better approximating the anarchic nature of the international system (Morelli 2009; Powell 2002). In doing so, we show that repeated interaction is a necessary but insufficient condition for cooperative treaties to emerge, and that reliable monitoring of behavior is a key component of such treaties. Moreover, as demonstrated below, economic and military power crucially shape states' incentives in these dynamic interactions.

The model outlined here considers continuous strategy spaces and flexible parameters for the costs and externalities from deployment and countries' ideal temperature points, which allow us to demonstrate the possibility of Pareto-inferior deployments by introducing countergeoengineering and the set of welfare-improving cooperative interventions. This extends Parker, J. B. Horton, and Keith (2018), whose sequential model has one state choose whether to geoengineer and another choose whether to countergeoengineer. In doing so, it provides insight on the magnitude of countries' deployments and costs. Weitzman (2015) offers some commentary on the strategic effects of countergeoengineering in formalizing the free-driver problem. Due to the assumption that solar geoengineering is costless, however, countergeoengineering in this model produces indeterminate outcomes. By incorporating a flexible cost parameter, our model formalizes Weitzman's observation about countergeoengineering but also characterizes a number of other equilibria with defined outcomes. The cost parameter also demonstrates how changes in the absolute and relative costs of geoengineering affect equilibrium outcomes.

The incorporation of countergeoengineering also suggests that the relationship between solar geoengineering and abatement may be more complex than that described by existing models. If the temperature effects of geoengineering become reversible, then countries sensitive to temperature changes may no longer be incentivized to abate in anticipation of future solar geoengineering, as suggested by Millar-Ball (2012) and Urpelainen (2012). Additionally, when countries are asymmetric, those that are more sensitive to mitigation costs or temperature changes may choose countergeoengineering as a less costly alternative to current or future mitigation, as suggested by Moreno-Cruz (2015).

As a further contribution to this literature, we incorporate conflict as a response to solar geoengineering. The possibility that countries may respond to solar geoengineering with military force has been explored by many scholars and journalists but we know of no formal models of solar geoengineering that accounts for the possibility of conflict (Gertner 2017; Victor et al. 2009). Like Urpelainen (2012) and Moreno-Cruz (2015), we consider only two actors, but the incorporation of conflict may address Barrett (2014)'s criticisms of multilateral governance arrangements that are not self-enforcing. It also changes the strategic incentives in models with self-enforcing arrangements and account for the possibility that conflict may sometimes prevent unilateral free-driver externalities. While federal agencies have started to explore the practical security consequences of solar geoengineering — the U.S. Central Intelligence Agency, for instance, funded a 2015 report on solar geoengineering by the National Academy of Sciences (National Research Council 2015) — to our knowledge no rationalist models of geoengineering have incorporated constraints that conflict imposes on deployment. As we show in the next section, the strategic impacts of conflict extend easily to a number of costly retaliatory policy responses that disable another country's capacity to geoengineer.

To construct a tractable model with countergeoengineering, we do not consider mitigation decisions, which are included in Millar-Ball (2012), Urpelainen (2012), Moreno-Cruz (2015) and Manoussi and Xepapadeas (2017). In this respect, our model is similar to Weitzman (2015)'s formalization of the free-driver effect. As opposed to other models that explain how solar geoengineering affects abatement, we focus on the relationship between conflict, solar geoengineering, countergeoengineering, and monitoring, which is increasingly relevant since humans already hold the technological ability to geoengineer.

A Model of Climate Tug-of-War

This section presents a simple model that shows how countries' divergent climate preferences could result in Pareto-inferior use of geoengineering and countergeoengineering, and increase the risk of conflict; and specifies conditions under which cooperation can emerge to provide Pareto-superior deployment alternatives. Consider two countries, A and B, who inhabit a climate with a normalized temperature of zero and with ideal temperature points τ_A and τ_B. The single round interaction involves countries simultaneously choosing g_i, for $i \in \{A, B\}$, the manner in which they manipulate the environment. Everything else being equal, engaging in geoengineering ($g_i < 0$) lowers the temperature, countergeoengineering ($g_i > 0$) increases it, and doing nothing ($g_i = 0$) leaves it unchanged. After deployment decisions, the new temperature is $\tau = g_A + g_B$. Each country's utility is described by the function

$$U_i = u_i(g_i, g_{-i}; k_i, s_i) = -(\tau_i - (g_i + g_{-i}))^2 - k_i(g_i)^2 - s_i(g_{-i})^2 \tag{1}$$

where $k_i > 0$ and $s_i > 0$, for $i, -i \in \{A, B\}$, are parameters that describe, respectively, the damages and costs from one's own deployment and that of others, respectively.[12] For each country, the first term $-(\tau_i - (g_i + g_{-i}))^2$ reflects its disutility from climate temperatures that diverge from its ideal point. As outlined in the introduction, costs from divergent temperatures may stem directly from lower crop yields, slower industry growth, and losses in economic productivity, and/or indirectly from resulting political or social unrest. The second term $-k_i(g_i)^2$ captures the additional direct or indirect, material or political costs and risks that a country bears from intervening. Despite the low cost of solar geoengineering relative to mitigation, experts estimate that fixed and annual costs will reach millions or billions of dollars, depending on how aerosols are deployed (McClellan et al. 2010). Actors excluded from the model, such as other countries or international organizations, may also impose costs on intervening countries in the form of retaliatory economic sanctions or conflict. Furthermore, if the decision to intervene alienates or polarizes the electorate, the intervening country may face further domestic costs from intervention. In addition, there could be environmental or climatic side effects from the country's intervention such as disruptions of precipitation patterns, which are captured by this cost parameter. Finally, the third term $-s_i(g_{-i})^2$ describes the costs and risks borne by a country from negative externalities

[12] These parameters make the model amenable to incorporating differences in actors' marginal costs and benefits and marginal rates of substitution.

that the other country's intervention imposes on it in the form of moral hazard problems, ozone depletion, disruptions in global hydrological cycles, and changes in temperature and precipitation patterns.[13]

We focus on the Nash Equilibria (NE) of this single round of interaction. Equilibrium strategies described in Proposition 1 imply that when the two countries' temperature preferences are sufficiently different, deployment patterns lead to a metaphorical "tug-of-war" between the two countries.[14] When the countries' ideal temperatures lie on opposite sides of the status quo, the equilibrium deployment levels will always be wasteful, canceling some of each other's efforts. Similar wastefulness may also occur when both agents prefer a cooler (or warmer) climate but their temperature preferences are sufficiently far apart:

Proposition 1 *In the single shot game without the possibility of conflict, the following strategy profile is the unique Nash equilibrium:*

$$g_A^{nc} = \frac{\tau_A(1+k_B) - \tau_B}{k_A k_B + k_A + k_B} \qquad g_B^{nc} = \frac{\tau_B(1+k_A) - \tau_A}{k_A k_B + k_A + k_B} \tag{2}$$

□

In equilibrium, the level of intervention depends on each country's costs and the difference between their ideal temperature points. When their ideal temperatures are identical but diverge from the status quo ($\tau_A = \tau_B \neq 0$), each country deploys in the same direction and in proportion to the other's per-unit cost of intervening. Under certain conditions, countries under-deploy in equilibrium, resulting in a net temperature that does not match their shared ideal point.[15] When both countries' ideal points fall on the same side of the status quo but differ from each other, the relative magnitude of their intervention and the extent of under-deployment depend on the difference in their preferences and their costs of intervention. Interestingly, when one country is more moderate than the other ($|\tau_i| < |\tau_{-i}|$, for $i, -i \in \{A, B\}$) and the more extreme country's costs are sufficiently low ($k_i < (\tau_i - \tau_{-i})/\tau_{-i}$), the moderate country may actually intervene in the opposite direction from its ideal point trying to shift the temperature toward the status quo. Furthermore, if countries' ideal points fall on opposite sides of the status quo, then countries always intervene in opposite directions. Finally, when the cost of deployment k_i, for $i \in \{A, B\}$, increases, state i's deployment in either direction decreases towards zero.

With the equilibrium deployment levels (g_A^{nc}, g_B^{nc}), the net temperature effect of interventions is $\tau^{nc} = g_A^{nc} + g_B^{nc}$, while the total magnitude of the intervention itself is $G^{nc} = |g_A^{nc}| + |g_B^{nc}|$. The net temperature effect and deployment levels will differ ($G^{nc} > |\tau^{nc}|$) when the countries counteract each other's interventions. Because each country wastes resources in trying to undo the other's intervention, any equilibria that involve interventions in opposite directions

[13] For presentational purposes, we assume that implementation of geoengineering and countergeoengineering are symmetrical, both in terms of deployment costs and overall side-effects. Allowing for differential costs and effects (through k_i and s_i) for each technology would not change the substantive conclusions as long as the remaining functional form assumptions are maintained.

[14] The proofs of the propositions are given in the online appendix.

[15] For instance, when $\tau_A = \tau_B = \bar{\tau} > 0$, and $k_A = k_B = \bar{k}$, both countries deploy $g_i = \frac{\bar{\tau}}{\bar{k}+2}$, resulting in a net temperature that is less than $\bar{\tau}$.

are Pareto-inferior, in the sense that there exist strategy profiles (e.g., $q' = (g_A^{nc} + g_B^{nc}, 0)$) that achieve the same temperature effect without wasting as many resources that are strictly preferred by both countries over the equilibrium strategy profile $q = (g_A^{nc}, g_B^{nc})$. In the non-wasteful alternatives, both states deploy less, resulting in lower costs from deployment given $k_i > 0$ and $s_i \geq 0$, and a strictly higher utility for each country.

How do these self-interested agents' equilibrium deployment levels compare to those of a socially optimal deployment regime? A social planner concerned with the collective welfare of both states would select g_A and g_B to solve the maximization problem $\max_{g_A, g_B}(U_A + U_B)$. In doing so, the social planner avoids Pareto-inferior deployments. Instead, she has both countries deploy in the same direction, regardless of where each of their ideal points is situated. The following proposition and corollary summarize the result:

Proposition 2 *To maximize the aggregate utility of states, an unbiased social planner selects the following levels of deployment:*

$$g_A^{sp} = \frac{(s_A + k_B)(\tau_A + \tau_B)}{(k_A + s_B + 2)(k_B + s_A + 2) - 4} \qquad g_B^{sp} = \frac{(s_B + k_A)(\tau_B + \tau_A)}{(k_A + s_B + 2)(k_B + s_A + 2) - 4} \tag{3}$$

□

The social planner's solution resembles implementation by a globally centralized government, or an international institution tasked with a compromise solution that takes into account countries' preferences, and is arguably unlikely to occur. Nonetheless, it provides an aspirational baseline. Unlike the solution described in Proposition 1, the social planner's deployment levels for each country also account for the negative externalities of implementation. In the appendix, we also provide deployment results for a biased social planner, which might represent an international organization that favor some countries more than others. Apart from the bias, the planner still avoids Pareto-inferior deployments, and the substantive comparisons to the self-interested deployment levels remain similar.

Possibility of Conflict

Could conflict be an alternative to Pareto-inferior deployment? We extend the model to include the possibility of conflict in a simple way. In this setting, A and B first sequentially choose whether to engage in conflict to disable the other's deployment capacities.[16] If neither country attacks, then they both play the geoengineering interaction described above by simultaneously choosing their deployment levels. For simplicity, we assume that states indifferent between attacking and not attacking choose peace, and that there are no first strike advantages. If either country chooses to attack, then conflict ensues. We model conflict as a costly lottery.[17] Country A wins with probability p, and country B with probability

[16] The choice of conflict to obtain unilateral control over deployment could be interpreted more broadly to include targeted strikes on deployment facilities, the imposition of sanctions on equipment needed to geoengineer, and information campaigns to mobilize domestic opposition to deployment.

[17] Modeling conflict in this manner as a costly lottery is standard in the international relations literature on bargaining and war. In line with this work, expected outcomes are a function of countries' balance of power and resolve, which are reflected in their probability of winning and costs of conflict. For examples, see Fearon (1995) and Bas and Coe (2012).

$1 - p$. Here, p captures the relative military strength of the two countries. Irrespective of who initiates the conflict, the costs of military conflict for countries A and B are c_A and c_B, respectively. The defeated country loses the ability to manipulate the temperature through deployment, whereas the winner can intervene unilaterally according to its temperature preferences. Let $g_i^W = \arg\max_{g_i} u_i(g_i, 0)$ represent country i's optimal intervention when $g_{-i} = 0$, for $i, -i \in \{A, B\}$, i.e., when the opponent cannot intervene. Then, each country's expected utility from military operations, denoted as W_i, equals

$$W_A = pu_A(g_A^W, 0) + (1-p)u_A(0, g_B^W) - c_A$$
$$W_B = pu_B(g_A^W, 0) + (1-p)u_B(0, g_B^W) - c_B$$

While conflict results in the fixed costs c_i that states incur from fighting, it avoids the potential waste of resources generated from peaceful deployment, since only one country can intervene after conflict. Thus, a state may prefer costly conflict to peaceful deployment when the latter entails both states deploying in opposite directions, resulting in waste of resources.[18] Proposition 3 summarizes the condition for peace in a Subgame Perfect Nash Equilibrium (SPNE):

Proposition 3 *In the unique SPNE of the conflict game, peace prevails and states deploy* (g_A^{nc}, g_B^{nc}) *if and only if* $u_A(g_A^{nc}, g_B^{nc}) \geq W_A$ *and* $u_B(g_A^{nc}, g_B^{nc}) \geq W_B$. *Otherwise, states engage in costly conflict.* □

In other words, in equilibrium, countries may engage in costly conflict to obtain unilateral control over the climate if doing so is less costly than the waste of resources resulting from peaceful deployment. One key assumption for this result is that costs of conflict are comparable, in scale, to direct and indirect costs from peaceful deployment. This is a reasonable assumption for three reasons. First, while the direct costs of deployment may appear negligible in comparison to military conflict when adjusted for duration, maintaining the same temperature over long periods of time would require continual deployment and thus accumulating costs, which may grow to become quite significant. Additionally, beyond its temperature effects, solar geoengineering may generate direct or indirect costs that are difficult to quantify (e.g., unpredictable weather patterns, acid rain, or ozone depletion). While the magnitude of such costs is uncertain, they must also be considered alongside direct deployment costs. Finally, a substantive body of research on the costs associated with climate change (Burke, Hsiang, and Miguel 2015b; Hsiang et al. 2017; Tol 2018) suggests that their magnitude is comparable to those of conflict, and, in countries most vulnerable to climate change, they may surpass military costs of conflict. If equilibria result in the "tug-of-war" behavior — wherein countries waste resources expended toward deployment yet fail to alter status-quo temperatures due to preference heterogeneity — unfavorable deployment may produce costs similar to those of climate change.[19]

[18] We model conflict only as it pertains to the issue of controlling the climate, so its winner only obtains unilateral control over deployment, nothing else. While we assume that the winner permanently prevents interventions by the opponent, our results extend to temporary controls. Finally, for simplicity of exposition, we do not model crisis bargaining, which can be captured by various cooperative equilibria we analyze in the next section.

[19] One may also argue that states would always seek to resolve Pareto-inferior deployments using peaceful

Repeated Interaction and Cooperation

We have shown above that outcomes in the form of under-deployment, wasteful deployment that cancel out states' efforts, or costly conflict could arise in equilibrium when countries disagree about the ideal temperature. Could repeated interaction between countries facilitate cooperative interventions that avoid such potentially Pareto-inferior outcomes? Suppose that the two countries share a common discount factor $\delta \in [0, 1)$ and play the game described above repeatedly in discrete time over an infinite number of periods. Below, we focus on the SPNE of this repeated interaction.

Per the standard result in such games, the repetition of stage game SPNE strategy profiles constitutes a SPNE in the repeated setting. We refer to this as the non-cooperative equilibrium of the repeated game and denote each state's utility from this equilibrium as U_i^s, for $i \in \{A, B\}$. As an alternative, in cooperative equilibria, we focus on cases where states achieve cooperation in SPNE by relying on grim-trigger strategies, in which both states agree upon a cooperative, non-wasteful, welfare-improving intervention strategy profile (g_A^c, g_B^c), and deploy these levels as long as neither country deviates from it. If either country deviates from the cooperative intervention, both revert indefinitely to the non-cooperative Nash equilibrium strategies, i.e., deploying g_A^{nc} and g_B^{nc} or engaging in conflict.

For cooperation to be sustained, the threat of reversion to the non-cooperative equilibrium must be deterrent, and more specifically, Pareto-inferior to the cooperative alternative being considered. As discussed above, one way this can happen is when states deploy in opposite directions in the stage game Nash equilibrium, resulting in waste of resources. Even when states deploy in the same direction, in some cases states may under-deploy, creating a possibility of welfare improvement through a cooperative arrangement. Finally, if the non-cooperative equilibrium has states engage in conflict, which is costly, Pareto superior alternatives may exist under certain conditions that avoid conflict and its associated costs. Proposition 4 establishes the conditions for such cooperative equilibria:

Proposition 4 *Assume that the non-cooperative SPNE of the game involves one of conflict, peaceful deployment in opposite directions, or under-deployment. Denote the per-period utility from this equilibrium as $U_i^s = u_i(g_A^{nc}, g_B^{nc})$. Consider a welfare-improving cooperative arrangement with deployment levels g_A^c and g_B^c, with each country's per-period utility from cooperation given as $U_i^c = u_i(g_A^c, g_B^c)$. Define U_i^d as the maximum one-period utility that i can achieve by deviating from the cooperative arrangement. When $\delta \geq (U_i^d - U_i^c)/(U_i^d - U_i^s)$, the following constitutes a SPNE of the game: at any period in the game, countries A and B do not attack and then deploy q_A^c and q_B^c, respectively, if neither country has attacked or deviated from (g_A^c, g_B^c) in the past. Otherwise, countries revert to the non-cooperative SPNE (either by attacking or deploying q_i^{nc}, depending on the nature of the Pareto-inferior non-cooperative equilibrium) for the remainder of the game.* □

alternatives to conflict, such as imposing economic sanctions or maintaining armament levels for deterrence. While such alternatives may seem less costly than conflict over comparable time-frames, research in international relations suggests that their costs may accumulate when they need to be adopted for long periods of time to maintain peace. See Coe (2019) for an analysis comparing the costs of containment and war prior to the Iraq War in 2003.

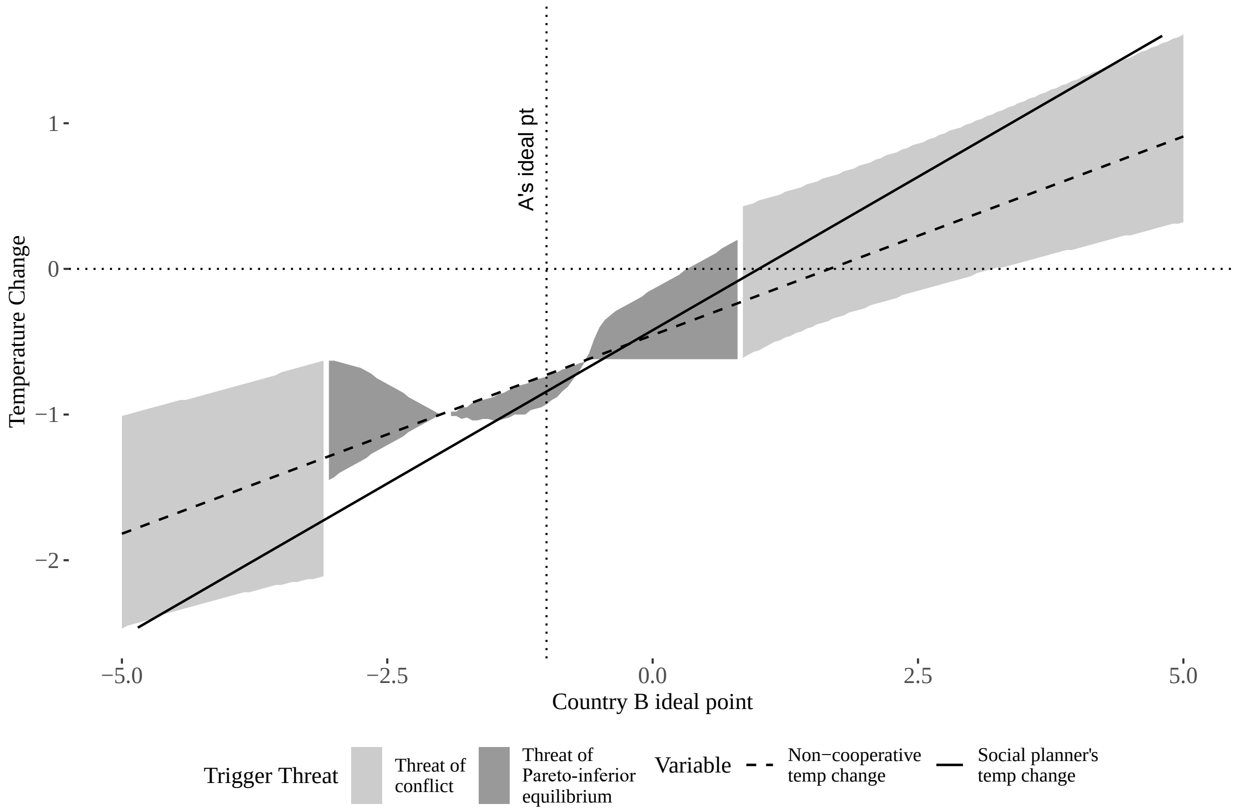
Thus, given that $U_i^d \geq U_i^c > U_i^s$, with repeated interaction, sufficiently patient states can sustain cooperation on Pareto-superior cooperative equilibria. Some observations are in order. First, when cooperation is possible, multiple intervention profiles (g_A^c, g_B^c) could satisfy its conditions. In many cases, multiple non-wasteful interventions can achieve the same temperature outcome in equilibrium as in the non-cooperative alternative. The existence of multiple cooperative intervention profiles also implies that a range of equilibrium temperatures, $\tau^c = g_A^c + g_B^c$, can be sustained with cooperation. When the non-cooperative equilibrium entails wasteful peaceful deployment, implying a unique non-cooperative equilibrium temperature, the most obvious set of alternative cooperative profiles are those that would sustain the same temperature but eliminate waste by having countries intervene in the same direction to achieve it. From a substantive viewpoint, international institutions and governments would likely find such cooperative equilibria, which generate immediate Pareto improvements resulting in the same temperature as the non-cooperative equilibrium, to be more viable than alternatives.

When there are cooperative equilibria that produce temperatures different from that produced by the non-cooperative equilibrium, it is also possible that for one of the countries, cooperative intervention may produce temperatures *further* from its ideal point than the non-cooperative intervention but may be still welfare-improving. One could imagine cases where the state with the strongest preference for the equilibrium temperature bears deployment costs to achieve its target in equilibrium; however, there are other less intuitive cost-sharing arrangements. For instance, for a target cooperative final temperature τ, cooperative arrangements may take the form $g_A^c = \alpha\tau$ and $g_B^c = (1-\alpha)\tau$, with $\alpha \in [0,1]$ representing the proportion of the total intervention allocated to country A under a given arrangement. While equilibrium selection in those cases is not the main focus of this paper, substantively, these cooperative arrangements would likely reflect countries' underlying bargaining power and existing institutional structures and biases.

Second, despite the attention devoted to the *least* wasteful cooperative equilibria — those in which countries do not counteract each other's interventions $(g_A^c + g_B^c = |g_A^c + g_B^c|)$ — there could also exist cooperative equilibria where states, if sufficiently patient, deploy in opposite directions, as long as the resulting wastefulness is less severe than those from the non-cooperative equilibrium. Finally, the set of possible interventions and the net temperatures from cooperative arrangements need not intersect with the social optimum. While cooperative equilibria produce Pareto gains over non-cooperative equilibria, socially optimal interventions are sometimes unsustainable, as they may contradict individual countries' interests, which lead them to defect at the expense of other cooperating countries.

As an example, Figure 1 illustrates the wide range of sustainable temperature changes at various ideal points for country B, while fixing country A's ideal temperature. In the figure, Country A's ideal temperature τ_A is -1, and its probability of victory, p, is 0.5. For A and B, the costs of deployment k_i are 0.6 and 1 respectively. The common discount factor δ is 0.9, and, for both countries, non-temperature externalities $s_i = 0$ and the costs of conflict $c_i = 1$. The example allows for different cost sharing arrangements in cooperative deals by varying α, where $\alpha \in [0,1]$ represents the proportion of the total intervention allocated to country A under a given arrangement. As described above, cooperation is sustained based on a threat

Figure 1. Temperature Changes Sustainable in Cooperative Equilibria for an Illustrative Set of Parameters.



of reversion to the non-cooperative equilibrium. The range of country B's ideal points for which the peaceful deployment equilibrium serves as a deterrent threat is represented by the dark shade of gray, spanning the center of the horizontal axis. The height of this region, describing the implied temperature range of sustainable deals, varies based on the magnitude of potential welfare gains from the non-cooperative equilibrium. When A and B's ideal points are close to each other, main improvement from the cooperative equilibria are due to underdeployment in equilibrium. As the ideal points diverge, states start deploying in opposite directions and counteracting each other's efforts. As wastefulness increases, the range of net temperature changes that can be sustained also expands. Given sufficient distance between the ideal points, the non-cooperative equilibrium becomes so wasteful that states prefer costly conflict to deployment in the non-cooperative equilibrium. Thus, the nature of the deterrent threat changes. Once their set of choices are limited to conflict and cooperative deals, the set of cooperative arrangements that countries can sustain expands, as represented by the lighter-shaded region. Greater divergence between ideal points in this region, however, does not expand the range of deals, since the cost of conflict remains constant.

Cooperation with Imperfect Monitoring

Compared to the one-period interaction, the repeated setting offers a promising possibility: if states are patient enough, they can cooperate to reduce inefficiencies stemming from conflict or peaceful deployment in opposite directions. This section presents an important caveat to this result. So far, we assumed that countries perfectly observe each other's deployment decisions. Consequently, they can condition their strategies on past behavior and can induce cooperation by levying the deterrent threat of switching to Pareto-inferior non-cooperative deployment levels or conflict.

However, some scholars raise the possibility that geoengineering deployment may not be perfectly observable, detectable, or attributable (Robock 2012; Svoboda and Irvine 2014). Even with strong scientific evidence establishing where deployment occurred, countries may be unable to determine whether it was sanctioned by the government or undertaken by a private or sub-state actor. Moreover, as with climate change, portions of the public may dismiss the validity of scientific evidence regarding solar geoengineering or disagree over what constitutes scientific consensus (Ding et al. 2011; Hornsey et al. 2016). Even policy makers who accept the scientific evidence themselves, then, may remain unsure over how the public would react to a costly response and the associated political costs of doing so, and this uncertainty may be compounded by foreign actors seeking to sway public opinion. Reaching domestic and international political consensus around attribution seems unlikely, and in light of uncertainty, states may use noisy indicators, such as the realized temperature or concentration of detected particles, to better assess others' past behavior. This noisy monitoring generates the possibility of false positives and negatives, as temperature and atmospheric concentrations may fluctuate exogenously due to other natural processes.

In this section, we revise the simple model described above to allow for imperfect public monitoring. As with perfect monitoring, repeatedly deploying at potentially Pareto-inferior stage game equilibrium levels is always an equilibrium strategy profile in the imperfect monitoring case as well, since such an equilibrium does not rely on monitoring past behavior. More pragmatically, we may ask: what are the conditions under which states can sustain cooperative equilibria that are Pareto-superior? How do the conditions for such equilibria compare to those described in Proposition 4, when there is noisy monitoring?

Given a pair of deployment levels, suppose that temperature follows a stationary i.i.d. stochastic process over time based on a symmetric, unimodal distribution $F(t \mid g_A, g_B)$, with mean $g_A + g_B$. In any given period, states first select their deployment levels. With z probability, Nature then reveals states' deployment levels as well as the realized temperature for that period, and with $1 - z$ probability, individual deployment levels remain private information to each state and Nature only reveals the temperature.[20]

Consider a cooperative pair of strategies, expected to result in an equilibrium temperature

[20] According to MacMartin et al. (2019), currently available technology limits the temperature effects of feasible deployment, making them indistinguishable from natural temperature fluctuations in the short run. Temperature changes attributable to deployment may thus require a window of multiple years before they are confidently detected. In our model, this would correspond to a situation with high levels of noise in temperatures, meaning that the variance of $T \sim F(t \mid g_A, g_B)$ is high.

t^c, that results in a Pareto improvement over the stage game equilibrium levels. If the countries can observe the past deployment levels, they can continue cooperation or revert to punishment as in the perfect monitoring repeated interaction case. Otherwise, if countries only observe the realized temperature from the previous period, they now need to account for the possibility that deviations from the expected temperature t^c may be due to cheating by one or both states, or to natural fluctuations in the temperature.

To sustain cooperation, we consider grim trigger strategies for comparability to the previous section. More specifically, we focus on states' use of threshold strategies, t_{lo} and t_{hi}, where $t^{lo} < t^c < t^{hi}$, when deployment levels remain private information and Nature only reveals temperature: if, in any such period, t falls outside the range (t^{lo}, t^{hi}), states take this as a signal that at least one state has deviated from the cooperative agreement. This triggers a reversion to the Pareto-inferior equilibrium.[21] Due to the stochastic nature of the temperature, states know that this reversion may occur when one or both states defect, or when both states abide by the cooperative agreement.[22] Thus, we find that, if states cannot always directly monitor each other's deployment, but instead form beliefs about one another's deployment based on imperfect signals, they find it more difficult to sustain cooperation due to the possibility of unnecessary punishment from false positives and lack of deserved punishment due to false negatives. Given sufficient noise in monitoring, cooperation becomes impossible. The derivations of the equilibrium under imperfect public monitoring are presented in the appendix, and the main result is summarized in the following proposition:

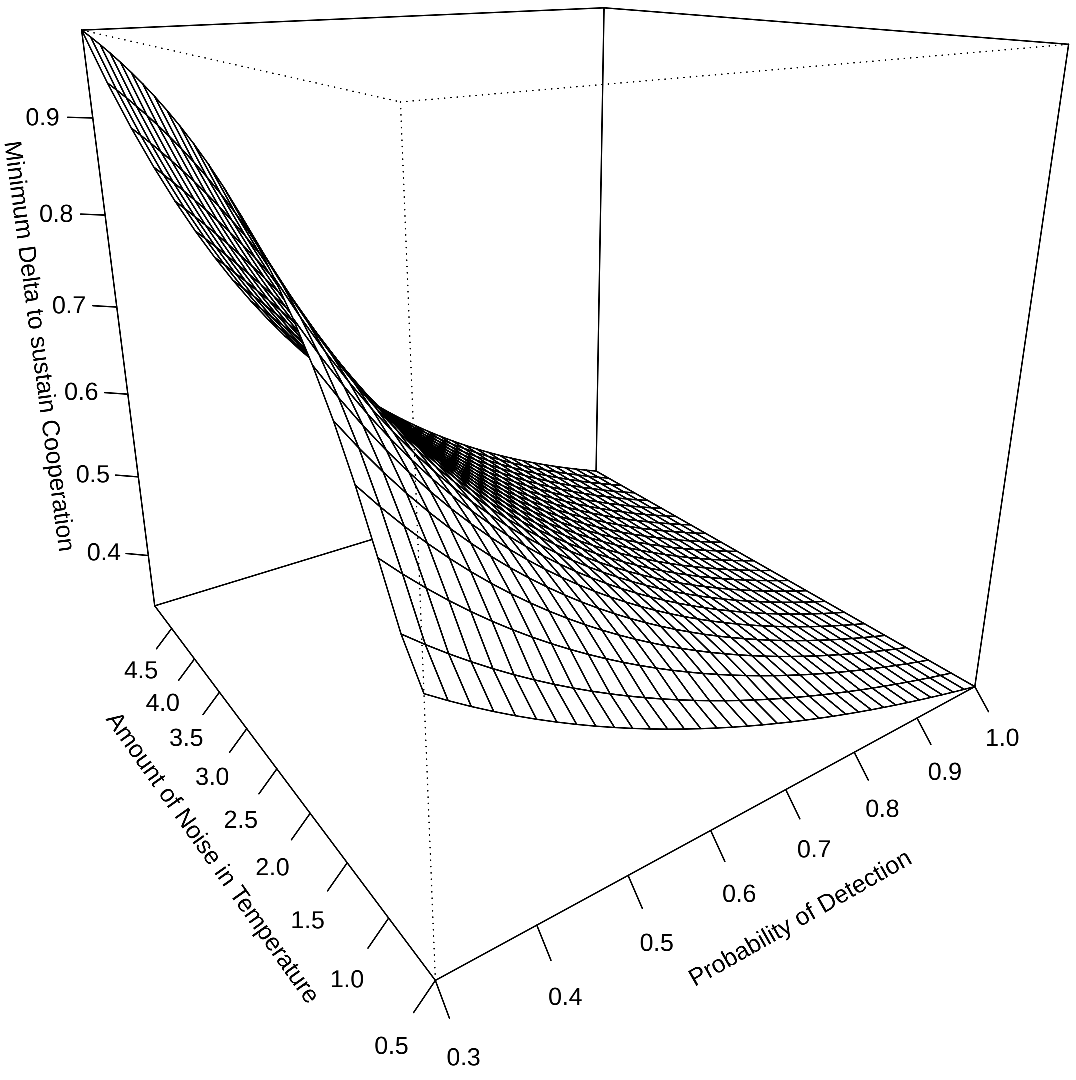
Proposition 5 *Assume that the non-cooperative equilibrium of the game is Pareto-inferior (due to conflict, under-deployment, or deployment in opposite directions). With sufficiently patient countries, if the noise in monitoring is low enough, the following strategies can constitute an equilibrium of the game: at any period t in the game, countries A and B do not attack and then deploy q_A^c and q_B^c, respectively, if neither country has attacked or been discovered to be cheating in the past, or in the absence of deployment information, temperatures have remained within the interval $(\overline{t}, \underline{t})$ in the previous period. Otherwise, states revert to the Pareto-inferior non-cooperative equilibrium deployment levels (either by attacking or deploying q_i^{nc}, for $i \in \{A, B\}$, depending on the nature of the non-cooperative equilibrium) for the remainder of the game.* □

To show how the conditions for cooperation compare to the perfect monitoring case, Figure 2 demonstrates the effect of the amount of imperfect monitoring on the likelihood of cooperation in a numerical example. In this example, A and B have ideal temperatures 1 and -1, respectively, and the unit cost of deployment is $k_i = 1$. The two states are equally powerful ($p = .5$) and the cost of conflict is $c_i = 4$. The noise in temperature, ϵ, follows a Normal distribution with mean 0 and standard deviation σ, and the realized temperature in a given

[21] In this section, we only consider equilibria in which states condition their behavior on temperatures from the previous period when there is no direct evidence of deployment. That being said, more complicated equilibria in which states make longer term observations of temperature trends over time to dynamically assess past defections from cooperation can also exist.

[22] Such conditional strategies based on simple cutpoints can trivially be a part of an equilibrium as the punishment itself is an equilibrium of the game.

Figure 2. Level of Patience Needed Sustain Cooperation with Imperfect Monitoring



period equals $g_A + g_B + \epsilon$. With these parametric assumptions, the non-cooperative equilibrium results in states deploying in opposite directions, canceling each others' efforts, and resulting in an expected temperature of 0 in each period. Thus, given the waste in the non-cooperative equilibrium, there exists a strategy profile that can achieve the same expected temperature 0 but without any deployment by either state, which should be preferred by both states to the non-cooperative alternative and can be the focus of the cooperative effort. In this candidate cooperative equilibrium, states use temperature thresholds (-1, 1) during periods without any deployment information.

In the figure, we vary z, the probability that Nature reveals A and B's deployment levels in a given period, as well as σ, the amount of noise in temperature in a given period. When $z = 1$, there is perfect monitoring, and the amount of noise in temperature is irrelevant. The minimum discount factor needed for cooperation in that case is $\delta = 1/3$. As detection becomes less likely (z decreases), or the temperature becomes more noisy (σ increases), however, the minimum discount factor needed for cooperation steadily increases. For low enough z and high enough σ, cooperation becomes impossible to sustain in equilibrium. Overall, the likelihood of cooperation decreases in the level of noise, but cooperation may nonetheless be possible in noisy environments if states are more likely to directly detect others' deployment levels.

Given technological limits on the extent to which states can currently change temperatures with geoengineering, the noise from naturally occurring temperature fluctuations remains significant (MacMartin et al. 2019). With temperatures providing limited information about deployment in the short run, cooperation would be less likely barring a sufficiently high probability of detecting deployment behavior directly. Technological developments to improve the efficiency of deployment may produce greater temperature shifts attributable to solar geoengineering. This would reduce the relative noise from natural variation in temperatures, making (imperfect) attribution more feasible on the basis of observed detection. Alongside possible improvements in direct detection of deployment, the prospects of sustaining cooperation would increase further in the future.

To sum up, imperfect monitoring and attribution erode both the benefits that countries obtain from cooperation and the costs from breaking agreements. As states recognize the possibilities of flouting cooperative agreements with impunity or facing reprisals despite adhering to them, sustaining cooperation becomes more challenging.

Conclusion

Climate warming over the coming century will produce global and unevenly dispersed economic and geopolitical impacts. While some countries, such as Russia, may benefit from warmer temperatures, most will face varying degrees of damage. New solar geoengineering and countergeoengineering technologies that allow countries to unilaterally change global temperatures suggest that these divergent preferences may have consequential impacts.

To better understand their security implications, we incorporated the possibility of conflict and countergeoengineering into a model of solar geoengineering and developed three main findings. First, when countries' temperature preferences diverge, even when they all prefer cooler temperatures, the possibility of countergeoengineering may produce a wasteful "tug-of-war" as different countries try to cool and warm the climate at the same time, canceling out the effects of each other's deployment of geoengineering or countergeoengineering. This finding suggests that the use of countergeoengineering may alter the dynamics of the free-driving effect often associated with solar geoengineering. With countergeoengineering, the inefficiency generated by free-driving may be compounded as countries counteract one another's deployment. Second, when there are significant differences in temperature preferences among countries, engaging in conflict to determine which country intervenes may be preferable to peacefully manipulating global temperatures.

Finally, welfare-improving cooperative intervention schemes that avoid waste or costs may emerge with sufficiently patient states, but these are not guaranteed, as countries face incentives to defect. In addition, imperfect monitoring and attribution of solar geoengineering would make cooperation more difficult. Because scientists have not yet tested the effects of the large-scale deployment of sulfate aerosols, it is currently unclear how accurately or precisely countries will be able to monitor or attribute its use. However, imperfect monitoring and attribution would decrease the likelihood of cooperation by eroding the benefits obtained from cooperation due to false positives, and increasing each country's incentives to defect from agreements and creating uncertainty about the efficacy of administering in costly punishment due to false negatives.

Author attribution

Both authors contributed equally to the writing of this manuscript and associated analysis. Authors' names are listed in alphabetical order.

References

Acemoglu, Daron and Will Rafey (2018). "Mirage on the Horizon: Geoengineering and Carbon Taxation Without Commitment". Working Paper.

Ahlvik, Lassi and Antti Iho (2018). "Optimal geoengineering experiments". In: *Journal of Environmental Economics and Management* 92, pp. 148–168.

Barrett, Scott (2014). "Solar Geoengineering's Brave New World: thoughts on the Governance of an Unprecedented Technology." In: *Review of Environmental Economics & Policy* 8.2.

Bas, Muhammet A and Andrew J. Coe (2012). "Arms Diffusion and War". In: *Journal of Conflict Resolution* 56.4, pp. 651–674.

– (2016). "A dynamic theory of nuclear proliferation and preventive war". In: *International Organization* 70.4, pp. 655–685.

Bollfrass, Alexander and Andrew Shaver (2015). "The effects of temperature on political violence: global evidence at the subnational level". In: *PLOS One* 10.5, e0123505.

Broecker, Wallace S (1985). *How to build a habitable planet*. Eldigio Press New York.

Buhaug, Halvard et al. (2014). "One effect to rule them all? A comment on climate and conflict". In: *Climatic Change* 127.3-4, pp. 391–397.

Burke, Marshall, Solomon M Hsiang, and Edward Miguel (2015a). "Climate and Conflict". In: *Annual Review of Economics* 7.1, pp. 577–617.

– (2015b). "Global non-linear effect of temperature on economic production". In: *Nature* 527.7577, pp. 235–239.

Caldeira, Ken, Govindasamy Bala, and Long Cao (2013). "The science of geoengineering". In: *Annual Review of Earth and Planetary Sciences* 41, pp. 231–256.

Coe, Andrew J. (2019). "Costly peace: A new rationalist explanation for war". Working paper available at https://eafdab1c-1639-49a6-b387-40b17853bd79.filesusr.com/ugd/c8f493_50287f9e40c843b7abcd24388f9f0c6a.pdf.

Crutzen, Paul J (2006). "Albedo enhancement by stratospheric sulfur injections: a contribution to resolve a policy dilemma?" In: *Climatic Change* 77.3, pp. 211–220.

Ding, Ding et al. (2011). "Support for climate policy and societal action are linked to perceptions about scientific agreement". In: *Nature Climate Change* 1.9, p. 462.

Egan, Patrick J and Megan Mullin (2016). "Recent improvement and projected worsening of weather in the United States". In: *Nature* 532.7599, pp. 357–360.

Fearon, James D (1995). "Rationalist explanations for war". In: *International Organization* 49.3, pp. 379–414.

Gautier, Donald L et al. (2009). "Assessment of undiscovered oil and gas in the Arctic". In: *Science* 324.5931, pp. 1175–1179.

Gertner, Jon (Apr. 2017). "Is It O.K. to tinker with the environment to fight climate change?" In: *The New York Times*. URL: https://www.nytimes.com/2017/04/18/magazine/is-it-ok-to-engineer-the-environment-to-fight-climate-change.html.

Harding, Anthony and Juan B Moreno-Cruz (2016). "Solar Geoengineering Economics: from Incredible to Inevitable and Half-way Back". In: *Earth's Future* 4.12, pp. 569–577.

Harvard Solar Geoengineering Research Program (2020). *Geoengineering*. URL: https://geoengineering.environment.harvard.edu/geoengineering (visited on 01/12/2020).

Heyen, Daniel, Joshua Horton, and Juan B Moreno-Cruz (2019). "Strategic implications of counter-geoengineering: clash or cooperation?" In: *Journal of Environmental Economics and Management* 95, pp. 153–177.

Hornsey, Matthew J et al. (2016). "Meta-analyses of the determinants and outcomes of belief in climate change". In: *Nature Climate Change* 6.6, p. 622.

Hsiang, Solomon M et al. (2017). "Estimating economic damage from climate change in the United States". In: *Science* 356.6345, pp. 1362–1369.

Keith, David W (2000). "Geoengineering the climate: history and prospects". In: *Annual Review of Energy and the Environment* 25.1, pp. 245–284.

Keith, David W et al. (2016). "Stratospheric solar geoengineering without ozone loss". In: *Proceedings of the National Academy of Sciences* 113.52, pp. 14910–14914.

Lloyd, Ian D and Michael Oppenheimer (2014). "On the design of an international governance framework for geoengineering". In: *Global Environmental Politics* 14.2, pp. 45–63.

MacMartin, Douglas G et al. (2019). "Technical characteristics of a solar geoengineering deployment and implications for governance". In: *Climate Policy* 19.10, pp. 1325–1339.

Mahajan, Aseem, Dustin Tingley, and Gernot Wagner (2019). "Fast, cheap, and imperfect? US public opinion about solar geoengineering". In: *Environmental Politics* 28.3, pp. 523–543.

Manoussi, Vassiliki and Anastasios Xepapadeas (2017). "Cooperation and competition in climate change policies: mitigation and climate engineering when countries are asymmetric". In: *Environmental & Resource Economics* 66.4, pp. 605–627.

McClellan, Justin et al. (2010). "Geoengineering cost analysis". Final report, Aurora Flight Sciences Corporation, Cambridge, Massachusetts.

Meirowitz, Adam et al. (2019). "Dispute resolution institutions and strategic militarization". In: *Journal of Political Economy* 127.1, pp. 378–418.

Millar-Ball, Adam (2012). "The Tuvalu syndrome". In: *Climatic Change* 110.3, pp. 1047–1066.

Morelli, Massimo (2009). "Institutional design and conflict: an introduction". In: *Review of Economic Design* 13.3, p. 167.

Moreno-Cruz, Juan B (2010). "Essays on the economics of geoengineering". PhD Dissertation.

– (2015). "Mitigation and the Geoengineering Threat". In: *Resource and Energy Economics* 41, pp. 248–263.

National Research Council (2015). *Climate Intervention: Reflecting Sunlight to Cool Earth.* National Academies Press.

Parker, Andy (2014). "Governing solar geoengineering research as it leaves the laboratory". In: *Phil. Trans. R. Soc. A* 372.2031, p. 20140173.

Parker, Andy, Joshua B Horton, and David W Keith (2018). "Stopping solar geoengineering through technical means: a preliminary assessment of counter-geoengineering". In: *Earth's Future.* URL: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018EF000864.

Parker, Andy and David W Keith (2015). "What's the right temperature for the earth". In: *Washington Post.*

Powell, Robert (1999). *In the shadow of power: states and strategies in international politics.* Princeton University Press.

Powell, Robert (2002). “Bargaining theory and international conflict”. In: *Annual Review of Political Science* 5.1, pp. 1–30.

Reynolds, Jesse L (2019). “Solar geoengineering to reduce climate change: a review of governance proposals”. In: *Proceedings of the Royal Society A* 475.2229, p. 20190255.

Ricke, Katharine L, Juan B Moreno-Cruz, and Ken Caldeira (2013). “Strategic incentives for climate geoengineering coalitions to exclude broad participation”. In: *Environmental Research Letters* 8.1, p. 014021.

Robock, Alan (2012). “Will geoengineering with solar radiation management ever be used?” In: *Ethics, Policy & Environment* 15.2, pp. 202–205.

Schelling, Thomas C. (1983). “Climatic Change: Implications for Welfare and Policy”. In: *Changing climate: Report of the carbon dioxide assessment committee.* Ed. by National Research Council, Carbon Dioxide Assessment Committee, et al. National Academies, pp. 449–497.

– (1996). “The economic diplomacy of geoengineering”. In: *Climatic Change* 33.3, pp. 303–307.

Shepherd, John G (2009). *Geoengineering the climate: science, governance and uncertainty.* The Royal Society.

Svoboda, Toby and Peter Irvine (2014). “Ethical and technical challenges in compensating for harm due to solar radiation management geoengineering”. In: *Ethics, Policy & Environment* 17.2, pp. 157–174.

Teller, Edward, Lowell Wood, and Roderick Hyde (1997). *Global warming and ice ages: prospects for physics based modulation of global change.* Tech. rep. Lawrence Livermore National Lab., CA (United States).

Tol, Richard SJ (2018). “The economic impacts of climate change”. In: *Review of Environmental Economics and Policy* 12.1, pp. 4–25.

Urpelainen, Johannes (2012). “Geoengineering and global warming: a strategic perspective”. In: *International Environmental Agreements: Politics, Law and Economics* 12.4, pp. 375–389.

Vaughan, Naomi E and Timothy M Lenton (2011). “A review of climate geoengineering proposals”. In: *Climatic Change* 109.3-4, pp. 745–790.

Victor, David G et al. (2009). “The geoengineering option: a last resort against global warming?” In: *Foreign Affairs*, pp. 64–76.

Waltz, Kenneth (1959). *Man, the State, and War: a Theoretical Analysis.* New York: Columbia University Press.

Weitzman, Martin L (2015). “A voting architecture for the governance of free-driving externalities with application to geoengineering”. In: *The Scandinavian Journal of Economics* 117.4, pp. 1049–1068.

Williamson, Phillip and Carol Turley (2012). “Ocean Acidification in a Geoengineering Context”. In: *Philosophical Transactions of the Royal Society* 370.1974, pp. 4317–4342.

Yumashev, Dmitry et al. (2017). “Towards a balanced view of Arctic shipping: estimating economic impacts of emissions from increased traffic on the Northern Sea Route”. In: *Climatic Change*, pp. 1–13.

Appendix A Properties of Stratospheric Aerosol Scattering

Throughout this section, we follow other related social science research in using the term "solar geoengineering" as a metonym for stratospheric aerosol scattering.

The assumptions in most rationalist models of solar geoengineering stem from five key properties: its speed, low-cost, non-excludability, unpredictability, and transience.

Solar geoengineering is inexpensive and has instantaneous and non-excludable effects that vary geographically. Additionally, aerosols are more transient than most GHGs, particularly carbon dioxide, and thus they must be replenished regularly to maintain a given temperature. Similarly, most countergeoengineering proposals are inexpensive and non-excludable, but their unpredictability, speed, and transience remain subject to debate. Some scholars speculate that countergeoengineering responses would be transient, though there is high variability in the lifetime of GHGs (Parker and Keith 2015). In this paper, we have limited our model to temporary countergeoengineering with characteristics similar to those of solar geoengineering.

Solar geoengineering is non-excludable, thereby generating externalities that may lead to over- or under-deployment relative to the socially optimal levels. The technology's possibly heterogeneous side-effects, low cost, and transience heighten the challenge of incentivizing efficient deployment. Moreover, the regulatory challenges posed by its low direct cost and non-excludability and by heterogeneous temperature preferences generate free-driving. Free-driving occurs when countries with extreme temperature preferences or high risk tolerance determine global solar geoengineering deployment (Schelling 1996; Weitzman 2015). If solar geoengineering is sufficiently cheap, actors with extreme preferences for low temperatures or high tolerance for associated risks may unilaterally over-deploy it, generating uninternalized free-driving externalities in the form of excessively low temperatures, climate destabilization, and negative side-effects.

Moral hazard occurs when governments and companies reduce costly and time-consuming mitigation efforts in anticipation of solar geoengineering, which offers a fast and inexpensive way to reduce temperatures (Keith 2000). This behavior exacerbates the traditional global public-goods problem associated with climate change. If solar geoengineering and mitigation were perfect substitutes, then moral hazard would not be problematic, but this is not the case. Sulfate aerosols decrease global temperatures, but do not prevent ocean acidification (Williamson and Turley 2012). By substituting solar geoengineering for mitigation, actors sensitive to mitigation costs and high temperatures — agricultural producers, for instance — can impose externalities on those vulnerable to ocean acidification, such as fish farmers. The interaction between moral hazard and the possibly heterogeneous effects of solar geoengineering (Harding and Moreno-Cruz 2016) further undercut the argument that solar geoengineering is a perfect substitute for mitigation.

Appendix B Formalized propositions and proofs

Proof of Proposition 1

Proof The first order condition for each country is

$$2\tau_i - 2(g_i + g_{-i}) - 2k_i g_i = 0$$

which reduces to

$$g_i(g_{-i}) = \frac{\tau_i - g_{-i}}{k_i + 1}. \tag{4}$$

The second order condition $-2 - 2k_i < 0$ indicates that the above is the best response to the opponent's deployment. To find the equilibrium deployment for each country, we solve the linear system of equations

$$\begin{bmatrix} k_A + 1 & 1 \\ 1 & k_B + 1 \end{bmatrix} \begin{bmatrix} g_A \\ g_B \end{bmatrix} = \begin{bmatrix} \tau_A \\ \tau_B \end{bmatrix} \tag{5}$$

which has the unique solution (g_A^{nc}, g_A^{nc}) given in the proposition for any $k_i > 0$. ■

B.1 Proof of Proposition 2

Proof The maximization problem facing an unbiased social planner concerned with total welfare is $\max_{g_A,g_B}(U_A+U_B)$ where

$$U_A + U_B = U^P = -\Big(\sum_{i\in\{a,b\}} (\tau_i - g_i - g_{-i})^2 + k_i(g_i)^2 + s_i(g_{-i})^2\Big) \tag{6}$$

The planner's selection of deployments for each country must then fulfill the first order conditions

$$\frac{\partial U^p}{\partial g_i} = 2(\tau_i - g_i - g_{-i}) + 2(\tau_{-i} - g_i - g_{-i}) - 2k_i g_i - 2s_{-i} g_i = 0 \tag{7}$$

which can be written in matrix form as

$$\begin{bmatrix} k_A + s_B + 2 & 2 \\ 2 & k_B + s_A + 2 \end{bmatrix} \begin{bmatrix} g_A \\ g_B \end{bmatrix} = \begin{bmatrix} \tau_A + \tau_B \\ \tau_A + \tau_B \end{bmatrix} \tag{8}$$

For any $k_i > 0$ and $s_i \geq 0$, this system has a unique solution given by

$$\begin{bmatrix} g_A^{SP} \\ g_B^{SP} \end{bmatrix} = \frac{1}{(k_A + s_B + 2)(k_B + s_A + 2) - 4} \begin{bmatrix} k_B + s_A + 2 & -2 \\ -2 & k_A + s_B + 2 \end{bmatrix} \begin{bmatrix} \tau_A + \tau_B \\ \tau_A + \tau_B \end{bmatrix} \tag{9}$$

which gives the expression in the proposition. Checking the second order condition at the solution,

$$H = \begin{bmatrix} -4 - 2(k_A + s_B) & -4 \\ -4 & -4 - 2(k_B + s_A) \end{bmatrix} \tag{10}$$

is negative definite for any $k_i > 0; s_i \geq 0$. ■

Let $r \in [0,1]$ capture the potential bias of the planner, where $r = 1/2$ corresponds to the impartial planner; $r > 1/2$ is a social planner in favor of State A, and $r < 1/2$ represents a planner biased towards State B.

Corollary 1 *To maximize weighted aggregate utility of states, the potentially biased social planner selects the following levels of deployment, where $r \in [0,1]$ reflects the amount of bias the planner has in favor of A:*

$$\begin{aligned} g_A^{bsp} &= \frac{(rs_A + (1-r)k_B)(r\tau_A + (1-r)\tau_B)}{(rk_A + (1-r)s_B + 1)((1-r)k_B + rs_A + 1) - 1} \\ g_B^{bsp} &= \frac{((1-r)s_B + rk_A)((1-r)\tau_B + r\tau_A)}{(rk_A + (1-r)s_B + 1)((1-r)k_B + rs_A + 1) - 1} \end{aligned}$$

The biased social planner's solution can be derived similarly to the unbiased case.

B.2 Proof of Proposition 3

PROOF First, suppose peace prevails in a Subgame Perfect Nash equilibrium. Then, it must be the case that neither country attacks at its initial decision node. Then, it must also be that each country i simultaneously deploys g_i^{nc} in accordance with Proposition 1. By the definition of the Nash equilibrium, country i can make no unilateral deviation that would produce a payoff greater than $u_i(g_i^{nc}, g_{-i}^{nc})$. Then, it must be true that $u_B(g_A^{nc}, g_B^{nc}) \geq W_B$ and $u_A(g_A^{nc}, g_B^{nc}) \geq W_A$, which is the condition specified in the proposition. Now, for the second part, suppose that the condition in Proposition 1 is met. As given in Proposition 1, the strategy profile (g_A^{nc}, g_B^{nc}) is the unique Nash equilibrium of the deployment subgame. Next, back up one step to the point when B decides whether to attack. Because the condition in Proposition 1 is met, $u_B(g_A^{nc}, g_B^{nc}) \geq W_B$, so B does not attack.[23] Using a similar argument, one can show that A chooses not to attack, and peace prevails in equilibrium. ■

[23] Per our earlier assumption, states choose peace when indifferent between attacking and not attacking.

B.3 Proof of Proposition 4

Proof With grim-trigger strategies, any deviation from the proposed cooperative equilibrium is punished by a permanent reversion to the Pareto-inferior SPNE. Thus, sustaining cooperation in a SPNE requires that any single-round deviations are not profitable for either state, or $U_i^c/(1-\delta_i) \geq U_i^d + \delta U_i^s/(1-\delta)$. This reduces to

$$\delta \geq \frac{U_i^d - U_i^c}{U_i^d - U_i^s} \tag{11}$$

as given in the Proposition. For the second part of the proof, the credibility of the punishment for deviations is established due to the nature of grim trigger strategies and the fact that the proposed punishment itself is a SPNE. The right hand side of the above condition (11) is less than 1 as long as the cooperative deployment provides a Pareto improvement over the non-cooperative equilibrium levels (e.g. when the non-cooperative equilibrium involves deployment in opposite directions.) ■

B.4 Proof of Proposition 5

Proof In any given period t, the final temperature τ_i is a function of g_A and g_B, but now also has a stochastic component. We write its probability density function as $f(\tau \,|\, g_A, g_B)$, with $f(\cdot)$ taken to be continuous with full support and mean $g_A + g_B$. When Nature does not reveal deployment levels at the end of a period (with probability $1 - z$), the probability that punishment is triggered in a given period is $F(\tau \leq \underline{t} \,|\, g_A, g_B) + F(\tau \geq \bar{t} \,|g_A, g_B)$. For notational convenience, define

$$F^C := F(\tau \leq \underline{t} \,|\, g_A^c, g_B^c) + F(\tau \geq \bar{t} \,|\, g_A^c, g_B^c) \tag{12}$$

$$F_i^D := F(\tau \leq \underline{t} \,|\, g_i^D, g_{-i}^c) + F(\tau \geq \bar{t} \,|\, g_i^D, g_{-i}^c) \tag{13}$$

where F^C is the probability that a punishment is triggered when both countries have complied with the cooperative arrangement and Nature did not reveal the deployment levels but the temperature, and likewise F_i^D is the probability that a punishment is triggered when country i defects to g_i^D in a given period from the cooperative arrangement but the deployment levels remain private. When $F^C = 0$ and $F_i^D = 1$, the information structure reduces to that of the perfect monitoring case.

For each country i, for $i \in \{A, B\}$, we first establish the payoff from cooperation, which includes a risk of punishment due to false positives. The punishment for deviating is a reversion to the non-cooperative equilibrium, so the expected payoff from cooperation is $E_i = U_i^C + \delta(1-z)F^C U_i^s/(1-\delta) + \delta(1-(1-z)F^C)E_i$ or,

$$E_i = \frac{U_i^C + \frac{\delta(1-z)F^C U_i^S}{1-\delta}}{1 - \delta(1-(1-z)F^C)} \tag{14}$$

where U_i^C and U_i^S are assessed in expectation due to the stochastic temperature. For cooperation to be sustained, one-period deviations should not be profitable. If i deviates, this can be punished in two ways: Nature reveals the deviation with z probability, or if the deployment level remains private, the revealed temperature exceeds the allowed zone. Thus, the total likelihood of punishment in the next period after a deviation equals $(z + (1-z)F_i^D)$. Expected payoff from deviation also includes a possibility of false negatives, i.e. the chance that Nature does not reveal deployment levels and the temperature stays within the allowed range despite deviation, which occurs with $(1-z)(1-F_i^D)$ probability. Given these, for cooperation to be successful, it should be preferred to the most attractive deviation for i. Thus,

$$\begin{aligned} E_i &\geq \max_{g_i^D} U_i^d + \delta(z + (1-z)F_i^D)U_i^s/(1-\delta) + \delta(1-z)(1-F_i^D)E_i \\ &\geq U_i^{d*} + \delta(z + (1-z)F_i^{D*})U_i^s/(1-\delta) + \delta(1-z)(1-F_i^{D*})E_i \end{aligned} \tag{15}$$

which is equivalent to:

$$E_i \geq \frac{U_i^{d*} + \frac{\delta(z+(1-z)F_i^{D*})U_i^s}{1-\delta}}{1 - \delta(1-z)(1-F_i^{D*})} \tag{16}$$

Substituting for E_i from (14) into (16)

$$\frac{U_i^C + \frac{\delta(1-z)F^C U_i^S}{1-\delta}}{1 - \delta(1-(1-z)F^C)} \geq \frac{U_i^{d*} + \frac{\delta(z+(1-z)F_i^{D*})U_i^s}{1-\delta}}{1 - \delta(1-z)(1-F_i^{D*})} \tag{17}$$

When $F^C = 0$ and $F_i^D = 1$, or when $z = 1$, this inequality reduces to the condition given in (11) for the cases with perfect monitoring. Moreover, as z decreases and the probability of false positives F^C increases, the discount factor δ required to sustain inequality (17) eventually exceeds 1, which eliminates the possibility of cooperation.[24] ■

[24] For the numerical example presented in the manuscript, the condition (17) becomes:

$$\frac{-1 - \sigma^2 + \frac{\delta(1-z)F^C(-2-\sigma^2)}{1-\delta}}{1 - \delta(1-(1-z)F^C)} \geq \frac{-(1-g_i^{D*})^2 - \sigma^2 + \frac{\delta(z+(1-z)F_i^{D*})(-2-\sigma^2)}{1-\delta}}{1 - \delta(1-z)(1-F_i^{D*})}$$

where $F^C = 1 + \Phi(\frac{\underline{t}}{\sigma}) - \Phi(\frac{\bar{t}}{\sigma})$ and $F_i^{D*} = 1 + \Phi(\frac{\underline{t}-g_i^{D*}}{\sigma}) - \Phi(\frac{\bar{t}-g_i^{D*}}{\sigma})$ and Φ is the Standard Normal CDF.