University of Heidelberg

Department of Economics

Discussion Paper Series | No. 644

A Model of Solar Radiation Management Liability

Tobias Pfrommer

January 2018

A Model of Solar Radiation Management Liability*

Tobias Pfrommer†
Heidelberg University

January 17, 2018

Abstract

Solar Radiation Management (SRM) is a set of potential technologies to counteract climate change. Liability regimes are one potential form of governance institution to avoid global externalities caused by the SRM "free-driver" problem. In this paper I examine the incentives structure and welfare consequences of SRM liability regimes. Characteristics specific to SRM impact on the incentives that liability regimes provide via the definition of harm and the liability standard. Consequently, a liability regime is defined as a combination of a definition of harm and a liability standard in the model. Providing several interpretations of these two dimensions adequate for the SRM context, I show that only one combination implements the social optimum. A numerical implementation of the model yields that the free-driver problem is moderate given a metric of mean temperature and extreme given a metric of mean precipitation. Furthermore, the implementation suggests that liability regimes are generally capable of mitigating the free-driver problem substantially and that the choice of the definition of harm is more consequential than the choice of the liability standard.

Keywords: Solar Radiation Management, Liability Regimes, Externalities, Climate Engineering, Free-Driver Scenario

JEL Codes: Q53, Q54, K13

---

*I want to thank Ulrike Niemeier and Ben Kravitz for providing me with the climate model data used in this study. Furthermore, I want to thank Timo Goeschl, Daniel Heyen, Johannes Lohse and John Stranlund, as well as conference participants at the EAERE 2016 in Zurich and the CEC17 in Berlin for helpful comments. I gratefully acknowledge funding from the Priority Programme 'Climate Engineering: Risks, Challenges, Opportunities?' (SPP 1689) of the German Research Foundation (DFG).

†Email: pfrommer@eco.uni-heidelberg.de. Postal address: Department of Economics, Bergheimer Str. 20, 69115 Heidelberg, Germany. Phone: +49 6221 548014, Fax: +49 6221 548020.

1 Introduction

Due to slow progress of climate change mitigation, techniques to increase the reflection of incoming solar radiation in the atmosphere, so-called Solar Radiation Management (SRM), have received increasing attention as potential means to reduce climate change risks. SRM is a potential high-leverage set of technologies which could be capable of lowering global temperatures within short time-scales (Keith et al. 2010). Under plausible assumptions, SRM seems to be cheap enough to be undertaken by a single country and with very small direct costs, compared to mitigation or unmitigated climate change damages (Barrett 2008, Keith et al. 2010). Since SRM also would have regionally different impacts (Lunt et al. 2008, Robock et al. 2008, Irvine et al. 2010, Ricke et al. 2010), it constitutes a "free-driver" problem (Weitzman 2015): Without any form of governance in place, the country with the strongest preferences for SRM has incentives to deploy SRM beyond the preferred provision point of all other countries. This free-driver outcome is highly undesirable from a social point of view and calls for some form of governance.

An emphasis on the need for governance for Geoengineering in general, and SRM in particular, is ubiquitous in the literature (Barrett 2008, Shepherd 2009, Keith et al. 2010, Rayner et al. 2013, Pasztor 2017). Liability regimes as potential tools for SRM governance have gained wide attention, with a focus on historical precedents, the applicability of existing international law to SRM, political feasibility and the issue of causation (Horton et al. 2014, Saxler et. al 2015, Reynolds 2015). From an economic point of view, the purpose of liability regimes is to solve incentive problems and liability regimes are a widely used and researched tool for internalizing environmental externalities.[1] In this paper I develop a theoretical model of SRM liability which I then numerically implement, in order to understand the basic incentive structure and to examine the extent to which different liability regimes can solve the free-driver incentive problem.

SRM has a key feature which sets it apart from more traditional domains of liability like car accidents or pollution problems. Following Weitzman's terminology, SRM is a public good-or-bad, a public good which benefits agents at some levels and harms the same agents at other levels: Studies focusing on two of the most important climate

[1] Prominent national, supranational and international examples include the Comprehensive Environmental Response, Compensation, and Liability Act in the US, the European Environmental Liability Directive 2004/35/EC, the International Convention on Civil Liability for Oil Pollution Damage and the Convention on International Liability for Damage Caused by Space Objects.

metrics, mean temperature and mean precipitation, suggest that moderate amounts of SRM would benefit most regions of the world (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015) and that SRM would only start to be detrimental to those regions' welfare if provided beyond those moderate amounts.[2] The public good-or-bad characteristic impacts on the incentives a liability regime provides in two ways. The first one is via the definition of harm, i.e. the question of which SRM impacts have to be compensated for. The second one is via the liability standards, i.e. the question of the circumstances under which harm from SRM has to be compensated for.

In this paper I focus on the implications of SRM's public good-or-bad characteristic for liability regimes. Consequently, a liability regime in the model consists of a definition of harm and a liability standard. There are n agents in the model, who can be thought of as countries or regions, having climate preferences in the form of convex damage functions. In order to reflect the good-or-bad characteristic, at least some agents benefit from moderate SRM levels. The agent with the strongest preferences for SRM is assumed to be the sole SRM provider. Direct costs are assumed to be negligible. I examine the equilibrium outcome both under no liability (the free-driver outcome) and under various liability regimes, each consisting of a definition of harm and a liability standard, relative to the social optimum defined by the minimization of aggregate damages.

The reference point against which harm is measured or should be measured is not self-evident for a public good-or-bad. One possibility is to use the victim's position in a world without any SRM as reference point. I call this the *absolute definition of harm*. A second possibility is to use the victim's preferred provision level as reference point, a world in which SRM is not provided beyond the victim's optimum. I call this the *marginal definition of harm*. In contrast, the two definitions of harm coincide for a pure bad like car accidents or pollution, since a victim's optimal provision level is then always zero.

Negligence, one of two fundamental types of liability standards, uses a behavioral standard in order to determine whether to assign liability. The traditional economic interpretation of the negligence standard is that it balances the marginal costs with the marginal benefits of avoiding harm (Posner 1972, Landes and Posner 1987): The injurer can forgo a reduction of own damages (and potentially those of some third parties) in

[2]Mean temperature and mean precipitation are of great relevance for impacts which could trigger a lawsuit: directly, since they, for example, greatly influence which types of agriculture are feasible in a given region and indirectly, since they are closely connected to the probability of occurrence of extreme events.

order to not increase damages of other agents. In a one-victim-one-injurer setting, there is only one way to trade off marginal costs and benefits from avoiding harm. However, the public good-or-bad SRM constitutes a multiple-victim-third-party-beneficiary setting, raising the question of whose costs and whose benefits are or should be traded off by a negligence standard. I will give three interpretations of the negligence standard.

From a normative welfare perspective all agents' welfare should be considered in the negligence standard. I call the standard emerging from considering all agents benefits and harms the *benefit-harm negligence* standard. However, consideration of effects on parties that are not part of the trial is generally not permissible in international law, probably the most important body of law for SRM, rendering the *benefit-harm negligence* standard unlikely to be applied in practice. The other two interpretations are designed to reflect potential scenarios of a trial and third-party beneficiaries are consequently excluded from the standard in these interpretations. The first scenario is a trial between all victims and the injurer. Here, the victims' harm is considered on aggregate, giving rise to the *aggregate harm negligence* standard. In the second scenario, there are individual trials between each victim and the injurer. Here, each victim's harm is considered individually, giving rise to the *individual harm negligence* standard, which sets a standard for each individual victim. I will consider these three negligence standards and the other fundamental type of liability standard, *strict liability*. Under *strict liability* an injurer is liable for all harm she causes irrespective of her behavior.

I find that only one liability regime implements the social optimum in general – the *marginal definition of harm* combined with the *benefit-harm negligence* standard. However, as already noted, the *benefit-harm negligence* standard is unlikely to be employed in a real-world scenario. All other liability regimes are biased. The direction of these biases is often ambiguous in general, since there are often multiple biases at play, which potentially pull into opposing directions.

Liability regimes employing the *absolute definition of harm* cannot implement the social optimum in general, since it only reflects increases in the victims' damage levels above the respective victim's damage level without any SRM at all. In contrast, the *marginal definition of harm* reflects all increases in the respective victim's damage levels due to increases in SRM provision. The former definition is therefore biased towards too high SRM provision levels, while the latter is unbiased. This result is of importance for SRM compensation regimes more generally, in that any compensation regime must define a reference point which is used to determine the amount of compensation to award. The

characteristics and incentive effects of the absolute and the *marginal definition* then carry over to their respective counterparts in any mechanism under which the SRM provider has to compensate victims.

Liability regimes employing the *benefit-harm negligence* standard can implement the social optimum in general. This standard is unbiased since it considers all agents' welfare. All other liability standards do not internalize the positive externality. *Strict liability* and the *aggregate harm negligence* standard both fully internalize the negative externality. They therefore implement the same SRM provision level in equilibrium and are biased to too low SRM levels. The *individual harm negligence* standard does not fully internalize the negative externality, since each victim's harm is balanced individually against the injurer's benefits. Its bias is therefore ambiguous in general. No liability implements the free-driver outcome.

I numerically implement the SRM liability model into the Residual Climate Response (RCR) model (Moreno-Cruz et al. 2012) which has been developed and used (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015) to examine regional SRM effects. I do so for two reasons. Firstly, in order to obtain an estimate of how severe the SRM governance problem is, I want to quantify the in the theoretical literature well-established (Weitzman 2015, Heyen 2016) free-driver-problem. Secondly, the numerical implementation of the liability model might help illuminate how the performance of the non-optimal liability regimes compared to the free-driver outcome and the social optimum is, whether there are major differences in performance between these regimes and whether the choice of the definition of harm and the choice of the liability standard are equally important. The RCR model is a simple framework for evaluating regional climate responses to SRM which uses quadratic regional damage functions in regional mean temperature and precipitation, with damages being minimal and normalized to zero at regional preindustrial conditions. For the implementation I use data from the G1 experiment of the Geoengineering Intercomparision Project (Kravitz et al. 2011).

In line with the literature (Moreno-Cruz et al. 2012, Yu et al. 2015), I find that socially optimal SRM is very effective at reducing residual damages for the temperature metric (0.2% of unmitigated climate change damages) and effective for the precipitation metric (5.1%). Concurrent research comes to the conclusion that the SRM governance problem might be substantial: Using an integrated assessment model approach for quantifying the free-driver outcome, Emmerling and Tavoni (2017) find SRM overprovision of a factor of eight. Using the much simpler RCR model approach, I find that the

the extent of the free-driver problem depends on the metric chosen: For a metric of mean temperature there is only moderate SRM overprovison in the free-driver outcome in which SRM is still capable to reduce damages effectively (1.8%). However, there is drastic overprovision for a metric of mean precipitation in the free-driver outcome, leading to damages 6.5 times higher than without any SRM. These findings confirm earlier results that regional differences in SRM impacts are larger for precipitation than for temperature (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015). In the free-driver outcome, the differences in residual damages between the two metrics are amplified, since SRM provision is according to the strongest preferences for SRM.

In the presence of liability regimes, SRM is also for the precipitation metric implemented in a welfare-enhancing way: All regimes reduce damages to at most 19.6% of unmitigated climate change damages for this metric. Liability regimes employing the *marginal definition of harm* virtually implement the social optimum for both metrics. For the temperature metric, the *absolute definition*'s bias renders liability regimes without any effect at all. Differences in outcomes across the definition of harm are larger than differences in outcomes across liability standards and liability regimes employing the *marginal definition of harm* do consistently better than regimes employing the *absolute definition*. Therefore, given the assumptions of this numerical implementation, the choice of the definition of harm is more consequential than the choice of the liability standard for the performance of a liability regime.

The paper proceeds as follows. Section 2 lays out the general SRM liability model. Section 3 discusses the definitions of harm, while section 4 the liability standards. Section 5 examines the performance of the various liability regimes. In section 6 the SRM liability model is implemented into the RCR model. Section 7 concludes.

2 The SRM Liability Model

I model SRM as a public good-or-bad which exhibits the free-driver characteristic. Besides the usual public good features of non-excludability and non-rivalry, being a public good-or-bad means that a marginal increase in the provision of the good-or-bad may be beneficial or harmful for the same agent, depending on amounts already provided. The free-driver characteristic implies that agents are heterogeneous in their preferences regarding the SRM provision level x and that SRM can be provided at negligible marginal costs.

These assumptions are reflected in the model set-up: I assume that there are n different agents and that each agent i has a well-defined and positive damage function

$$d_i(x),$$

depending on the SRM provision level x. Each damage function is convex and continuous in x. Furthermore, each damage function is increasing beyond some provision level. This implies that each agent i has a unique optimal SRM level x_i. In line with SRM's good-or-bad characteristic I assume that $x_i > 0$ for at least some agents.

I assume the social welfare criterion to be the minimization of total damages

$$\min_{x \in [0, x_n]} \sum_i d_i(x).$$

Due to the individual damage functions' characteristics this problem has a unique solution which is denoted by x^*. I assume that it is always the n-th agent who has the greatest incentives to provide SRM at the margin and that agent n is the sole SRM provider.[3]

There is a liability regime in place which makes the SRM provider pay for the harm she causes to other agents according to some liability function $L(x)$. The liability function determines the amount of compensation the SRM provider as to pay given her behavior. The SRM provider knows the liability function and minimizes her own damage function plus the liability function:

$$\min_{x \in [0, x_n]} \left[d_n(x) + L(x) \right].$$

The liability function $L(x)$ depends on two dimensions in this model, harm to other parties and the liability standard. The liability standard determines whether the SRM provider has to make liability payments to other parties. Harm determines the amount of compensation a party receives, in case the SRM provider has to compensate the party according to the prevailing liability standard.

[3]The domain in the minimization problem can be restricted because SRM levels beyond the free-driver outcome x_n are never optimal and no agent has an incentive to provide SRM beyond x_n.

2.1 Definition of Harm

There are two salient reference points for measuring SRM harm: The first one is the potential victim's condition without any SRM provision. I call this definition of harm the *absolute definition of harm*. The second one is the victim's optimal condition or preferred provision level, in other words, the level from which on SRM is indeed a bad for the agent in question. I call this definition of harm the *marginal definition of harm*.

According to the *absolute definition of harm*, an agent i is harmed by SRM if her damage level is above her damage level in the complete absence of SRM. The reference point here is the damage level at zero SRM provision, i.e. harm is

$$h_i^A(x) = \max\{0, d_i(x) - d_i(0)\}.$$

According to the the *marginal definition of harm*, an agent i is harmed if her damage level would be lower under some smaller SRM level than the actual one. The reference point here is the damage level at her optimal provision point x_i, i.e. harm is

$$h_i^M(x) = d_i(\max\{x, x_i\}) - d_i(x_i).$$

Since harm is always positive, the definition of harm only impacts on the internalization of the negative externality. In theory, a definition of harm could also be employed to internalize the positive externality of SRM provision, by allowing for negative harm for some provision levels x. Such 'negative liability' does not correspond to the institutional reality (Dari-Mattiacci 2009) and is therefore not considered in this paper.

2.2 Liability Standards

There are two traditional types of liability standards, *strict liability* and *negligence standards*. Under *strict liability*, the SRM provider has to compensate for any harm inflicted on any agent according to the prevalent definition of harm. Liability payments to be made by the SRM provider are then

$$L_{SL}(x) = \sum_{i \neq n} h_i(x).$$

Under negligence, the provider has to pay damages in accordance with the prevalent definition of harm, if she fails to meet a certain behavioral standard. The SRM provider's

behavior is characterized by the provision level x. In the law-and-economics literature, the behavioral standard is conceived as a level of (costly) precaution which reduces harm to other agents. Its standard economic interpretation is that it provides a balancing of the marginal harm and marginal costs of preventing harm (Posner 1972, Landes and Posner 1987). Translated into the context of SRM, the costs of refraining from increasing the SRM level are the forgone benefits in form of reduced damages to the SRM provider and, potentially, other agents. The costs of preventing harm are weighted against the prevented harm from not increasing the SRM level. Since there is SRM overprovision in absence of governance, the behavioral standard is conceptualized as a maximum level of SRM provision in this model. Liability payments then depend on the SRM level x chosen by the provider:

$$L_N(x) = \begin{cases} 0 & \text{if} \quad x \leq x_N \\ L_{SL}(x) & \text{if} \quad x > x_N \end{cases}$$

Here, x_N is the behavioral standard. If the SRM provider complies with the standard, she is absolved from liability. If she does not comply she has to pay for all harm caused, i.e. she faces liability payments equivalent to those under *strict liability*.

In traditional liability settings, in which a single injurer's actions unambiguously harm a single victim, there is only one way how the behavioral standard can trade off the two parties' interests. However, in the multi-agent context of the public good-or-bad SRM, there are several potential options for defining the behavioral standard. I give three different interpretations of the behavioral standard, one guided by the normative criterion of welfare maximization and two reflecting potential institutional realities.

From a normative welfare perspective, the weighting underlying the behavioral standard should reflect the consequences of the SRM provision level on all agents' welfare: This includes the harm inflicted on other parties, as well as the benefits, in form of damage reduction, conveyed to other parties as positive externality and the SRM provider's damage reduction. I call the behavioral standard emerging from this interpretation the *benefit-harm negligence* standard: This behavioral standard x_{BHN} is the unique solution[4] to

$$\min_{x \in [0, x_n]} \left[L_{SL}(x) + \sum_{i \neq n} d_i(\min\{x_i, x\}) + d_n(x) \right].$$

[4]There exists a unique solution since this is a continuous and convex optimization problem on a compact set.

Under *benefit-harm negligence*, the SRM provider has then to compensate either all victims or none, depending on whether she complied with the *benefit-harm negligence* standard or not.

While appealing from a normative point of view, the *benefit-harm negligence* standard, however, is likely to be incompatible with institutional reality. Consideration of effects on parties that are not part of the trial is generally not permissible in international law. This is likely to prevent third-party beneficiaries from being considered in the behavioral standard and makes the *benefit-harm negligence* standard unlikely to be employed in a real-world scenario.

Interpretations of negligence focusing on the parties harmed and the SRM provider are arguably more in line with institutional reality. Two different potential settings arise: In the first one the parties harmed sue jointly and are part of the same trial. In the second one they sue individually and there are separate trials for each party harmed. The former scenario suggests an interpretation of negligence under which the victims' harm is considered on aggregate in the weighting process. I call this standard the *aggregate harm negligence* standard. The behavioral standard x_{AHN} is defined as the unique solution to

$$\min_{x\in[0,x_n]} \Big[L_{SL}(x) + d_n(x) \Big].$$

Under *aggregate harm negligence* the SRM provider has then to compensate either all victims or none, depending on whether she complied with the *aggregate harm negligence* standard or not.

In the latter scenario, there are as many potential trials as potential victims. In each case the court balances the victim's harm individually with the SRM provider's damage reduction from increasing SRM provision. I call this the *individual harm negligence* standard under which there is a standard $x_{ILN}(i)$ for each potential victim i, where each standard the solution to the respective minimization problem

$$\min_{x\in[0,x_n]} \Big[h_i(x) + d_n(x) \Big].$$

Under *individual harm negligence*, the SRM provider has then to compensate victims on an individual basis, depending on whether she complied with the standard corresponding to the respective victim or not.

3 Assessment of the Definitions of Harm

For a liability regime to induce the socially optimal SRM provision level, it must make the SRM provider internalize the negative and the positive externalities on the other $n-1$ agents. Harm determines how large the compensation is which the SRM provider has to pay to victims, given that she has to compensate according to the liability standard. Since this compensation is always positive, the definition of harm only impacts on the internalization of the negative externality. I will now examine the marginal and *absolute definition of harm* with regard to their ability to be part of a SRM liability regime which internalizes the negative externality.

Fully internalizing the negative externality means that any welfare-reducing effect of further provision is reflected in the SRM provider's optimization problem. Under all liability regimes, the occurrence of harm is a necessary condition to award compensation. Any negative change in welfare to third parties can only be internalized by a liability regime to the extent that the negative change is reflected in what is understood to be harm, i.e. to the extent that there is a corresponding positive change in the prevalent definition of harm. In case of the *absolute definition of harm*

$$h_i^A(x) = \max\{0, d_i(x) - d_i(0)\},$$

there are settings in which a negative change in third parties' welfare does not correspond to an increase in harm: Assume that $x_i > 0$ for some agent i and that the current provision level is x_i. Consider a marginal increase in the provision level. Agent i will clearly be worse-off by this marginal increase, since x_i is her optimal provision point. However, given the *absolute definition of harm*, harm is only positive if agent i's damages are larger compared to the her damages without any SRM at all. Since her damages, given the provision level x_i, are even smaller than those in the complete absence of SRM, a marginal increase in the provision level cannot render her damages larger than those in complete absence of SRM. This effect disappears as soon as the actual damage $d_i(x)$ is larger than the initial damage level $d_i(0)$ in absence of SRM, in particular it is non-existent if the agent's optimal provision level x_i is zero. I denote the largest provision level such that absolute harm is zero for agent i by x_i^A. Furthermore, given a specific provision level x, I define the victim set at a provision level x as the set of agents for whom a marginal increase in the provision level is detrimental: $V(x) = \{i \mid x_i < x\}$.

I have just argued that for all agents with $x_i > 0$, there is a provision interval $[x_i, x_i^A]$

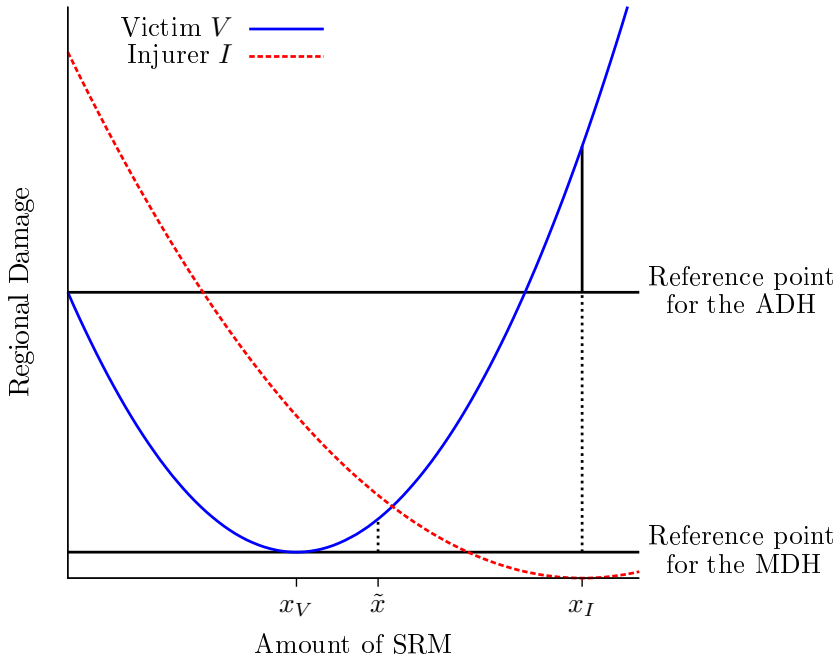

Figure 1: Reference points for the *marginal definition of harm* (MDH) and the *absolute definition of harm* (ADH). Victim V's damage function is in blue (solid curve), the injurer I's damage function is in red (dashed curve). The *marginal definition of harm* is represented by the combination of the dotted and the solid vertical lines. The *absolute definition of harm* is represented by the solid vertical line. At the provision level $\tilde{x}$ the *marginal definition of harm* is positive, while the *absolute definition of harm* is zero.

in which the agent i's marginal damage from SRM provision is positive, while marginal harm[5], given the *absolute definition of harm*, is zero. Therefore, the negative impacts on agent i from further provision in that interval can never be reflected in the SRM provider's private maximization problem by means of any liability regime employing the *absolute definition of harm*: Employing the *absolute definition of harm* introduces a bias towards too high SRM provision levels x in equilibrium.

In contrast, in case of the *marginal definition of harm*

$$h_i^M(x) = d_i(\max\{x, x_i\}) - d_i(x_i),$$

[5]The harm's derivative does not exist at all points. However, since the harm function is convex, the one-sided derivatives exist, in particular the right derivative. Throughout the paper I am interested in the changes of an increase in SRM, i.e. the right derivative. In cases in which the derivative does not exist, be it for the harm function or any other function, I mean "right derivative" when referring to the derivative or marginals.

marginal harm and marginal damage coincide for all agents in the victim set $V(x)$:

$$\frac{d}{dx}h_i^M(x) = \frac{d}{dx}d_i(x) \qquad \text{if} \quad x \geq x_i$$

Any negative change (and only negative changes) in third-party welfare is reflected in the *marginal definition of harm*. Therefore, employing the *marginal definition of harm* does not introduce a bias towards too high SRM provision levels x in equilibrium. Whether the negative changes in welfare on third parties is actually internalized by a specific liability regime employing the *marginal definition of harm* is then up to the specific liability rule employed.

Proposition 1.

1. *The absolute definition of harm does in general not fully reflect the negative externality from increases in the SRM provision x. For*

$$\min\{x_i \mid x_i > 0\} \leq x < \max\{x_i^A \mid i \neq n\},$$

the sum of marginal damages for agents in the victim set $V(x)$ is larger than the sum of marginal harm:

$$\sum_{i \in V(x)} \frac{d}{dx}d_i(x) > \sum_{i \in V(x)} \frac{d}{dx}h_i^A(x).$$

2. *The marginal definition of harm fully reflects the negative externality from increases in the SRM provision x. For all x, the sum of marginal damages for agents in the victim set $V(x)$ and the sum of marginal harm coincide:*

$$\sum_{i \in V(x)} \frac{d}{dx}d_i(x) = \sum_{i \in V(x)} \frac{d}{dx}h_i^M(x).$$

3. *No liability regime employing the absolute definition of harm can in general implement the socially optimal SRM provision level. If there are two liability regimes employing the same liability standard, one using the marginal definition and the other one using the absolute definition of harm, the former implements a (weakly) higher SRM provision level than the latter.*

The assumption that $x_i > 0$ for more than one agent is crucial for this result. In

substance, there is not any difference between the *marginal* and the *absolute definition of harm*, if this assumption is not fulfilled: If for all agents (except for the SRM provider) $x_i = 0$, we have a traditional setting of harm in which the first unit of an activity directly harms all potential victims. In such a setting, the distinction between the *marginal* and the *absolute definition of harm* becomes meaningless. The *marginal definition of harm* essentially reestablishes such a traditional setting of harm: By setting the reference point to the agent's optimal provision level, it ignores all changes in an agent's welfare before the public good-or-bad unambiguously becomes a bad for the agent in question. This allows the *marginal definition of harm* to reflect all negative changes in the victim's welfare.[6]

The *absolute definition of harm*'s bias does not imply that a specific liability standard in combination with the *absolute definition of harm* does always worse than the same standard in combination with the *marginal definition of harm*, since there is also a positive externality at play. If the positive externality is as well not (fully) internalized given the regime's liability standard, which would give rise to a bias in the opposite direction, the two biases (partially) cancel out. Which of the two liability regime then entails the larger bias is ambiguous and depends on the specific case at hand.

Proposition 1 is of importance for SRM compensation regimes more generally. Any compensation regime (e.g. insurance provided by the SRM provider) has to define a reference point used to ascertain the amount of compensation to be paid. The counterparts of the *marginal* and the *absolute definition* in such a compensation regime then have the same characteristics as those stated in proposition 1.

4 Assessment of the Liability Standards

In this section I discuss the liability standards, employing generic harm functions $h_i(x)$ which can stand for both definitions of harm. I denote the equilibrium SRM provision level under a liability standard S by $\hat{x}_S$. The equilibrium provision level also depends on the definition of harm. However, all statements made in this section, in particular statements about the equilibrium provision levels, hold for both definitions of harm. Liability standards can only make the SRM provider internalize the negative externality to the extent that it is reflected in the definition of harm $h_i(x)$. They can therefore only

[6]The distinction between the *marginal* and the *absolute definition of harm* is related to a legal and philosophical discussion on the nature of harm (Feinberg 1986, Perry 2003), which differentiates between a 'worsening' notion and a 'counterfactual' notion of harm.

internalize harm, but not the negative externality as such.

4.1 No Liability

In case of no liability, the SRM provider does neither face direct costs nor liability payments. Acting in self-interest, she provides SRM up to her personal optimum. SRM provision in equilibrium is then the free-driver outcome $\hat{x}_{FD} = \max_i x_i > x^*$.

4.2 Strict Liability

Under *strict liability*, liability payments are the sum of the individual agents' harm:

$$L_{SL}(x) = \sum_{i \neq n} h_i(x).$$

The SRM provider minimize the sum of her liability payments and her own damage:

$$\min_{x \in [0, x_n]} [L_{SL}(x) + d_n(x)].$$

At any provision level, the provider faces the trade-off between a marginal decrease in her own damages and a marginal increase in her liability payments. The liability payments reflect the increases in third-party harm and the SRM provider therefore internalizes the full harm externality. However, positive externalities are not captured in her minimization problem. Since the negative externality is fully captured, but the positive externality is not captured, *strict liability* carries a bias towards too low SRM provision levels. A liability regime employing *strict liability* can therefore in general not implement the socially optimal outcome x^*.

4.3 Negligence Rules

Negligence rules set a behavioral standard to which the provider must adhere in order to escape liability payments. The behavioral standard is some maximum SRM provision level x_N. The SRM provider faces damages of

$$L_N(x) = \begin{cases} 0 & \text{if} \quad x \leq x_N \\ L_{SL}(x) & \text{if} \quad x > x_N \end{cases}$$

and her minimization problem accordingly is

$$\min_{x\in[0,x_n]} \left[d_n(x) + \begin{cases} 0 & \text{if} \quad x \leq x_N \\ L_{SL}(x) & \text{if} \quad x > x_N \end{cases} \right]$$

If the SRM provider is better-off by complying with the standard compared to her optimal choice under *strict liability*, she will choose the provision level x_N in equilibrium.

Benefit-Harm Negligence

The *benefit-harm negligence* standard is guided by the normative approach of balancing costs and benefits of all agents. The behavioral standard x_{BHN} is defined as solution to

$$\min_{x\in[0,x_n]} \left[L_{SL}(x) + \sum_{i\neq n} d_i(\min\{x_i, x\}) + d_n(x) \right].$$

This minimization problem includes the same components as the SRM provider's minimization problem under *strict liability* plus the terms representing the positive externality. It therefore holds that the behavioral standard under *benefit-harm negligence* is larger than the equilibrium outcome under *strict liability*: $x_{BHN} \geq \hat{x}_{SL}$. Complying with the *benefit-harm negligence* standard, the SRM provider does not face liability payments and her damage level is $d_n(x_{BHN})$. Under *strict liability* she faces liability payments and her damage level is $d_n(\hat{x}_{SL})$. Since $x_n \geq x_{BHN} \geq \hat{x}_{SL}$, she is better-off by complying with the *benefit-harm negligence* standard. It follows that $x^*_{BHN} = x_{BHN} \geq x^*_{SL}$. Since the *benefit-harm negligence* standard takes into account both the positive and the negative externality, it is not biased and a liability regime employing this standard may implement the social optimal outcome x^* in general.

Aggregate Harm Negligence

The *aggregate harm negligence* standard reflects a setting in which all agents harmed jointly sue the SRM provider. In this setting all agents harmed are party to the trial and their harm is taken into consideration on aggregate. The resulting *aggregate harm negligence* standard x_{ALN} is defined as solution to

$$\min_{x\in[0,x_n]} \left[L_{SL}(x) + d_n(x) \right].$$

Since this is identical to the private minimization problem the SRM provider faces under *strict liability*, we have $x_{ALN} = x^*_{SL}$. It directly follows that the SRM provider chooses $x^*_{ALN} = x_{ALN} = x^*_{SL}$ in equilibrium in order to avoid paying damages. The *aggregate harm negligence* standard leads to the same outcome as *strict liability*: It carries a bias towards too low SRM provision levels and fully internalizes the negative externality while not capturing the positive externality at all. However, note that the *aggregate harm negligence* standard has other distributional effects: While the agents harmed receive compensation under *strict liability*, there are no liability payments under the *aggregate harm negligence* standard in equilibrium.

Individual Harm Negligence

The *individual harm negligence* standard reflects a setting in which agents harmed individually sue the SRM provider. In this setting there is an individual for each victim and their harm is taken into consideration individually. Therefore, there is an individual behavioral standard $x_{IHN}(i)$ for each agent i (except for the SRM provider) under the *individual harm negligence* standard. The individual standard for agent i is defined as solution to:

$$\min_{x \in [0, x_n]} \left[h_i(x) + d_n(x) \right].$$

For a given provision level x the liability payments are

$$L_{IHN}(x) = \sum_{i \neq n} l_{IHN}(i, x) \qquad \text{with}$$

$$l_{IHN}(i, x) = \begin{cases} 0 & \text{if} \quad x \leq x_{IHN}(i) \\ h_i(x) & \text{if} \quad x > x_{IHN}(i) \end{cases}.$$

The SRM provider's minimization problem then is

$$\min_{x \in [0, x_n]} \left[d_n(x) + \sum_{i \neq n} \begin{cases} 0 & \text{if} \quad x \leq x_{IHN}(i) \\ h_i(x) & \text{if} \quad x > x_{IHN}(i) \end{cases} \right].$$

Since there is an individual behavioral standard for each potential victim, the SRM provider will in general adhere to some of these behavioral standards and not to others. Since these behavioral standards only consider one victim's harm at a time, they fail to

internalize the harm of all other victims in their balancing process: Consider the smallest of the individual standards. At this standard's provision level, the marginal benefit to the SRM provider and the marginal harm to the victim in question are balanced, but the marginal harm to all other victims is neglected. Taking this harm to the other victims into account shows that the aggregate marginal harm at this provision level outweighs the SRM provider's marginal benefit. Therefore, the victims' harm is only partially internalized under the *individual harm negligence* standard. This implies that the *individual harm negligence* equilibrium provision level is larger than under *strict liability* and the *aggregate harm negligence* standard: $\hat{x}_{IHN} \geq \hat{x}_{AHN} = \hat{x}_{SL}$. However, not only the victims' harm is not fully internalized, but also the positive externality is not internalized at all under the *individual harm negligence* standard: The *individual harm negligence* standard is biased, but the direction of the bias is ambiguous in general. It is therefore also ambiguous whether the *individual harm negligence* equilibrium provision level is larger or smaller than the *benefit-harm negligence* equilibrium provision level. The answer to this depends both on the definition of harm and the agents' damage functions. Due to the standard's bias, a liability regime employing the *individual harm negligence* standard is in general not able to implement the socially optimal SRM provision level.

The results about liability standards are summarized in

Proposition 2.

1. *No liability regime employing one of the liability standards of strict liability, the aggregate harm negligence standard or the individual harm negligence standard can in general implement the socially optimal SRM provision level.*

2. *For both definitions of harm, it holds that*

$$\hat{x}_{FD} \geq \hat{x}_{BHN} \geq \hat{x}_{SL} = \hat{x}_{AHN} \quad \textit{and} \quad \hat{x}_{FD} \geq \hat{x}_{IHN} \geq \hat{x}_{SL} = \hat{x}_{AHN}.$$

   The ordering of $\hat{x}_{BHN}$ and $\hat{x}_{IHN}$ is ambiguous in general and depends on the agents' damage functions $d_i(x)$ and the prevalent definition of harm.

The *benefit-harm negligence* standard is the only liability standard which can be part of a liability regime which implements the socially optimal SRM provision level in general. However, as already mentioned, the *benefit-harm negligence* standard is unlikely to be employed in a real-world scenario, since consideration of effects on parties that are not part of the trial is generally not permissible in international law, which

is probably the most important body of law for SRM. Furthermore, the *benefit-harm negligence* standard imposes the highest informational requirements on a court, since for the determination of the standard the welfare of all regions would have to be considered.

5 Assessment of Liability Regimes

I now assess the performance of liability regimes, each consisting of a definition of harm and a liability standard. I denote the equilibrium SRM provision level under a liability regime consisting of liability standard S and definition of harm H by $\hat{x}_S(H)$. The behavioral standard corresponding to a negligence rule S in combination with a definition of harm H is accordingly denoted by $x_S(H)$.

The *marginal definition of harm* and the *benefit-harm negligence* standard both do not carry a bias. A liability regime employing those two components indeed succeeds in implementing the socially optimal SRM level: The social optimum x^* is defined as the solution to

$$\min_{x\in[0,x_n]} \sum_i d_i(x).$$

Given the *marginal definition of harm*

$$h_i^M(x) = d_i(\max\{x, x_i\}) - d_i(x_i),$$

the behavioral standard under the *benefit-harm negligence* standard is defined as the solution to

$$\min_{x\in[0,x_n]} \left[\sum_{i\neq n} (d_i(\max\{x, x_i\} - d_i(x_i))) + \sum_{i\neq n} d_i(\min\{x_i, x\}) + d_n(x) \right].$$

Since $d_i(\max\{x, x_i\}) + d_i(\min\{x, x_i\}) = d_i(x) + d_i(x_i)$, this is equivalent to

$$\min_{x\in[0,x_n]} \left[\sum_{i\neq n} d_i(x_i) + d_n(x) \right],$$

and therefore equivalent to the minimization problem which defines the social optimum. From the discussion of the *benefit-harm negligence* standard we know that the SRM provider adheres to that standard in equilibrium. It follows that $\hat{x}_{BHN}(M) = x_{BHN}(M) = x^*$.

Knowing the biases, or absence of biases, of the different definitions of harm and of the different liability standards, one can infer the performance of the other potential liability regimes relative to the social optimum.

Proposition 3.

1. *A liability regime employing the marginal definition of harm and the benefit-harm negligence standard implements the socially optimal SRM provision level in equilibrium.*

2. *In combination with the marginal definition of harm, strict liability and the aggregate harm negligence standard implement too low SRM provision levels in equilibrium compared to the social optimum, whereas the marginal definition of harm combined with the individual harm negligence standard may lead to too high or too low provision levels in equilibrium compared to the social optimum.*

3. *In combination with the absolute definition of harm, the benefit-harm negligence standard implements too high SRM provision levels in equilibrium compared to the social optimum, whereas strict liability, the aggregate harm negligence standard and the individual harm negligence standard in combination with the absolute definition of harm implement too high or too low SRM provision levels in equilibrium compared to the social optimum.*

Summary of Model Results

	Bias: Liability Standards	MDH	ADH
Bias: Definitions of Harm		o	+
BHN Standard	o	o	+
SL & AHN Standard	-	-	?
IHN Standard	?	?	?

Table 1: The second row and the second column report the biases for the definitions of harm and the liability standards, respectively, alone. The net bias for the liability regimes, each consisting of a definition of harm and a liability standard, are reported in the third and fourth column. An 'o' marks the absence of a bias, a '+' one towards too high SRM provision levels, a '–' one towards too low ones and a '?' marks a bias whose direction is ambiguous in general. Abbreviations: *Marginal definition of harm* (MDH); *Absolute definition of harm* (ADH); *Benefit-harm negligence* (BHN); *Aggregate harm negligence* (AHN); *Individual harm negligence* (IHN); *Strict Liability* (SL).

6 Numerical Implementation

I numerically implement the SRM liability model for two main reasons: Firstly, various studies have numerically examined the regional effects of SRM (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015), using regional mean temperature and regional mean precipitation as metrics. These studies have so far focused on Pareto-optimal and socially optimal SRM provision levels, finding that the socially optimal provision of SRM reduces damages at the regional level compared to climate changes damages substantially (Moreno-Cruz et al. 2012, Yu et al. 2015) and that Pareto-optimal SRM reduces regional damages considerably at least for the temperature metric (Kravitz et al. 2014). However, these studies ignored the underlying incentive structure. The free-driver's incentive to provide SRM up to her private optimum is well-established in the literature (Weitzman 2015, Heyen 2016). I quantify the free-driver problem in order to estimate the extent of the SRM governance problem.

Secondly, in the theoretical part of this paper I found that only one liability regime implements the social optimum. However, this regime employs the liability standard arguably least likely to be employed in the real world. All other liability regimes fail to implement the social optimum. Whether these liability regimes implement too much or too little SRM compared to the social optimum is often ambiguous, due to the presence of multiple biases, which potentially pull into opposing directions. The numerical implementation of the liability model might help illuminate how the performance of these non-optimal, but more likely to be employed, liability regimes compares to the free-driver outcome and the social optimum is, whether there are major differences in performance between these regimes and whether the choice of the definition of harm and the choice of the liability standard are equally important.

Moreno-Cruz et al. (2012) have developed a simple framework for evaluating regional effects of SRM, the Residual Climate Response (RCR) model. The RCR model uses quadratic regional damage functions in regional mean temperature and precipitation, with damages being minimal and normalized to zero at regional preindustrial conditions. These quadratic damage functions are one specific instance of the more general regional damage functions used in the theoretical part of this paper. Using preindustrial climate conditions as a baseline to evaluate regional SRM impacts, and thereby assuming that any deviation from that baseline inflicts damage, has been criticized as unrealistic in the literature (Heyen et al. 2015). However, since there is no obvious way which baseline

to employ instead, I follow the existing studies (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015) and hold on to using the preindustrial baseline.

6.1 The Residual Climate Response Model

The RCR model uses 22 geographic regions (as defined in Giorgi and Francisco 2000). The relevant climate metric is either regional mean temperature or regional mean precipitation. Let M be either of these metrics. A region's climate preferences are determined by a regional damage function. Regional damage is quadratic in the regional deviation $\Delta\mathcal{M}^i(x)$ from the preindustrial mean:

$$d_i(x) = -\Delta\mathcal{M}^i(x)^2.$$

These preferences imply that regional damage is lowest (i.e. zero) for preindustrial regional means.

The regional deviation from the preindustrial mean is the sum of the individual regional deviations due to climate change ($\Delta\mathcal{M}^i_{CO2}$) and SRM ($\Delta\mathcal{M}^i_{SRM}(x)$). Both of these deviations are normalized by preindustrial regional interannual variability $\sigma^i_{M,pre}$. The SRM provision level's impact is assumed to be linear[7]:

$$\Delta\mathcal{M}^i(x) = \Delta\mathcal{M}^i_{CO2} + \Delta\mathcal{M}^i_{SRM}(x) = \Delta\mathcal{M}^i_{CO2} + x \cdot \Delta\mathcal{M}^i_{SRM}.$$

$\Delta\mathcal{M}^i_{CO2}$ is the normalized difference between the pure climate change regional mean M^i_{CO2} and the preindustrial regional mean M^i_{pre}:

$$\Delta\mathcal{M}^i_{CO2} = \frac{M^i_{CO2} - M^i_{pre}}{\sigma^i_{M,pre}}.$$

$\Delta\mathcal{M}^i_{SRM}$ is the normalized difference between the regional mean M^i_{SRM} in the SRM climate, in which global mean temperature is restored to the preindustrial level, and the pure climate change regional mean M^i_{CO2}:

$$\Delta\mathcal{M}^i_{SRM} = \frac{M^i_{SRM} - M^i_{CO2}}{\sigma^i_{M,pre}}.$$

[7]Moreno-Cruz et al. (2012) and Kravitz et al. (2014) provide evidence for the reasonableness of this linear climate response assumption.

For all regions, M^i_{pre}, M^i_{CO2}, M^i_{SRM} and $\sigma^i_{M,pre}$ have to be calculated from climate model data. $\Delta\mathcal{M}_{CO2}$ is called the CO_2 vector and $x \cdot \Delta\mathcal{M}_{SRM}$ is called the SRM vector. For a given SRM provision level x, the residual vector $\Delta\mathcal{M}(x)$ contains all regions' normalized deviations from the preindustrial mean.

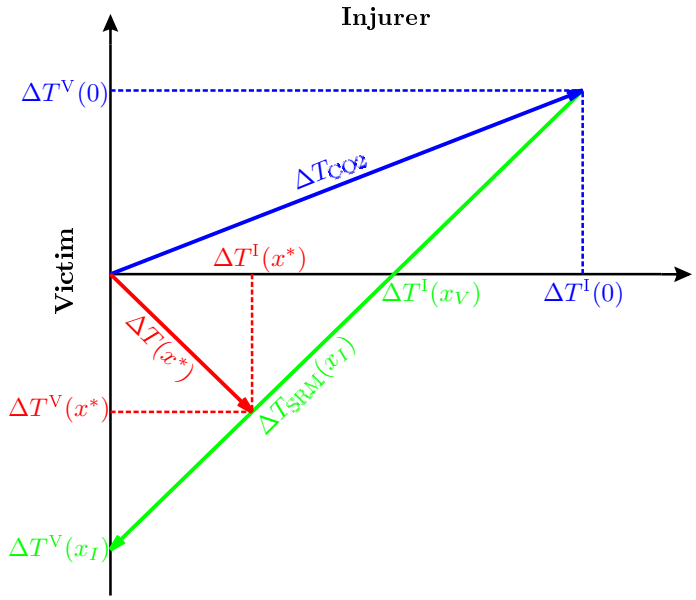

Figure 2: Residual Climate Response model. The horizontal axis shows changes in temperature for the injurer I, the vertical axis shows changes in temperature for the victim V. The blue CO_2 vector represents the temperature change due to climate change. The green SRM vector represents the temperature change due to SRM implemented according to the injurer's preferences. The red vector is the residual vector in the social optimum, pointing to regional temperatures under the socially optimal amount of SRM. Since regional damages are quadratic, the squared length of the residual vector represents the residual damages in the social optimum. In absence of governance, the injurer has incentives to provide SRM up to her preferred provision level x_I, the free-driver outcome, implying a larger residual vector than in the social optimum. The victim's preferred provision level x_V is attained at the intersection of the SRM vector with the horizontal axis.

The measure of global welfare is residual damages $D(x)$, i.e. the sum of regional residual damages $d_i(x)$, normalized to units of unmitigated climate change damages:

$$D(x) = \frac{\sum_i d_i(x)}{\sum_i d_i(0)}$$

The theoretical minimum of residual damages is zero (for preindustrial climate conditions), while residual damages for pure climate change conditions (zero SRM) are one.

Moreno-Cruz et al. (2012) use three different ways of weighting a region's damages. In this paper all regions' damages are accorded the same weight.

I use data from the G1 experiment as defined in the Geoengineering Model Intercomparision Project (Kravitz et al. 2011). The G1 experiment consists of a preindustrial model run, a pure climate change model run with elevated CO2 levels and a model run in which SRM is deployed on top of the elevated CO2 climate in order to restore the preindustrial global mean temperature. This implies that a SRM provision level of $x = 1$ in the model corresponds to restoring preindustrial global mean temperature. M^i_{pre}, M^i_{CO2}, M^i_{SRM} and $\sigma^i_{M,pre}$ can be calculated from the runs of the G1 experiment. I did so for each of the thirteen climate models participating in G1 individually and averaged the results. I then carried out the numerical implementation based on the averaged M^i_{pre}, M^i_{CO2}, M^i_{SRM} and $\sigma^i_{M,pre}$. I used the average across climate models, since the free-driver scenario reflects the strongest preferences for SRM and is therefore prone to outliers.

6.2 Results from the RCR Model

I report the equilibrium SRM level and the associated residual damages for the social optimum, for the free-driver outcome under no liability and for each liability regime. For comparison I also report the results for the Pareto optimum. I find an optimal SRM level of 0.99 for the temperature metric and of 0.80 for the precipitation metric. These optimal SRM levels entail residual damages of 0.2% and 5.1% of unmitigated climate change damages. Pareto-optimal SRM levels are 0.93 for the temperature metric and zero for the precipitation metric. These results are in line with the findings of Moreno-Cruz et al. (2012), Kravitz et al. (2014) and Yu et al. (2015).[8]

Concurrent research comes to the conclusion that the SRM governance problem might be substantial: Using an integrated assessment model approach for quantifying the free-driver outcome, Emmerling and Tavoni (2017) find SRM overprovision of a factor of eight. Using the simpler RCR model approach, I find that the the extent of the

[8]Moreno-Cruz et al. (2012) report residual damages of 1% for the temperature metric (independent of the weighting) and a range of $3\% - 15\%$ for the precipitation metric. Kravitz et al. (2014) and Yu et al. (2015) use data from the climate models participating in the G1 experiment. Yu et al. (2015) report residual damages of 0% (independent of the climate model) for the temperature metric and an average of 14% with a standard deviation of 14% for the precipitation metric in the social optimum. Note that in this study the results for the individual models were calculated and then averaged, while in the present paper the averaging is done for the parameters M^i_{pre}, M^i_{CO2}, M^i_{SRM} and $\sigma^i_{M,pre}$. For the median climate model, Kravitz et al. (2014) report a Pareto-optimal SRM level of 0.91 for the temperature metric and of zero for the precipitation metric.

free-driver problem depends on the metric chosen: For the temperature metric, there is moderate SRM overprovison in the free-driver outcome (13% higher compared to the social optimum; total damage is 1.6 percentage points of unmitigated climate change damages higher than in the social optimum) and SRM still reduces regional damages very effectively. However, for the precipitation metric, overprovision in the free-driver outcome is 362% compared to the social optimum and total damage is 658 percentage points of unmitigated climate change damages. While the inefficiencies due to the free-driver outcome are small for the temperature metric, these results suggest that the free-driver problem for SRM is devastating if mean precipitation is the relevant metric.

These results reflect the findings from earlier studies that regional differences in SRM effects are more pronounced for precipitation than for temperature (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015). The findings for the free-driver outcome confirm the direction of these results, but are more incisive: For the temperature metric SRM is even in the free-driver outcome very effective at reducing unmitigated climate change damages. However, SRM does not only become ineffective in the free-driver outcome under the precipitation metric, but it produces damages more than a factor 6.5 higher compared to unmitigated climate change. The reason that the difference in results for the two metrics in the free-driver outcome become so extreme is that the region with the strongest preferences for SRM determines the level of SRM. In this scenario the larger regional SRM disparities for precipitation have a much stronger effect on welfare, compared to scenarios in which the socially optimal or the Pareto-optimal SRM level are deployed.

Liability regimes employing the *marginal definition of harm* do very well for both the temperature and the precipitation metric. They reduce damages compared to the free-driver outcome very effectively, both in absolute and in relative terms: The only standards for which the absolute difference between the residual damages under a liability regime employing the *marginal definition of harm* and the damage level in the social optimum is larger than 0.1% percentage points of unmitigated climate change damages are *strict liability* and the *aggregate harm negligence* standard in combination with the precipitation metric. The absolute difference in residual damages here is 0.7% percentage points of unmitigated climate change damages. This difference corresponds to 13.7% higher damages under the liability regime compared to the social optimum. At the same time it corresponds to only 1‰ of the difference in residual damages between the free-driver outcome and the social optimum, which implies that the liability regime achieves

Panel A: Temperature Metric

	Marginal Definition of Harm		Absolute Definition of Harm	
	SRM Level	Damages	SRM Level	Damages
Social Optimum	0.99	0.2	0.99	0.2
Pareto Optimum	0.93	0.5	0.93	0.5
Free-Driver Outcome	1.12	1.8	1.12	1.8
BHN Standard	0.99	0.2	1.12	1.8
SL & AHN Standard	0.96	0.3	1.12	1.8
IHN Standard	0.96	0.3	1.12	1.8

Panel B: Precipitation Metric

	Marginal Definition of Harm		Absolute Definition of Harm	
	SRM Level	Damages	SRM Level	Damages
Social Optimum	0.81	5.1	0.81	5.1
Pareto Optimum	0.00	100	0.00	100
Free-Driver Outcome	2.93	658	2.93	658
BHN Standard	0.81	5.1	1.11	18.7
SL & AHN Standard	0.74	5.8	1.03	12.1
IHN Standard	0.80	5.1	1.12	19.6

Table 2: SRM levels are given as a fraction of the SRM level which restores global mean temperature to preindustrial. Residual damages in percent of unmitigated climate change damages. No SRM therefore corresponds to residual damages of 100%.

99.9% of the possible reduction in residual damages.

Liability regimes employing the *absolute definition of harm* do less well. For the temperature metric, liability regimes employing the *absolute definition of harm* are without any effect at all. This shows that even in the free-driver outcome there is no harm to other agents according to the *absolute definition*, given the temperature metric and the climate preferences assumed in the RCR model. However, for the precipitation metric, liability regimes employing the *absolute definition of harm* achieve a substantial reduction in damages: The absolute differences in residual damages between social optimum and liability regime range from 7.0% (SL and AHN), through 13.6% (BHN) to 14.5% (IHN). This corresponds to 137%, 167% and 184% higher damages under the respective liability regime compared to the social optimum and to 1.9%, 2.1% and 2.2% of the difference in

residual damages between the free-driver outcome and the social optimum. While the differences in damages to the social optimum are not trivial, all of these liability regimes achieve at least 97.8% of the possible reduction in residual damages.

The differences in outcomes across liability standards are comparatively small. For the temperature metric, the liability standards are irrelevant given the *absolute definition of harm*, since harm is then zero even in the free-driver outcome. In combination with the *marginal definition of harm*, the different liability standards still lead to almost the same outcomes. For the precipitation metric, residual damages under the *benefit-harm negligence* and the *individual harm negligence* standard are very similar: For the *marginal definition of harm*, the *benefit-harm negligence* standard implements the social optimum (residual damages of 5.1%) and the SRM equilibrium provision level under the *individual harm negligence* standard is close enough to the social optimum that the absolute difference in residual damages to the social optimum is smaller than 0.1%. For the *absolute definition of harm*, residual damages under the *benefit-harm negligence* standard are 18.7% and 19.6% for the *individual harm negligence* standard. The results for the precipitation metric confirms that the ordering in terms of SRM equilibrium provision level of the *benefit-harm* and the *individual harm negligence* standards is in general ambiguous. Under *strict liability* and the *aggregate harm negligence* standard, residual damages are somewhat higher for the *marginal definition of harm* (5.8%) compared to the other two standards, but substantially smaller for the *absolute definition of harm* (12.1%). The reason is that these two standards are biased towards too low SRM provision levels. Since the *marginal definition of harm* has no bias, the standards' bias drives the SRM equilibrium provision level away from the social optimum. However, the *absolute definition of harm* is biased towards too high SRM provision levels and the biases at play then partially cancel out.

The results show that, at least under the assumptions of the RCR model, liability regimes employing the *marginal definition of harm* do in every instance better than regimes employing the *absolute definition*. Furthermore, the only instance in which the performance between regimes employing different liability standards differs noticeably, is for the precipitation metric in combination with the *absolute definition of harm*. However, even in this case, the differences in residual damages between liability regimes employing different definitions of harm are larger than between liability regimes employing different liability standards. The results of the implementation therefore suggest that the choice of the definition of harm is more consequential for a liability regime's perfor-

mance than the choice of the liability standard and that the *marginal definition of harm* is generally superior to the *absolute definition of harm*. Liability regimes always lead to a significant reduction in residual damages with the exception of those employing the *absolute definition of harm* in case of the temperature metric. In particular, this means that liability regimes always achieve a significant reduction in residual damages under the precipitation metric, the case in which the free-driver outcome leads to devastating damage levels.

7 Conclusion

SRM is a set of techniques which has received increasing attention as potential means to offset climate change. Governance is a key issue for SRM, since it is likely to be cheap and would have regionally different impacts, giving rise to the "free-driver" problem (Weitzman 2015): In the absence of governance, the country with the strongest preferences for SRM has incentives to deploy SRM beyond the preferred provision point of all other countries. This paper focuses on liability regimes as a potential governance instrument. In the paper I developed a framework to understand the basic incentives SRM liability regimes provide. Furthermore, I implemented the model numerically in order to obtain a first-order estimate of the extent of the SRM governance problem and the capability of liability regimes to solve it in a simplified setting.

SRM is a public good-or-bad, a public good which benefits agents at some levels and harms the same agents at other levels. This feature sets SRM apart from more traditional domains of liability. The public good-or-bad characteristic is in two ways relevant for the incentives a liability regime provides. The first one concerns the definition of harm, which is about which SRM impacts have to be compensated for. The second one concerns the liability standards, which are about the circumstances under which harm from SRM has to be compensated for. The liability model of SRM in this paper puts the definition of harm and the liability standards center stage in order to focus on the specific incentives arising for a SRM provider from SRM's good-or-bad characteristic under a liability regime. A liability regime in the model consequently consists of a definition of harm and a liability standard.

I give two definitions of harm. As liability standards I consider *strict liability* and three interpretations of the negligence standard. Only one definition of harm, the *marginal definition*, and only one liability standard, the *benefit-harm negligence* stan-

dard, are unbiased. Therefore, only the liability regime employing the *marginal definition of harm* and the *benefit-harm negligence* standard implements the social optimum. However, the *benefit-harm negligence* standard is the one least likely to be employed in a real-world scenario due to the legal institutional reality. All other liability regimes do in general not implement the social optimum and carry a bias towards too low or high SRM provision levels. The direction of this bias is often ambiguous, since a regime's net bias is generally the result of multiple biases which may pull into opposing directions. This highlights the difficulties in deciding which liability regime to pick in a real-world scenario and shows that the choice of one component of a liability regime should in general depend on the other component. Lastly, it should be noted that the results for the definition of harm are of relevance for any SRM compensation mechanism.

I numerically implement the theoretical model into the Residual Climate Response (RCR) model (Moreno-Cruz et al. 2012), a framework for investigating regional impacts of SRM, using climate model data on regional mean temperature and precipitation from the G1 experiment of the Geoengineering Model Intercomparision Project (Kravitz et al. 2011). It should be kept in mind that the RCR model is a simple framework and that the results from using it serve as first-order estimation of the effects examined. Concurrent research comes to the conclusion that the SRM governance problem might be substantial: Using a more sophisticated integrated assessment model approach for quantifying the free-driver outcome, Emmerling and Tavoni (2017) find SRM overprovision of a factor of eight. Using the simpler RCR model approach, I find that the the extent of the free-driver problem depends on the metric chosen: For the temperature metric, there is moderate SRM overprovison in the free-driver outcome and SRM still reduces damages in the free-driver outcome down to 1.8% of unmitigated climate change. It is extreme for the mean precipitation metric and SRM increase damages by more than a factor of 6.5 in the free-driver outcome compared to unmitigated climate change. These findings suggest that, from an economic point of view, the SRM governance problem is very severe in case precipitation is the relevant metric, but rather benign in case temperature is the relevant metric. This reflects earlier findings (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015) that regional SRM differences are more pronounced for precipitation than for temperature. The difference between the two metrics is amplified in the free-driver outcome, since the region with the most extreme preferences for SRM then determines the SRM provision level.

Liability regimes lead to a welfare-enhancing implementation of SRM for the pre-

cipitation metric: All regimes reduce damages to at most 19.6% of unmitigated climate change damages for this metric. Liability regimes employing the *marginal definition of harm* virtually implement the social optimum for both metrics. For the temperature metric, the *absolute definition*'s bias renders liability regimes without any effect at all. Differences in outcomes across the definition of harm are larger than differences in outcomes across liability standards and the *marginal definition of harm* always performs better than the *absolute definition*. Given the assumptions of the RCR model, all liability regimes drastically mitigate the extreme free-driver problem found for the precipitation metric, the *marginal definition of harm* is generally superior to the *absolute definition* and the choice of the definition of harm is of greater importance than the liability standard for the performance of a SRM liability regime.

This paper has focused on the specific incentives liability regimes provide in light of SRM's good-or-bad characteristic. I therefore abstracted from various other SRM aspects, which are important, but do not lie at the heart of the SRM-specific incentives liability regimes provide. These aspects include uncertainty about SRM impacts, other potential SRM side-effects like ozone loss or potential health impacts and potential coalitions among agents (compare Ricke et al. 2013). Furthermore, I abstracted from issues of causation. There are literatures dealing both with issues of causation in the context of SRM (Horten et al. 2014 and Saxler et al. 2015) and with the general law-and-economics implications of uncertain causation (Shavell 1985). Lastly, the model presupposes an existing SRM liability regime. Currently, there is no dedicated SRM liability regime in place and it is far from clear whether there will be such a regime in the future. In any case, even in the absence of a dedicated SRM liability regime, customary international law may provide a legal basis for SRM liability (Saxler et al. 2015).

There are valuable extensions future research could pursue. The first potential extension concerns the formation of coalitions. Agents who benefit greatly from SRM could decide to form a coalition (Ricke et al. 2013) in order to provide SRM jointly, while sharing the expected liability payments, thereby partly internalizing the positive externalities from SRM provision. Taking coalitions into account has therefore the potential to alter the assessment of the liability regimes presented in this paper. The second potential extension is the consideration of treaty formation, asking the questions of whether and under which conditions a liability regime could emerge as the result of a negotiation and bargaining process. The framework presented in this paper and its insights regarding

the incentives potential liability regimes provide are ideal starting points for approaching these two extensions.

Lastly, future research could focus on extending the RCR model. At the moment, agents in the RCR model have preferences for a preindustrial climate, an assumption which does not seem to be very realistic (Burke et al. 2015, Heyen et al. 2015). An extension of the RCR model, which allows for climate preferences diverging from preindustrial climate conditions, would be a valuable contribution for the assessment of regional SRM impacts in general and as a result also for the assessment of the performance of SRM liability regimes.

Appendix

Proof of Proposition 1. Only the second statement of the third part remains to be shown. Liability standards are defined via the balancing of a subset of victims' harm and a subset of beneficiaries' benefits (including the SRM provider). At any x and for any subset of victims the respective sum of marginal harm is weakly smaller for the *absolute definition of harm* than for the *marginal definition of harm*. Therefore, the SRM provider is for a given liability standard at a given provision level x never liable given the *absolute definition of harm* if she is not liable given the *marginal definition of harm*. If she is not liable given the *absolute definition of harm*, her marginal costs of SRM provision are zero under the *absolute definition*. If she is liable, she is also liable under the *marginal definition* and marginal liability payments, her marginal costs of SRM provision, are weakly higher given the *marginal definition* than given the *absolute definition*. □

Proof of Proposition 2. Only the second statement of the second part remains to be shown. Consider two settings in which the *marginal definition of harm* is the relevant definition and in which there are four agents. Setting 1: $d_1(x) = 0.5x^2$; $d_2(x) = 0.5(7.5-x)^2$; $d_3(x) = 0.5(5-x)^2$; $d_4(x) = 0.5(10-x)^2$. Here, we have $\hat{x}_{BHN} = 5.625$ and $\hat{x}_{IHN} = 5$. Setting 2: $d_1(x) = 0.5x^2$; $d_2(x) = 0.5(2.5-x)^2$; $d_3(x) = 0.5(5-x)^2$; $d_4(x) = 0.5(10-x)^2$. Here, we have $\hat{x}_{BHN} = 4.325$ and $\hat{x}_{IHN} = 5$. □

Proof of Proposition 3. Only the statements about the SL, the AHN and IHN standards in part three remain to be shown. Setting 1: $d_1(x) = 0.5(5-x)^2$; $d_2(x) = 0.5(10-x)^2$. Absolute harm here is zero even for $x = 10$. All liability regimes employing the *absolute definition* therefore implement a too high SRM provision level. Setting 2: $d_1(x) = 0.5x^2$; $d_2(x) = 0.5(8-x)^2$; $d_3(x) = 0.5(10-x)^2$. Here, all three liability standards in combination with the *absolute definition of harm* implement a SRM provision level of 5, which is below the socially optimal level $x^* = 6$. □

References

[1] Barrett, S. (2008). "The incredible economics of geoengineering". In: *Environmental and resource economics* 39.1, pp. 45–54.

[2] Burke, M., Hsiang, S. M., and Miguel, E. (2015). "Global non-linear effect of temperature on economic production". In: *Nature* 527.7577, pp. 235–239.

[3] Emmerling, J. and Tavoni, M. (2017). "Quantifying non-cooperative climate engineering". In: *FEEM Working Paper Series* 058.2017. Available at https://ssrn.com/abstract=3090312.

[4] Feinberg, J. (1986). "Wrongful life and the counterfactual element in harming". In: *Social Philosophy and Policy* 4.1, pp. 145–178.

[5] Giorgi, F. and Francisco, R. (2000). "Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM". In: *Climate Dynamics* 16.2-3, pp. 169–182.

[6] Heyen, D. (2016). "Strategic conflicts on the horizon: R&D incentives for environmental technologies". In: *Climate Change Economics* 7.04, p. 1650013.

[7] Heyen, D., Wiertz, T., and Irvine, P. J. (2015). "Regional disparities in SRM impacts: the challenge of diverging preferences". In: *Climatic Change* 133.4, pp. 557–563.

[8] Horton, J. B., Parker, A., and Keith, D. (2014). "Liability for Solar Geoengineering: Historical Precedents, Contemporary Innovations, and Governance Possibilities". In: *NYU Envtl. LJ* 22, p. 225.

[9] Irvine, P. J., Ridgwell, A., and Lunt, D. J. (2010). "Assessing the regional disparities in geoengineering impacts". In: *Geophysical Research Letters* 37.18.

[10] Keith, D. W., Parson, E., and Morgan, M. G. (2010). "Research on global sun block needed now". In: *Nature* 463.7280, pp. 426–427.

[11] Kravitz, B., MacMartin, D. G., Robock, A., Rasch, P. J., Ricke, K. L., Cole, J. N., Curry, C. L., Irvine, P. J., Ji, D., Keith, D. W., et al. "A multi-model assessment of regional climate disparities caused by solar geoengineering". In: *Environmental Research Letters* 9.7: 074013.

[12] Kravitz, B., Robock, A., Boucher, O., Schmidt, H., Taylor, K. E., Stenchikov, G., and Schulz, M. (2011). "The geoengineering model intercomparison project (GeoMIP)". In: *Atmospheric Science Letters* 12.2, pp. 162–167.

[13] Landes, W. M. and Posner, R. A. (1987). *The economic structure of tort law*. Harvard University Press.

[14] Lunt, D. J., Ridgwell, A., Valdes, P. J., and Seale, A. (2008). ""Sunshade World": A fully coupled GCM evaluation of the climatic impacts of geoengineering". In: *Geophysical Research Letters* 35.12.

[15] Moreno-Cruz, J. B., Ricke, K. L., and Keith, D. W. (2012). "A simple model to account for regional inequalities in the effectiveness of solar radiation management". In: *Climatic change* 110.3-4, pp. 649–668.

[16] Pasztor, J. (2017). "The Need for Governance of Climate Geoengineering". In: *Ethics & International Affairs* 31.4, pp. 419–430.

[17] Perry, S. (2003). "Harm, history, and counterfactuals". In: *San Diego L. Rev.* 40, pp. 1283–1314.

[18] Posner, R. A. (1972). "A theory of negligence". In: *The Journal of Legal Studies* 1.1, pp. 29–96.

[19] Rayner, S., Heyward, C., Kruger, T., Pidgeon, N., Redgwell, C., and Savulescu, J. (2013). "The oxford principles". In: *Climatic Change* 121.3, pp. 499–512.

[20] Reynolds, J. L. (2015). "An Economic Analysis of Liability and Compensation for Harm from Large-Scale Field Research in Solar Climate Engineering". In: *Climate Law* 5.2-4, pp. 182–209.

[21] Ricke, K. L., Moreno-Cruz, J. B., and Caldeira, K. (2013). "Strategic incentives for climate geoengineering coalitions to exclude broad participation". In: *Environmental Research Letters* 8.1: 014021.

[22] Ricke, K. L., Morgan, M. G., and Allen, M. R. (2010). "Regional climate response to solar-radiation management". In: *Nature Geoscience* 3.8, pp. 537–541.

[23] Robock, A., Oman, L., and Stenchikov, G. L. (2008). "Regional climate responses to geoengineering with tropical and Arctic SO2 injections". In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 113.D16.

[24] Saxler, B., Siegfried, J., and Proelss, A. (2015). "International liability for transboundary damage arising from stratospheric aerosol injections". In: *Law, Innovation and Technology* 7.1, pp. 112–147.

[25] Shavell, S. (1985). "Uncertainty over causation and the determination of civil liability". In: *The Journal of Law and Economics* 28.3, pp. 587–609.

[26] Shepherd, J. G. (2009). *Geoengineering the climate: science, governance and uncertainty*. Royal Society.

[27] Weitzman, M. L. (2015). "A Voting Architecture for the Governance of Free-Driver Externalities, with Application to Geoengineering". In: *The Scandinavian Journal of Economics* 117.4, pp. 1049–1068.

[28] Yu, X., Moore, J. C., Cui, X., Rinke, A., Ji, D., Kravitz, B., and Yoon, J.-H. (2015). "Impacts, effectiveness and regional inequalities of the GeoMIP G1 to G4 solar radiation management scenarios". In: *Global and Planetary Change* 129, pp. 10–22.