A Voting Architecture for the Governance of Free-Driver Externalities, with Application to Geoengineering

Martin L. Weitzman*

November 28, 2013
comments appreciated

Abstract

Climate change is a global "free rider" problem because significant abatement of greenhouse gases is an expensive public good requiring international cooperation to apportion compliance among states. But it is also a global "free driver" problem because geoengineering the stratosphere with reflective particles to block incoming solar radiation is so cheap that it could essentially be undertaken unilaterally by one state perceiving itself to be in peril. This exploratory paper develops the main features of a "free driver" externality in a simple model motivated by the asymmetric consequences of type-I and type-II errors. I propose a social-choice decision architecture embodying the solution concept of a supermajority voting rule and derive its basic properties. In the model this supermajority voting rule attains the socially optimal cooperative solution, which is a new theoretical result around which the paper is built.

1 Introduction via Climate Change

Perhaps the most striking aspect of the economics of climate change is the enormity of the international public-goods problem that it presents. Overcoming the free rider problem on a global externality of such immense scope represents a world governance challenge of

*Department of Economics, Harvard University (mweitzman@harvard.edu). Without tying them to the contents of this paper or implying that they neccesarily agree with it, I am grateful for useful critical comments to Scott Barrett, Richard Cooper, Vincent Crawford, Jerry Green, David Keith, Robert Keohane, Eric Maskin, Antony Millner, Hervé Moulin, Wilfried Rickels, Thomas Schelling, Robert Stavins, Thomas Sterner, and Gernot Wagner.

unprecedented proportions. Not infrequently, one encounters statements in the literature such as "climate change is the biggest market failure the world has ever seen" or "climate change is the mother of all externalities" or the like.

This paper begins with the realization that there are really *two different* externalities involved in the climate change problem, that they have near-opposite properties, that they interact, and that it seems difficult to say offhand which one is more threatening than the other. The first externality, described by the above quotes, comes in the usual familiar form of a public goods problem whose challenge is enormous because so much is at stake and it is so difficult to reach an international governing agreement that divides up the relatively *expensive* sacrifices that would be required by each nation to really make much of a dent in greenhouse gas (GHG) concentrations. The classic governance problem here is to limit the under-provision of a public good from free riding.

A second less-familiar externality shows up in the scary form of geoengineering the stratosphere with reflective particles to block incoming solar radiation. This geoengineering-type externality is so relatively *cheap* to enact that it might in principle effectively be undertaken unilaterally by one nation feeling itself under climate siege, to the detriment of other nations. The challenge with this second global externality also appears to be enormous, because here too so much is at stake and it also seems difficult to reach an international governing agreement. If the first externality founders on the "free rider" problem of under-provision, then the second externality founders on what might be called the "free driver" problem of over-provision. If the first externality is the "mother of all externalities," then the second externality might be called the "father of all externalities." These two powerful externalities appear to be almost polar opposites, with the world forced to confront both.

This paper concentrates on the second or free-driver externality. The next section describes in an extremely compressed form some of the most salient features of geoengineering that are relevant for motivating the abstract model of this paper. Among the questions needing to be addressed are the following. Who is allowed to do geoengineering? Under what circumstances? What is the relevant solution concept? Is there any recognizable decision mechanism, however hypothetical, abstract, and seemingly unrealistic, that theory suggests? And might this theory actually form the backbone of a governance architecture?

The paper treats geoengineering as one particular motivating example from a more general family of public-good-like externalities, whose generic properties are the main subject of investigation. This abstraction of geoengineering is called a "free driver" externality for reasons that will become apparent. Governance is the key issue for a free-driver externality. For example, geoengineering without a proper governance architecture could become a major global threat with the potential to cause serious international frictions and even outright

conflicts. Designing a social-choice architecture to deal with this free-driver governance dilemma is the central theme of the paper. Alas, while the problem is important, it is also difficult to model. I am forced to beg the reader's indulgence for an analysis that is only partial (and therefore criticizable) in favor of providing some new insights on an important subject.

I present the simplest formal analytical representation of a free-driver externality that I can imagine. The model is based on a very crude analogy to the asymmetric consequences of type-I and type-II errors as manifested in a kinked loss function with different right-side and left-side slopes. The socially optimal solution is derived. I attempt constructively to sketch the theoretical outlines of a possible governance architecture for dealing with a free-driver externality. The point of departure is the insight that a free-driver externality does not confront the thorny issue of assigning compliance costs, which hobbles resolution of a free-rider externality. I show that a free-driver externality may perhaps be more amenable to a reasonable resolution than a free-rider externality because some of its worst features might be ameliorated by a relatively simple voting mechanism that does not involve transfer payments.

In this paper I propose a social-choice decision architecture based on a supermajority[1] voting rule for the free-driver problem and I examine its basic properties. In the model this supermajority voting rule attains the socially optimal cooperative solution, which is a new theoretical result around which the paper is built. To be sure, this proposed solution concept is presented and analyzed here only under very strong assumptions and at such a high level of abstraction that it might seem remote from geoengineering. The paper is frankly exploratory and intended to be thought provoking. Nevertheless, my hope is that the derived supermajority voting rule might serve as a template for a future governance architecture that is at least worth thinking about and discussing.

2 Geoengineering as a Free-Driver Externality

I now want to describe very briefly some aspects of the spectre of geoengineering that are relevant to this paper.[2] Suppose, for the sake of argument, that the world is unable to rise

[1] In the name of convenience I abuse terminology throughout this paper because the "supermajority" rule is not applied symmetrically, but refers to one direction only (typically "upward changes" whereas "downward changes" are made by minority rule).

[2] There is a sizable literature on this subject, which is readily available on the internet by searching the word "geoengineering." In particular, Wikipedia provides a decent summary of the main issues with an extensive bibliography for further reference. See also U.K. Royal Society (2009), U.S. Academy of Sciences (2010), and Bipartisan Policy Center (2011). While emphasizing his own viewpoint, the recent book by David Keith (2013) covers much of the latest thinking on geoengineering.

to the free-rider global public-good challenge represented by excessive GHG emissions and that we continue more or less along the same lines of business as usual. Suppose, further for the sake of argument, that some kind of a tipping event like massive methane or carbon dioxide releases with strong bad feedbacks begins in earnest a half-century or so from now. In this low-probability science fiction story we might then become very scared that we were riding along a trajectory leading to a climate disaster. A high-temperature trajectory might be accompanied by the threat of a rapid rise of sea level, altered oceanic and atmospheric circulation patterns, harmful regional weather changes, and so forth. There could well be other nasty tipping-point surprises, some of which are "unknown unknowns" in the form of events that we cannot now even imagine. What might we then do? In the face of rapidly rising temperatures some might be tempted to try to deliberately geoengineer the planet as a quick fix, which would be sufficient to restore temperatures to safer levels at least temporarily while we try, this time hopefully seriously, to cut back drastically on greenhouse gas emissions and to undertake other, more permanent if much more slower-acting, measures.

A U.S. National Academy of Sciences (2010) study defined geoengineering as "options that would involve large-scale engineering of our environment in order to combat or counteract the effects of changes in atmospheric chemistry." Similarly, a study of the U.K. Royal Society (2009) defined geoengineering as "the deliberate large-scale manipulation of the planetary environment to counteract anthropogenic climate change." There are several possible forms of geoengineering. But as of now it seems that there is only one type that would offer a quick fix to the problem of increasing temperatures. This form of geoengineering would create an artificial sunshade by injecting reflective particles into the stratosphere that block out a small but significant fraction of about a percent or so of incoming solar radiation. Henceforth in this paper I abuse terminology by identifying the term "geoengineering" specifically with providing an artificial sunshade, which more technically is sometimes called "solar radiation management" (SRM).[3]

The planet itself naturally geoengineers a temporary sunshade when there is an explosive volcanic eruption involving sulfur dioxide (SO_2). The resulting aerosol particles that coalesce around SO_2 in the stratosphere reflect back incoming sunlight, thereby lowering the Earth's surface temperatures almost immediately. The last time this naturally occurring phenomenon transpired was during the eruption of Mount Pinatubo in 1991, which was estimated to have lowered the average surface temperature of the earth by about 0.5°C during the subsequent year or two, returning to its baseline temperature shortly thereafter.

[3]Other examples of geoengineering might include ocean fertilization, direct removal of atmospheric CO_2, creating low-level clouds from ocean spray, and so forth. Injecting reflective particles into the stratosphere is only one form of SRM, but here I blur this distinction.

For better or for worse, discussion about researching a geoengineered sunshade has grown enormously in the past five years or so. It is an extraordinarily controversial idea. A geoengineered sunshade of particles placed in the stratosphere introduces immense difficulties, dangers, uncertainties, and dilemmas of its own making. Almost no serious observer is advocating a geoengineered sunshade as a first line of defense against climate change. But it might have an important niche role as an emergency fallback component in a complete portfolio of options to deal with global warming. This might prove to be significant if very little is done about averting climate change by way of curtailing GHG emissions until noticeably disastrous effects are first bearing down upon us seriously.

A geoengineered sunshade is now the only known measure that can lower worldwide surface temperatures immediately, and therefore it represents as of now the only human response that might quickly ward off catastrophic impacts of accelerating-temperature trajectories. By comparison, carbon dioxide emissions reductions are extremely slow acting on climate change due to very long inertial lags. Even if it could be so ordained instantaneously, a complete cessation of CO_2 emissions would be unlikely to fend off many catastrophes by the time that they appeared.[4] Given the magnitude of the global public goods problem involved, many observers reluctantly consider it unlikely that significant worldwide GHG reductions will begin in earnest until and unless the threat of dangerous climate change is perceived as being tangible and imminent at the grassroots level. If this is an accurate appraisal, catastrophic climate outcomes have a built-in endogenous component and it becomes less a question of whether or not they will occur than when they will occur.

The setting for this paper's problem of geoengineering is a future world that has accumulated high enough GHG concentrations for a long enough time that some countries are feeling under severe threat from climate changes. Perhaps Bangladesh is threatened by inundation from melting ice sheets. Or maybe Indian agriculture is starting to wilt from high temperatures and monsoon alterations. Or other countries like China are beginning to be concerned with damaging climate change for other reasons. Suppose that the governments of one or more such concerned countries feel themselves under such intense domestic political pressure that they cannot wait for gradual diminishment of GHG emissions, but must come

[4]Solomon et al (2009) calculated how concentrations of CO_2 would be expected to fall off over time if *all* anthropogenic emissions were to cease immediately, following a future 2% annual growth rate of emissions up to peak concentrations of 450, 550, 650, 750, 850 and 1,200 ppm. As the authors state: "The example of a sudden cessation of emissions provides an upper bound to how much reversibility is possible, if, for example, unexpectedly damaging climate changes were to be observed." Results differed for different trajectories and scenarios, but a crude rule of thumb seemed to be that approximately 70% of the peak enhancement level over the preindustrial level of 280 ppm persevered after 100 years of *zero* emissions, while approximately 40% of the peak enhancement level over the preindustrial level of 280 ppm persevered after 1,000 years of *zero* emissions.

out in favor of geoengineering lower temperatures immediately (at very little cost to them). Suppose that much of the rest of the world fears geoengineering and opposes anyone doing it. What is the outcome?

A geoengineered sunshade has a long list of things going against it. It is scary and potentially dangerous. Some of the negatives include continued ocean acidification, depletion of stratospheric ozone, dependency effects, changed regional weather patterns, a possible weakening of resolve to cut GHG emissions, and so forth. My purpose here is not to discuss in detail the pros or cons of an engineered sunshade approach to the climate change problem. I merely want to convey the most rudimentary knowledge of the basic underlying idea for the primary purpose of motivating the model of this paper.

The economics of geoengineering have been called "incredible."[5] It appears that the direct costs of putting up a geoengineered particulate sunshade by modified high-altitude airplanes or balloon-tethered hoses or other means are extraordinarily cheap relative to the costs of mitigating GHG emissions.[6] Essentially any determined country with even a medium-sized economy could, if unopposed, put up a geoengineered sunshade on its own, in answer to its own perceived need to lower global temperatures and change its own climate quickly.

This is a true "twin externality" to the conventional externality of emitting greenhouse gases. The conventional CO_2 emissions externality is sometimes called colorfully the "mother of all externalities" because curtailing GHGs is sufficiently *expensive* that it is difficult to attain meaningful global agreement on apportioning compliance costs. But then a geoengineered sunshade might be called (also colorfully) the "father of all externalities" because knocking down global average temperatures is so *cheap* that in principle one country could do it unilaterally to fit its own particular perceived needs, thereby imposing a dangerous "public bad" on a multitude of other nations. So the world faces not one, but two global externalities from climate change.

The first, conventional, externality of curtailing GHGs is already familiar as a global public goods issue having a serious free-rider problem. The second, geoengineered-sunshade-type externality, is less familiar. I next move towards addressing this "free driver" externality in a formal model. At the center of this formal model will be a loose generalization of the

[5] The term is due to Scott Barrett (2008), who drew attention to this aspect. See also the more recent papers of Klepper and Rickels (2012) and Goes, Keller and Turano (2011). These papers contain a more detailed description of the economics of geoengineering than this paper, along with references. Thomas Schelling (1996) should be credited with first articulating the idea that the low cost of geoengineering turns the climate-change externality problem on its head.

[6] Ballpark estimates of annual geoengineering costs of offsetting projected heating in this century might be in the neighborhood of around eight billion dollars or so per year (McClellan, Keith, and Apt (2012)). The leading technology being discussed is a fleet of high-altitude airplanes specially modified to emit sulfates.

idea of type-I and type-II errors, as extended to a continuum of possible choices. The next section is intended to motivate this generalization by first discussing errors of type I and type II in a simpler and more standard discrete binary setting.

3 Background: Errors of Type I and Type II

This section is frankly metaphorical and suggestive. The purpose is to motivate envisioning geoengineering as involving two types of risks – the risk of overdoing it (here analogous to a type-I error), and the risk of underdoing it (here analogous to a type-II error). In the paper, these two errors or mistakes will have asymmetric expected losses as in a two-part tariff or a kinked loss function.

This section exposits the simplest zero-one binary choice model in a decision-theoretic context with different penalties for type-I and type-II mistakes. I give two examples. The first involves a familiar aspect of the criminal justice system and is used primarily as a conceptual device to motivate further applications. The second example involves a simple discrete-choice version of a geoengineering decision, which will serve as a more direct motivation for the more general continuous version of a free-driver externality that constitutes the core model of this paper. The aim here is to convey the loose imagery of a familiar analogy, absent any pretense that this motivating metaphor provides a rigorous foundation for the general model.

A type-I error is the rejection of a null hypothesis that is actually true. It is a false positive. By contrast, a type-II error is the acceptance of a null hypothesis that is actually false, or a false negative. These two types of mistakes may have very different risk consequences with very different penalty losses. The goal, which will later be made more explicit, is to minimize some risk-weighted sum of the two types of losses.

Consider first a binary choice example from the legal system. Let the null hypothesis be that the accused is innocent. Let x be a binary variable reflecting the judgment of a hypothetical outside social observer representing the justice system as a whole. If x=0, the outside observer believes that the accused is innocent. If x=1, this outside observer believes that the accused is guilty.

Let y be a binary variable representing the actual verdict. If y=0, the accused is found not guilty and is acquitted. If y=1, the accused is found guilty and receives a severe punishment.

A type-I error occurs when x=0 and y=1. To the hypothetical outside observer representing the justice system as a whole, an innocent person has wrongfully been found guilty. Suppose the outside observer attaches a social-penalty loss of Λ_I to this false-positive out-

come.

A type-II error occurs when x=1 and y=0. To the outside observer representing the justice system as a whole, a guilty person has erroneously been acquitted. Suppose the outside observer attaches a social-penalty loss of Λ_{II} to this false-negative outcome.

Throughout this paper it is more convenient to think in terms of *relative* penalty losses, which are normalized so that

$$\lambda \equiv \frac{\Lambda_I}{\Lambda_I + \Lambda_{II}} \tag{1}$$

is the relative penalty weight attributed to an error of type I, while $1 - \lambda$ is the relative penalty weight attributed to an error of type II. These two types of errors are unlikely to be equally costly. In the justice example a type-II error is like a disturbing error of omission, whereas a type-I error is more like a horrifying error of commission. Therefore, in this example, λ is large while $1 - \lambda$ is small.

In some sense yet to be made precise, the social observer wishes to design an optimal voting-like decision mechanism for a hypothetical jury that reflects the relative weights of the two penalty losses for errors of type I and type II.

The second example concerns a vastly oversimplified and highly abstract formulation of geoengineering as a binary choice problem. This discrete example will serve as a transition bridge to the more general continuous version of the free-driver problem, which is the main subject of the paper.

The null hypothesis here is that geoengineering is undesirable. Let x be a binary variable reflecting the opinion of an interested party about whether or not geoengineering should be undertaken. If x=0, the interested party believes that geoengineering is harmful to them and should not be undertaken. If x=1, the interested party believes that geoengineering is beneficial for them and should be undertaken.

Let y be a binary variable representing the actual outcome of geoengineering. If y=0, geoengineering is not undertaken. If y=1, geoengineering is undertaken.

A type-I loss occurs when x=0 and y=1. In this case geoengineering is undertaken despite the fact that it harms the interested party. From the point of view of the interested party, geoengineering is *over*done here, resulting in a type I error. Suppose that in this situation the relative social loss is λ.

A type-II loss occurs when x=1 and y=0. In this case geoengineering is not undertaken despite the fact that it benefits the interested party. From the point of view of the interested party, geoengineering is *under*done here, resulting in a type II error. In this situation the relative social loss is $1 - \lambda$.

As with the justice example, it seems reasonable to suppose that a type-II loss (geoengi-

neering is underdone) might be disturbing to the interested party because their welfare is sub-optimal, whereas a type-I loss (geoengineering is overdone) might be horrifying to the interested party because it represents a relatively much riskier strategy with a relatively much more heavily weighted downside. Therefore, in this binary geoengineering example, λ is relatively large while $1 - \lambda$ is relatively small.

In a sense that is about to be made precise within a more general setting, the social planner wishes to design a constitution for an optimal voting-like decision mechanism that reflects the relative riskiness-weights of the two penalty losses. It is to this more general formulation that we now turn.

4 The Pure Theory of a Free-Driver Externality

Geoengineering represents a kind of perverse public good having some distinctive properties. I feel that the role of the geoengineering externality will be better appreciated when it is first studied in its abstract pure form – as a particular example belonging to the public-good-like family, but having special features whose generic properties warrant attention in their own right.

A pure public good is typically defined as a commodity that is both non-excludable (no one can be excluded from consuming it) and non-diminishable (one person's consumption does not alter the amount available to others). A public good is standardly considered to be *good*, meaning that almost everyone thinks more of it is better, at least throughout the domain having policy relevance. Usual examples are police and fire protection, national defense, weather predictions, and the like.

A pure public bad is typically defined as a commodity that is both non-exemptable (no one can be exempted from consuming it) and non-diminishable (one person's consumption does not diminish the amount that others must consume). A public bad is standardly considered to be *bad*, meaning that almost everyone thinks more of it is worse, at least throughout the domain having policy relevance. A standard example of a public bad is pollution.

Because it is costly to increase the level of a public good or to decrease the level of a public bad, such situations are plagued by the free-rider problem. Instead of paying their fair share, everyone wants to free ride off the payments of everyone else. The problem of a geoengineering externality has a different structure.

I now want to introduce the idea of a *gob*.[7] A "gob" is a commodity that may be *good or*

[7] I think it is somewhat clearer for a reader if I use fresh terminology rather than attempting to shoehorn this problem into the already existing terminology of public goods when this problem is not a fully comfortable

bad depending on who is consuming it and how much they are consuming. A *pure public gob* is a pure public *good* (more of it is better) for some people under some circumstances and a pure public *bad* (more of it is worse) for some other people under some other circumstances. Throughout this paper, the primary example of a pure public gob is geoengineering in a future world sufficiently impaired by climate change that some countries would want to do some of it on their own if allowed to act unilaterally. The key issue is that parties *differ* in their attitudes toward whether more or less gob is desirable and some mechanism is required to reconcile these differences.

A *free-driver externality* is a pure public gob whose production happens to be *free* (or, in practice, is sufficiently inexpensive to be considered free). In this paper the inspiration for, and primary application of, a "free-driver externality" is geoengineering the stratosphere with reflective particles to reflect back incoming solar radiation. This would be so relatively cheap that many nations could afford to do it unilaterally.

The key abstraction about being "free" in a "free-driver externality" is that, absent the rules of some overarching governance structure, each agent is in principle *free to choose* (at zero cost to itself) the gob level that will be imposed on itself and all of the other agents. Depending on the nature of the gob and its reversibility by other agents, this leads either to anarchy with an undefined outcome (for free reversibility by other agents) or to an extreme outcome dominated by the agent with the most extreme preferences (for complete irreversibility by other agents). I assume that the latter situation is relevant for geoengineering because it is difficult to do counter-geoengineering.

The theoretical core of this paper characterizes the socially optimal level of gob production in an abstract setting and shows that (under a particular piecewise-linear specification) it can be implemented by a relatively simple supermajority-type voting rule. It is possible to pose the free-driver externality problem in somewhat more general form than I do here, but only at the expense of dulling a sharp simple result. In this paper I aim for sharpness and simplicity. Therefore, in what follows, I abstract heroically – to put it mildly. At the very least, the crisp formulation of this paper can serve as a benchmark point of departure for more complicated and fuller analyses.

Let there be n "nations" indexed by $i = 1, 2, ..., n$. There are m_i "citizens of nation i" and a total of m "citizens of the world," where

$$m = \sum_{i=1}^{n} m_i. \tag{2}$$

fit with the existing terminology.

The citizens of each nation have identical preferences with each other but (possibly) different preferences from the citizens of other nations. In this metaphor each citizen will have one vote and it will not matter whether citizens of nation i vote individually or as a bloc with m_i votes. At the highest level of abstraction, the $\{m_i\}$ are given equity-welfare voting weights that have already been assigned on the basis of some or another criterion.

Suppose that a citizen of nation i prefers the gob level $x_i \geq 0$ to any other level. In the case of geoengineering, this "preference" is essentially for a level of geoengineering (as measured, say, by sulfate levels) that best offsets the deleterious effects of climate change being experienced by nation i. Without loss of generality, nations are arranged in ascending order of gob preference so that

$$i < j \implies x_i \leq x_j. \tag{3}$$

Let $y \geq 0$ be the *actual* level of worldwide gob production. Let $L_i(y)$ be the loss function for a citizen of nation i. This paper considers a very simple loss function, which embodies the concept of constant per-unit penalties for type-I and type-II mistakes. When $y \geq x_i$, the citizens of nation i suffer what to them is a type-I error of magnitude $y - x_i$ (geoengineering is overdone). When $y < x_i$, the citizens of nation i suffer what for them is a type-II error of magnitude $x_i - y$ (geoengineering is underdone).

All citizens of all nations have the same per-unit penalty of λ for a type-I error, and the same per-unit penalty of $1 - \lambda$ for a type-II error. Therefore,

$$y \geq x_i \implies L_i(y) = \lambda(y - x_i), \tag{4}$$

and

$$y < x_i \implies L_i(y) = (1 - \lambda)(x_i - y). \tag{5}$$

The loss function (4), (5) is of a simple piecewise-linear form with a single kink at x_i. In this sense, x_i acts as a kind of reference point for nation i. The upward per-unit loss aversion is λ for an error of type I, while the downward per-unit loss aversion for an error of type II is $1-\lambda$. Citizens differ only by their preferred reference level of the geoengineering gob, with the per-unit loss aversion for *over*doing geoengineering being identically λ for an error of type I and the per-unit loss aversion for *under*doing geoengineering being identically $1 - \lambda$ for an error of type II. This is a very strong assumption. For sure, the crisp voting result of the paper depends on this simple kinked penalty function with the same slopes for everyone, where the only difference is the location of the kink. I cannot provide a strong foundation for this assumption, but must instead rely on heuristics and the fact that it gives a sharp result which might serve as a point of departure for further discussion.

Without loss of generality it is assumed that $\lambda > 1/2$. (The case $\lambda < 1/2$ involves a symmetric treatment, while the case $\lambda = 1/2$ is familiar from median-voter theory.) Thus, in what follows a gob level above the desired reference level involves a type-I penalty that is greater than the type-II penalty for a gob level equally far below the desired reference level. In the situation of geoengineering, λ might well be deemed to be larger than $1 - \lambda$ because overdone geoengineering involves risks that are potentially more dangerous than underdone geoengineering (although this logic could be reversed for some scenarios).

If states are sovereign and do not have binding treaty obligations they can, at least in principle, act unilaterally in their own self interest by choosing their own favorite amount of reflective particles to place in the stratosphere. (This is an abstraction of a more complicated situation where states have some responsibility not to harm other states, are not giving or receiving behavior-altering side payments, and so forth.) The non-cooperative Nash-equilibrium outcome $\widetilde{y}$ is then the maximum of the preferred geoengineering-gob level among all nations. By (3), the nation who favors the most gob is nation n. Therefore

$$\widetilde{y} = x_n = \max_{1 \leq i \leq n} \{x_i\}. \tag{6}$$

The nation n, which favors the most gob, is called the *dominant free driver*.

Even without yet defining formally the socially optimal gob level, what leaps out of (6) is the strong degree of non-optimality of $\widetilde{y}$. In the Nash equilibrium, free-driven gob is *oversupplied* because only the dominant driver is fully satisfied with the outcome – everyone else wants less gob but is forced to accept a large per-unit type-I loss of λ. In this setup there are not just winners and losers. Only the dominant free driver is a winner – everyone else is a loser relative to their preferred gob level because they are exposed to the excessive risk of a type-I error.[8]

Of course, this model is just a particularly heroic abstraction of a much more complicated situation. Even so, the message would appear to be that geoengineering looks like a dangerous global externality accident waiting to happen, which has the potential to cause serious international frictions and even outright conflicts if it is left to simmer away on its own. Overall, the externality-governance issues raised by geoengineering look severe enough to warrant being addressed by the international community long before the problem might actually raise its ugly head.

What is the socially-optimal level of geoengineering gob? To answer this question requires a bit more notation.

[8] This does not exclude the possibility that society might be better off at the extreme point $y = x_n$ than at the extreme point $y = 0$.

For any nonnegative gob level y, let $F(y)$ be the cumulative distribution function, meaning the fraction of the population whose preferred gob level is less than or equal to y. Thus,

$$x_i \le y < x_{i+1} \implies F(y) = \frac{1}{m}\sum_{j=1}^{i} m_j. \tag{7}$$

The worldwide *social loss function* $\mathcal{L}_\lambda(y)$ is postulated to be the utilitarian sum of each citizen's loss function. From (4), (5) this means that

$$\mathcal{L}_\lambda(y) = \lambda \int_0^y (y-x)\, dF(x) + (1-\lambda) \int_y^\infty (x-y)\, dF(x), \tag{8}$$

where the integration in (8) refers to a Riemann-Stieltjes integral that accommodates $F(x)$ being a step function.

A λ-*optimal* gob level y^* satisfies

$$\mathcal{L}_\lambda(y^*) = \min_{0 \le y < \infty} \{\mathcal{L}_\lambda(y)\}, \tag{9}$$

where existence of such a minimizing y^* is guaranteed because $\mathcal{L}_\lambda(y)$ is continuous in $y \ge 0$ and $\mathcal{L}_\lambda(\infty) = \infty$.

The next task is to show that a socially optimal gob level y^* is supported as a supermajority voting equilibrium and vice versa.

Much of the paper to this point has been devoted to justifying (8) with a story exposited in terms of geoengineering gob. An alternative route would have been to *begin* with (8) as a social loss function, leaving its justification in the background since (8) might apply for many situations (with or without free driving). This alternative route might focus an even sharper spotlight on the key analytical result of the paper, which is to show that there is a tight duality connection between optimized social welfare (9) (when expressed by the particular loss function (8)) and a simple supermajoritarian voting implementation mechanism. In other words, the pure theory has a stand-alone quality that does not require the motivational example of free-driving geoengineering, although, in my opinion, it greatly enhances the telling of the story.

5 The Socially-Optimal Gob as a Voting Equilibrium

We seek a robust governance architecture with "good" properties that can react automatically to balance ever-changing opinions and attitudes about individually-desired levels of geoengineering gob $\{x_i\}$. In other words, the individual $\{x_i\}$ might change over time

depending upon circumstances, but the ideal governance constitution should automatically select the gob level y^* that is optimal for these changed values of $\{x_i\}$.

Consider any two levels of gob y' and y''. Suppose $y' < y''$. Consider the following asymmetric pairwise θ-*voting* rule. To *raise* the level of geoengineering gob from y' to y'' requires the approval of at least the fraction θ of voters, in which case we write $y'' \succsim_\theta y'$. In the other direction, to *lower* the level of geoengineering gob from y'' to y' requires the approval of at least the fraction $1-\theta$ of voters, in which case we write $y' \succsim_\theta y''$.

A θ-*voting equilibrium* is a value $\widehat{y}$ that defeats (or at least ties) every other possible candidate in a θ-voting binary comparison – i.e., for all $y \geq 0$ it holds that

$$\widehat{y} \succsim_\theta y. \tag{10}$$

In this setup with type-I and type-II errors, what is the relationship between a voting equilibrium and a social optimum? The following proposition is a generalization of the median-voter theorem. (The median-voter theorem corresponds to the special case $\theta = \lambda = 1/2$.) The result presented in the following theorem is new and constitutes the main theoretical contribution of this paper.[9]

Theorem 1 *The gob level y^* is λ-optimal if and only if y^* is a λ-voting equilibrium.*
Proof. *Differentiating (8) from the right, the right hand side derivative of $\mathcal{L}_\lambda(y)$ is*

$$\frac{d\mathcal{L}_\lambda}{dy+} = \lambda \int_0^{y+} dF(x) - (1-\lambda) \int_{y+}^{\infty} dF(x) = F(y) + \lambda - 1. \tag{11}$$

For all $y>0$, define

$$F^-(y) \equiv \lim_{\epsilon \to 0+} F(y-\epsilon), \tag{12}$$

and define $F^-(0) = 0$.

[9] For the technically minded reader, the model is a special variant of choosing a one-dimensional public outcome when preferences are single peaked. The voting mechanism I recommend is one of the classic "positional dictator" mechanisms (Moulin (1991), section 10.2). Such mechanisms have the good properties of group-strategy-proofness, efficiency, and fairness. My special contribution is to motivate and study a special asymmetric pair of marginal disutilities as one moves away from the peak in each direction. (The symmetric case is standard and associated with median-voter theory.) Extending the argument for a well-known result from the median voter $\theta = 1/2$ literature (see, e.g., Easley and Kleinberg (2010), section 23.6) to the case $\theta \neq 1/2$ can be used to prove that a θ-voting rule applied to all pairs of alternatives produces a group voting-preference relation that is complete and transitive.

Differentiating (8) from the left when $y > 0$, the left hand side derivative of $\mathcal{L}_\lambda(y)$ is

$$\frac{d\mathcal{L}_\lambda}{dy-} = \lambda \int_0^{y-} dF(x) - (1-\lambda) \int_{y-}^{\infty} dF(x) = F^-(y) + \lambda - 1. \tag{13}$$

Both $F(y)$ and $F^-(y)$ are monotone non-decreasing in y with $F^-(y) \leq F(y)$, signifying from (11) and (13) that the function $\mathcal{L}_\lambda(y)$ is convex. The necessary and sufficient condition for $\mathcal{L}_\lambda(y)$ to be minimized is therefore

$$0 \leq \frac{d\mathcal{L}_\lambda}{dy+} \tag{14}$$

for $y = 0$, and

$$\frac{d\mathcal{L}_\lambda}{dy-} \leq 0 \leq \frac{d\mathcal{L}_\lambda}{dy+} \tag{15}$$

for $y > 0$.

Combining (11) and (13) with (14) and (15), gob level y^* minimizes $\mathcal{L}_\lambda(y)$ if and only if it satisfies the condition

$$F^-(y^*) \leq 1 - \lambda \leq F(y^*). \tag{16}$$

We next show that (16) implies that y^* is a λ-voting equilibrium.

Pick any $y'' > y^*$. Then at least the fraction $F(y^*)$ of voters are "closer" to y^* than to y'', and therefore prefer y^* to y''. Equivalently, no more than the fraction $1 - F(y^*)$ prefers y'' to y^*. But from (16), $1 - F(y^*) \leq \lambda$, which then means that no more than the fraction λ of voters prefers y'' to y^*. This implies, by the λ-voting rule, that $y^* \succsim_\lambda y''$.

Pick any $y' < y^*$ (if $y^* > 0$). Then at least the fraction $1 - F^-(y^*)$ of voters are "closer" to y^* than to y', and therefore prefer y^* to y'. Equivalently, no more than the fraction $F^-(y^*)$ prefers y' to y^*. But from (16), $F^-(y^*) \leq 1 - \lambda$, which then means that no more than the fraction $1 - \lambda$ of voters prefers y' to y^*. This implies, by the λ-voting rule, that $y^* \succsim_\lambda y'$.

To show that y^* being a λ-voting equilibrium implies (16), we employ a local small-perturbation argument. Let $\epsilon = 0^+$ be an arbitrarily small positive number.

The fraction of voters who prefer $y^* + \epsilon$ to y^* is $1 - F(y^*)$. But $y^* \succsim_\lambda y^* + \epsilon$ implies by the λ-voting rule that no more than the fraction λ of voters prefers $y^* + \epsilon$ to y^*. Thus, $1 - F(y^*) \leq \lambda$. In the other direction, if $y^* > 0$ then the fraction of voters who prefer $y^* - \epsilon$ to y^* is $F^-(y^*)$. But $y^* \succsim_\lambda y^* - \epsilon$ implies by the λ-voting rule that no more than the fraction $1 - \lambda$ of voters prefers $y^* - \epsilon$ to y^*. Thus, $F^-(y^*) \leq 1 - \lambda$. Combining these two conditions yields (16). ■

The θ-voting rule corresponds to a form of supermajoritarianism that already exists and

is familiar for special situations throughout the real world. One might then invert Theorem 1 to ask the following question. Given some value of θ, for what class of preferences is the θ-voting rule socially optimal? Theorem 1 states that the θ-voting rule is socially optimal for preferences that are "as if" given by (4), (5) with $\lambda = \theta$. To go beyond this "as if" characterization to a more general description of preferences for which the θ-voting rule is socially optimal is an interesting subject of future research that would take this paper too far afield. Here I can only hope that Theorem 1 gives some broad insights as an approximation that extends beyond its restrictive preference structure. What is remarkable here, I think, is not that Theorem 1 is restrictive and criticizable, but that one can obtain such a social-optimality result at all from a θ-voting rule. I view Theorem 1 as a point of departure for further discussion, not the final word on a very complicated subject.

6 A Naive Geoengineering-Governance Proposal

The idea of geoengineering is not about to go away any time soon. If anything, interest in solar radiation management is likely to grow over time. Geoengineering is simply too cheap and too tempting for it to recede politely from public view. My basic premise is that we must do some serious thinking about the architecture of a geoengineering governance structure – sooner, rather than later.

What are we to make of Theorem 1? Can it be taken seriously? I guess the answer depends, at least in part, on the alternatives. An old adage has it that "you can't beat something with nothing." Suppose we allow a willing suspension of disbelief. In the spirit of putting something constructive on the discussion table, I propose the following idea.

Yes, we need advisory commissions with public participation for the governance of geoengineering. And yes, we need to balance standards of oversight with international political reality and with principles of transparency and accountability. But at the end of the day this is all too vague. At the end of the day we need to have some concrete governance structure with specific rules concerning how to make final decisions about geoengineering levels that differentially impact parties having different interests. Otherwise, with a free-driver externality, we risk paralysis and conflict.

For the sake of specificity, I somewhat arbitrarily propose that a type-I error of overdone geoengineering be given a relative penalty weight three times that for a type-II error of underdone geoengineering. In the notation of this paper, I am setting $\lambda = 3/4$. This value corresponds to a voting system that requires a 3/4 majority. (A more cautious person who puts a heavier weight on a type-I error of over-geoengineering relative to a type-II error of under-geoengineering might prefer a value of $\lambda = 4/5$, say, while a less cautious person might

prefer $\lambda = 2/3$, say.[10])

A permanent "international governance structure for geoengineering" is established, at the core of which is a body acting like a legislative general assembly. Each country has a metaphorical representation in the general assembly, with voting weight proportional to its population, say. Any proposal to increase the level of geoengineering requires at least a 3/4 supermajority of the general assembly. Any proposal to decrease the level of geoengineering requires at least a 1/4 "superminority" of the general assembly. An executive arm is empowered to carry out decisions of the general assembly and to assess penalties for noncompliance. A judicial arm adjudicates conflicts.

Is this proposal naive? Almost surely yes. To begin with, there are very few precedents of international voting outcomes applying with binding force. More generally, I am simplistically brushing aside a great many truly important aspects of the real world of international agreements.

There is already a sizable literature concerning the nuances and difficulties of geoengineering governance written by distinguished experts on international law and politics.[11] The tone of this literature is grounded in the realities of global politics and is largely pessimistic about the prospects for workable geoengineering governance. This paper has somewhat different aims, being more theory-based, more speculative, more heroic, and more naive. The proposal of this section is being mooted not so much as a reality-encapsulated operational plan, but more in the spirit of a theory-based point of departure for further discussion. Maybe the world is not yet ready for such a heroic international voting governance structure. On the other hand, maybe we need to start thinking along such radical lines and the threat of geoengineering provides the impetus to try.

Why would countries voluntarily accede to a voting limitation on their sovereignty? That is, why would a country agree in the first stage to participate in such a second-stage voting architecture? I do not have a good answer to this question except to ask another question. What are the alternatives for geoengineering governance and on what alternative theory are they based? The paper blindly assumes that the geoengineering free-driver problem is sufficiently threatening to encourage first-stage participation and concentrates on examining the second-stage voting consequences. There is a tension here, which I am unable to resolve, between presenting a specific constructive voting proposal for addressing an important ex-

[10] Ideally, the appropriate value of λ is thrashed out at some kind of constitutional convention of the parties that occurs during a prequel when some "veil of ignorance" might apply and well before free-driver geoengineering becomes an actual threat. This is yet another real-world detail that I am putting aside in favor of focusing on the big picture.

[11] See, e.g., Parson and Ernst (2012), Bodansky (2011), Victor (2008), Horton (2011) and the many references cited therein.

ternality problem and being an easy target for criticism on the grounds that participation is impractical.

What comes out of this model, I think, is a loose sense that the free-driver problem of geoengineering may be ever-so-slightly easier to resolve by a voting architecture than the free-rider problem of GHG abatement. In the latter problem, the participants have first to agree on the fraction of abatement that each will bear – before ever getting to second-stage voting on the level of overall abatement. This first-stage complication is absent from a free-driver problem like geoengineering because participants need not negotiate what is effectively a payment-transfer agreement. Intuitively, it may be relatively easier (although undoubtedly still very difficult) to reach agreement on a voting architecture when the parties do not have to first argue about who pays what. Admittedly this argument is informal, but I think it carries some weight.

7 Concluding Comments

At the beginning of this paper I posed the basic question of whether or not there exists a solution-theoretic concept, however hypothetical and abstract, that might form the backbone of a governance architecture for a free-driver externality. I think that Theorem 1 is giving an affirmative answer to this question in the form of a supermajority voting rule. But of course readers are free to form their own opinions and to have their own answers. It is true that a number of very strong assumptions have been made to obtain the basic result. The model is subject to all of the many caveats that apply to the median voter theorem (which is a special case of Theorem 1 corresponding to $\theta = \lambda = 1/2$). That is on the one hand. On the other hand, the conclusion of Theorem 1 is quite striking. There is not really a comparable voting-optimality result available for a free-rider externality because the problem of apportioning compliance costs adds an extra dimension of strategic complexity.[12] In this sense I think that a free-driver externality may be more "solvable" than a free-rider externality. Even though it may not be easy to apply the principle of Theorem 1 in practice, at least there *exists* such a principle.

The assumptions behind Theorem 1 are very restrictive. Preferences take the specific form of a piecewise-linear loss function with everyone having the same relative-penalty slopes for errors of type I and type II. As usual, however, the restrictiveness of the assumptions behind the model must be weighed against the power of the results coming out of the model.

[12] But see Bergstrom (1979) for discussion of some special cases of apportioning costs in a majority voting context. Ricke, Moreno-Cruz, and Caldera (2013) investigate strategic coalition formation in a geoengineering climate game that is different from the setup of this paper.

Here a relatively simple supermajoritarian rule overcomes the free-driver externality to obtain the socially optimal solution. Readers must judge for themselves the relevance of conclusions based upon this model in a domain where strong results are scarce. Unsurprisingly, my own conclusion would be that Theorem 1 may serve as a useful starting point for concentrating the mind on a serious discussion of a decent architecture for the governance of geoengineering.

References

[1] Barrett, Scott (2008). "The incredible economics of geoengineering." *Environmental and Resource Economics* (39), 45-54.

[2] Bergstrom, Theodore (1979). "When Does Majority Rule Supply Public Goods Efficiently?". *Scandinavian Journal of Economics*, 216-226.

[3] Bipartisan Policy Center (2011). "Report of the Task Force on Climate Remediation Research." Bipartisan Policy Center, Washington DC..

[4] Bodansky, Daniel (2011). "Governing Climate Engineering: Scenarios for Analysis." Belfer Center Discussion Paper 2011-47.

[5] Easley, David, and Jon Kleinberg (2010). *Networks, Crowds, and Markets.* Cambridge University Press.

[6] Goes, M., Keller, K., Turana, N. (2011). "The economics (or lack thereof) of aerosol geoengineering." *Climatic Change*, 109, 719-744.

[7] Horton, Joshua B. (2011). "Geoengineering and the Myth of Unilateralism: Pressures and Prospects for International Cooperation." *Stanford Journal of Law, Science, and Policy*, vol. IV, 56-69.

[8] Keith, David. A Case for Climate Engineering. MIT Press.

[9] Klepper, Gernot, and Wilfried Rickels (2012). "The Real Economics of Climate Engineering." *Economics Research International*, vol. 2012, article ID 316564.

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

[11] Moulin, Hervé (1991). *Axioms of Cooperative Decision Making.* Cambridge University Press.

[12] Parson, Edward A., and Lia N. Ernst (2012). "International Governance of Climate Engineering." UCLA Law School Research Paper no. 12-23 (forthcoming in 2013 *Theoretical Inquiries in Law*).

[13] Ricke, Katharine L, Juan B Moreno-Cruz and Ken Caldeira (2013). "Strategic incentives for climate geoengineering to exclude broad participation." *Environ. Res. Lett.* 8 (2013) 014021 (8 pp.).

[14] Schelling, Thomas (1996). "The economic diplomacy of geoengineering." *Climatic Change* 33, 303-307.

[15] Solomon, Susan, Gian-Kasper Plattner, Reto Knutti, and Pierre Friedlingststein (2009). "Irreversible climate change due to carbon dioxide emissions." *Proceedings of the National Academy of Sciences* 106 (6): 1704-1709.

[16] U.K. Royal Society (2009). *Geoengineering the Climate: Science, Governance and Uncertainty.* Mimeo.

[17] U.S. National Academy of Sciences (2010). *Advancing the Science of Climate Change.* Mimeo.

[18] Victor, David G. (2008). "On the regulation of geoengineering." *Oxford Review of Economic Policy*, 24(2), 322-336.