Four Essays in the Law and Economics of Novel Technologies

Dissertation

zur Erlangung des akademischen Grades
doctor rerum politicarum

an der Fakultät für Wirtschafts- und Sozialwissenschaften
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von
Tobias Oliver Pfrommer
geboren am 13. Oktober 1983 in Tettnang

im Juni 2018

Acknowledgments

Over the years of writing this dissertation, I have received support from many different people in many different ways and on many different levels. To each of them I am deeply grateful.

First and foremost, I would like to thank Prof. Timo Goeschl for supervising this thesis, for being an inspiring co-author, for helping me to develop my thoughts and for great support.

Furthermore, I am grateful to Prof. Martin Quaas for serving as the second referee for this dissertation.

I thank my co-authors from the research projects CEIBRAL and CELARIT for many inspiring discussions and for providing me with fascinating insights into their respective fields.

I want to thank Johannes Diederich for being an enjoyable officemate and for great discussions. Thanks also go to my other current and former colleagues at the Chair of Environmental Economics for help, advice and fun. Specifically, I would like to thank Daniel Heyen and Johannes Lohse, who are not only insightful colleagues, but also great friends.

My friends in Heidelberg and elsewhere enrich my life in the most different of ways. Thank you all for many memorable moments, for fun and laughter, for listening, for cheering me up and simply for being part of my life.

The house I shared during the years of writing this dissertation, Villa 47, is a very special place to me. I want to thank all my housemates over the years and all the many others, friends and acquaintances, who make Villa 47 such a wonderful place.

A big thank you goes to my parents and my family for their continuous support, patience and caring, as well as for providing me with many of the opportunities I had and have in life.

Finally, I wish to thank my dear friend Agnes for her support, for great fun, for caring, for companionship and for believing in me when I did not. Without her, this dissertation would never even have been started.

Contents

Chapter 1

Introduction

Introduction

Novel technologies have changed and continue to change modern life. The pervasive use of industrial chemicals, advances in human medicine and the rises of biotechnology and nanotechnology are just some prominent examples which have changed human life on many levels and provide enormous benefits to society. At the same time, novel technologies often involve unknown risks – a fact to which the victims of the morning sickness drug thalidomide[1] or the early use of X-radiation can testify. In the risks and benefits of novel technologies lies a tension that society has to resolve one way or the other.

When agents bear both the full risks and benefits of their actions, the tension between risks and benefits is not societal in nature – it is rather one to be addressed by individual decision-making. In contrast, when the risks fall on other agents, as is often the case, the problem of risk externalities emerges. The traditional law and economics approach to such problems is to analyze the incentives that the relevant agents face, the information they possess and the transaction costs involved under different legal institutions. Law and economics sees legal matters through the lens of resource allocation and one of its central tenets is that institutional evolution is towards efficiency. On this basis, law and economics derives positive and normative conclusions about behavior and institutional design. While law and economics scholars acknowledge the existence and relevance of other rationals for legal rules, involving notions of justice, fairness, responsibility, equity and morality, these rationals are not the subject of investigation in law and economics (Calabresi 1970).

The law and economics approach has proven fruitful for understanding and describing the structure of legal institutions. Coase (1960) observed that in the absence of transaction costs the only relevant institution for resolving problems of allocation is property rights. With that observation in mind, it is not surprising that other means to decide on legal entitlements usually only arise in settings in which transaction costs are not trivial (Calabresi 1968, Calabresi and Melamed 1972). The economic role of these other means is to decide on legal entitlements in a way such that the same outcome emerge as it would in the absence of transaction costs. A classic example of substantial, and usually prohibitive, transaction costs are risk externalities. For example, the driver of a car cannot in any meaningful sense negotiate ex-ante with pedestrians about how much and how fast she drives. Liability is one of the two main instruments for regulating risk externalities and, as the reader will find out, the instrument that will be the main

[1] Thalidomide was first marketed in Germany under the trade-name Contergan. It turned out that, when taken during pregnancy, the drug causes birth defects.

subject of this dissertation.[2] In the following, I will therefore briefly outline the basic approach and findings in the law and economics analysis of liability rules.

There are two basic types of liability rules, strict liability and negligence rules. Strict liability always holds the injurer liable for the harm she causes. In contrast, negligence only holds the injurer liable if she was negligent, i.e. if she did not conform to a certain behavioral standard. The economic purpose of liability is to reduce total accident costs. The total costs of accidents consist of both the direct costs of accidents and the precaution costs (or care costs) incurred in order to avoid accidents (Calabresi 1970). The implication is that the optimal number of accidents is not zero and that there is a finite socially optimal level of precaution. When the behavioral standard is defined in terms of the socially optimal level of precaution, both rules incentivize the injurer to take socially optimal care (Calabresi 1970, Posner 1972, Brown 1973).[3]

There are two parts to this dissertation. The first part examines the relationship between liability rules and novel technologies from a general perspective. The second part concentrates on a specific novel technology, solar geoengineering. I will introduce the relevant context for each part's contributions in sequence, starting with the first part.

The analysis of liability rules is usually undertaken in static settings. Novel technologies, by their very nature, elude a static analysis and characterization. In contrast, their novelty introduces a dynamic relationship between liability regimes and the technologies they govern. Law and economics has investigated this dynamic relationship mainly from two related angles. The first is an innovation perspective, where innovation can either refer to innovation in safety technology or to product innovation. With respect to the former, Endres and Bertram (2006) analyze the relationship between liability rules and endogenous investments in care costs reducing technology. Endres and Friehe (2011) study the incentives for the diffusion of such a technology. Parchomovsky and Stein (2008) highlight the disincentive on the innovation and adoption of care cost reducing technologies originating from the role of custom in determining negligence in court. This literature mainly finds that strict liability provides optimal incentives, while negligence does not. With regard to product innovation, there is an empirical literature which concludes that high liability costs are detrimental to product innovation (Viscusi and Moore 1993, Finkelstein 2004). Immordino et al. (2011) employ a model in which investment into product innovation can produce either safe or harmful innovation and in which risk information is private. They find that both direct and liability regulation reduce investment in product innovation as regulation becomes more stringent. The

[2]Direct regulation is the other. While some risks are usually governed by direct regulation, other risks are usually governed by liability. Shavell (1984a, 1984b) analyzes the use of direct regulation and liability along four dimensions, drawing conclusions about the settings in which each is usually applied.

[3]In general, both injurer and victim may be able to reduce accident risks by taking precautions. Whether liability rules should financially burden the injurer or victim in specific settings then becomes a question of identifying the 'least cost avoider'. However, such settings are not examined in the articles contained in this dissertation.

bigger picture suggests that stringent liability regimes impede product innovation, but not the innovation and diffusion of safety technology.

Learning about uncertain risks is the second angle from which law and economics has investigated the dynamic relationship between liability regimes and novel technologies: Risks emerging from novel technologies are uncertain and poorly understood. Shaping the incentives for the potential use of novel technologies, liability regimes influence if and how novel technologies are adopted. Consequently, a dynamic interaction arises between how liability regimes govern novel technologies and the learning that takes place about novel technologies' potential risks. The classic approach in the literature is to consider a learning opportunity before the technology might be used (Shavell 1992, Ben-Shahar 1998). The classic result is that when injurers can acquire perfect information at a fixed cost, strict liability provides adequate incentives. A recent contribution by Baumann and Friehe (2016) employed a learning-by-doing mechanism. They find that strict liability is not necessarily optimal anymore in such a setting. The first part of this dissertation (article 1) contributes to this literature on the relationship between liability regimes and learning. Its main contribution is the introduction of a learning mechanism operating at the post-market stage that can explain the widespread use of exemptions for harm from novel risks found in the US, the EU and beyond.[4]

The second part of this dissertation investigates the relationship between liability and a specific novel technology – solar geoengineering. In order to provide for the proper context, I will first give a brief general introduction to the topic. Climate change poses substantial risks to many aspects of human life (IPCC 2014), but mitigation efforts have so far been insufficient for meaningfully curbing future climate change risks. In that respect, the 2015 Paris Agreement was widely celebrated as a significant step into the right direction. However, an initial assessment yields a warming of 2.6°C to 3.1°C in 2100, based on the Intended Nationally Determined Contributions submitted by the participating States (Rogelj et al. 2016). Furthermore, since the climate response is uncertain, "it is also possible that substantially higher temperatures will materialize with compelling likelihoods" (Rogelj et al. 2016). The perils of such 'catastrophic' climate change, when temperature increases realize at the upper end of the probability range, have been the subject of substantial discussion (Weitzman 2009, Nordhaus 2011, Pindyck 2011, Weitzman 2014).

In the face of these bleak prospects, discussions about geoengineering have gained more and more traction. Geoengineering is often defined as the "deliberately large-scale manipulation of the planetary environment to counteract anthropogenic climate change" (Shepherd 2009). Geoengineering approaches fall into two categories, carbon dioxide removal (CDR) and solar geoengineering (SG). CDR methods aim at lowering atmospheric greenhouse gas concentrations. While CDR methods are very likely to play an important role in climate policy, they are in many respects similar to mitigation efforts

[4] A more detailed summary of the article follows later on.

(Klepper and Rickels 2012) and not the subject of this dissertation. SG methods aim at reducing the amount of solar radiation absorbed by the earth. SG methods have the potential to quickly (Keith et al. 2010) and substantially (IPCC 2013) change global temperatures. 'Stratospheric aerosol injection' is the most widely discussed SG method (Keith 2013). Although SG methods have not been developed yet, observations following volcanic eruptions, e.g. Mount Pinatubo in 1991, confirm that the method of stratospheric aerosol injection works in principle (Crutzen 2006).

Potential risks from SG are a widespread concern, one of the most prominent being that SG changes the hydrological circle (Trenberth and Dai 2007, Robock et al. 2008, Kravitz et al. 2013). Among the relevant implications are reduced precipitation and an increased potential for droughts in some regions (Trenberth and Dai 2007, Robock et al. 2009). Furthermore, the method of stratospheric aerosol injection is known to lead to ozone depletion (Tilmes et al. 2008, Keith et al. 2010, Pitari et al. 2014). Generally, unanticipated risks from SG are a serious concern (Keith et al. 2010) and there remain large gaps in knowledge regarding the effects of SG (MacMartin et al. 2016). As is the case with other novel technologies, there is the potential that some risks remain unanticipated, irrespective of the amount of research prior to potential use, rendering any large-scale solar geoengineering deployment inextricably connected to the potential of unanticipated risks and side effects.

In light of these substantial risks, governance is an overarching topic in the SG literature (Barrett 2008, Shepherd 2009, Rayner et al. 2013, Parker 2014, Pasztor 2017). Liability regimes, specifically, have received considerable attention in the debate. Horton et al. (2014) propose liability regimes as a potential means of SG governance, drawing comparisons to the Space Liability Convention[5] which they invoke as a successful precedent for such a regime. Saxler et al. (2015) discuss the suitability of existing international liability regimes for governing SG. They conclude that existing international liability treaties do not cover SG, but that these treaties can offer valuable guidance to a SG regime. Lastly, they reference international customary law as a potential legal source for SG liability governance. Reynolds (2015) analyzes liability as an instrument specifically for governing large-scale SG field research from a mainly economic perspective, taking into account the public good character of such research. A common point of concern in the literature is causation. All three studies cited in this paragraph stress that establishing the causal link between a SG intervention and a potentially ensuing harm presents a major challenge to any SG liability regime: Due to the stochastic nature of the climate system, the traditional legal 'but for' test of deterministic causation is infeasible (Allen et al. 2007). Legal attribution based on probabilistic notions appears to be the only possible option. However, this approach presents a considerable challenge from a legal perspective (Saxler et al. 2015).

[5]The convention's full name is 'Convention on International Liability for Damage Caused by Space Objects'. More information can be retrieved from `http://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/introliability-convention.html`.

The substantial risks of SG are exacerbated by the unique incentive structure underlying potential SG deployment. On the one hand SG methods are likely to be very cheap and can probably be implemented by a single actor (Barrett 2008). On the other hand they have regionally diverging impacts (Ban-Weiss and Caldeira 2010, Ricke et al. 2010). The combination of these two factors gives rise to what Weitzman (2015) coined the free-driver incentive structure: In absence of governance, the agent or region with the strongest preferences for SG determines the amount of SG deployed, with detrimental effects on all other agents or regions. However, only the last units of SG are a cost to all other agents. Since SG represents technologies which aim at curbing climate change risks, the initial units of SG provided by the 'free-driver' benefit at least a decent number of other agents. This implies that SG provision by a free-driver is both a 'good' and a 'bad' to at least some of the agents. From a law and economics perspective, such a setting provides for an interesting research opportunity: It is far from clear how one should interpret the concept of harm in such a setting. Since the concept of harm is usually taken for granted in law and economics, this setting presents a challenge to the existing law and economics literature.

A literature quantitatively examining the extent of regional differences in SG impacts is of general relevance to SG liability, and more generally, to SG governance. Moreno-Cruz et al. (2012) developed the Residual Climate Response (RCR) model for assessing such regional differences which was subsequently used in the literature (Kravitz et al. 2014, Yu et al. 2015). The common finding is that regional differences in temperature are small to moderate, while regional differences in precipitation are substantial. However, Heyen et al. (2015) criticize the common assumption of these studies that regional climate preferences are determined by a historic climate, showing that regional differences may also be substantial for temperatures when relaxing that assumption. The literature is relevant to SG liability and SG governance for two reasons. Firstly, the RCR model can be used for numerically investigating strategic SG settings: Ricke et al. (2013) employ the RCR model in order to construct a 'coalition game', based on the strategic incentives to build coalitions, when SG deployment requires a minimum amount of international power. Secondly, the literature may give more structure to the governance problem beyond the free-driver characterization in identifying specific ways and dimensions in which regional SG impacts are heterogeneous: A well-established result from climate science is that SG undercompensates for climate change driven temperature increases at high latitudes, while overcompensating at low-latitudes (Ban-Weiss and Caldeira 2010, Ricke et al. 2010).

The second part of this dissertation (articles 2, 3 and 4) contributes to various aspects important to SG liability. Article 2 examines the challenge of causation in climate litigation in an interdisciplinary research effort. In particular, article 2 takes into account the incentive structure of evidence production in a SG liability trial and its repercussions for overcoming the challenge of causation. Article 3 provides an analytical model of SG

liability, considering the unique SG incentive structure. Article 4 extends the Regional Climate Response model, introducing climate preferences diverging from historic climate conditions. In that, article 4 sheds further light on structural dimensions of regional SG disagreement, contributing to the structural understanding of SG governance problems.

The two parts of the dissertation jointly contribute to the law and economics of novel technologies. The first part contributes to the general understanding of the relationship between novel technologies and liability rules in that it furthers the understanding of learning about novel and uncertain risks in liability contexts (article 1).[6] The second part concerns the specific novel technology of solar geoengineering. In providing a model of SG liability, which takes into account the unique incentive structure of the setting, it contributes to the law and economic literature on liability rules (article 3). Beyond that, it extends the understanding of the structure of regional SG disagreement, providing valuable input for the analysis of SG liability and SG governance (article 4). Lastly, it provides an analysis of the challenge of causation in the specific context of climate litigation (article 2).[7] This article links the the two parts of the dissertation, in that it takes up an issue of general importance in the liability law (more specifically, the tort law) of novel risks and analyzes it in the specific context of climate litigation.

Synopsis

The first article **"Experimentation, Torts, and the Value of Information"** (with Timo Goeschl) focuses on the relationship between liability regimes and experiential learning at the post-market about the potential risks emerging from novel technologies. The article is motivated by a startling observation. Legal rules usually provide for exemptions for harm arising from novel, uncertain or unknowable risks. Such 'state-of-the art defenses' are "in principle available in all but a very small number of jurisdictions" in the world (Reimann 2003). However, the classic economics of tort law arrives at different conclusions, finding that strict liability provides optimal incentives when tort law deals with uncertain risks (Shavell 1992, Ben-Shahar 1998). The article's main contribution is to deliver an economic rationale for the exemptions tort law provides for novel and uncertain risk.

The rationale rests on two assumptions. The first assumption is that experiential learning about uncertain risks at the post-market stage is essential. This sort of learning is public and creates information spillovers to other potential injurers. The assumption is contrary to the classic literature on uncertain risks (Shavell 1992, Ben-Shahar 1998) where injurers can ex-ante privately acquire perfect information about the risk. Our notion of learning has more in common with models in which the injurer's care choice in

[6] Joint with Timo Goeschl.

[7] Joint with Martin Carrier, Timo Goeschl, Johannes Lenhard, Henrike Martin, Ulrike Niemeier, Alexander Proelß, Hauke Schmidt.

period 1 determines learning about the true care-harm technology for a second period (Baumann and Friehe 2016). Here, strict liability can be inferior to a negligence rule that employs a dynamically optimal due care level, given the presence of other potential injurers. In our setting, injurers are firms that market products and learning occurs about the risk properties of a novel technology underlying those products. Since the products are based on the same novel technology, they share their risk characteristics.

For the proper understanding of the article, some information on the relationship between strict liability, negligence and decisions on bringing a product to the market is helpful. When marketing-decisions are relevant, strict liability incentivizes the right number of firms to go to the market, while negligence, or any rule providing for exemptions from liability, incentivize marketing from too many firms from a *static* perspective (Shavell 1980, Polinsky 1980). The simple reason is that exemptions lead to the firm not internalizing the full total accident costs.

The second assumption is that learning depends on the cumulative market experience with products based on the novel technology. This assumption connects learning about novel risks at the post-market stage to the existing exemptions from liability for such risks. We contend that the role of these exemptions is to offer firms that market products based on novel technologies a form of discount on the expected harm. This 'experimentation discount' induces marketing beyond the static optimum, thereby supporting experimentation with the novel technology. Experimentation produces public and socially valuable information on the risk characteristics of the technology. In this, our characterization of the learning environment differs from the existing literature on uncertain risks (Shavell 1992, Ben-Shahar 1998, Baumann and Friehe 2016) that focuses on the care level and does not consider the relationship between the marketing-decision and learning. In that bringing the novel technology to the market entails a positive information externality for other firms considering to market a product based on the novel technology, it is this exact link that enables us to give an explanation for the exemptions tort law provides for novel and uncertain risks.

The article's second contribution is of methodological nature. The article is the first to formalize the post-market learning mechanism in a tort context. Based on the two assumptions just outlined, we provide a framework in which the broader issues of regulation can be explored when learning at the post-market stage about uncertain risks is relevant. For example, the framework can be connected to the literature on private pre-market learning about uncertain risks (Shavell 1992, Ben-Shahar 1998).

We find that when liability regimes can be tailored to the specific novel technology at hand, the social optimum can be implemented. However, the exemptions provided for uncertain risks are general rules and not technology-specific. When the experimentation discount cannot be tailored to specific novel technologies, which seems to be the more realistic assumption, the choice of the discount has to trade off too little and too much

experimentation across different novel technologies. Optimal experimentation for an individual novel technology is determined by the prior probability that the technology is hazardous, the information rate from marketing products based on the novel technology and firms' static benefits from marketing.

Lastly, the article is tightly linked to the literature on tort law and innovation. We offer a new mechanism to support the longstanding claim that strict tort regimes can cause innovative and novel products to be withheld from the market (Burk and Boczar 1993, Viscusi and Moore 1993, Finkelstein 2004) and the first one to explain this claim with specific reference to state-of-the-art defenses and unknowable risks (Connolly 1965, O'Reilly 1987, Fondazione Rosselli 2004).

The second article **"Establishing Causation in Climate Litigation: Admissibility and Reliability"** (with Martin Carrier, Timo Goeschl, Johannes Lenhard, Henrike Martin, Ulrike Niemeier, Alexander Proelß, Hauke Schmidt) deals with the challenge of causation in climate litigation. A key challenge in climate litigation is to establish a causal link between a climate alteration, be it through greenhouse gas emissions or solar geoengineering, and an adverse event, usually an extreme weather event. The issue of causation is of preeminent concern in the climate litigation literature (Horton et al. 2014, Reynolds 2015, Saxler et al. 2015, Marjanac et al. 2017). When the causal link between action and effect cannot be correctly established in trials, liability regimes are not able to provide correct incentives to the relevant agents and liability loses its economic function. Establishing causation is not only a challenge in climate litigation, but in litigation concerning novel technologies and risks – by the very nature of novel risks and technologies – generally. This article thus connects article 1 and article 3, by considering an issue of general importance in the tort law of novel risks and analyzing it in the specific context of climate litigation.

Before I proceed to the paper's contributions, I provide some background on the issue at hand: Due to the inherent stochasticity of the climate system, attributing a single extreme event to human intervention into the climate system with certainty is impossible (Allen et al. 2007). The traditional legal 'but for' test of deterministic necessary causation is therefore not suitable here. However, attribution science may provide a possibility to overcome that challenge (Allen 2003, Horton et al. 2014, Marjanac et al. 2017). Attribution science is concerned with making quantitative statements about the relationship between human influence on the climate and the probability of occurrence of specific extreme events (NAS 2016). However, in order to be considered by a court, evidence based on attribution science must be legally admissible, otherwise it is excluded from the trial, denying it a role in resolving causation. The specific type of attribution science evidence we consider in this article are so-called Fraction of Attributable Risk (FAR) estimates. We focus on FAR estimates, since they are the statistical equivalent to the traditional 'but for' test (Hannart et al. 2016).

Against this background, the paper makes two contributions. The first is to demonstrate that evidence not only needs to be legally admissible, but also needs to be epistemologically reliable for resolving the challenge of causation. Otherwise the causal link between action (human climate alteration) and effect (extreme event) will typically not be correctly established in a trial. We show that rules on scientific evidence need to strike a balance between, on the one hand, the admissibility of FAR evidence and, on the other, maintaining the epistemological quality, i.e. the reliability, of FAR evidence. This means that FAR can only be an effectual tool for resolving questions of causation if and to the extent to which evidentiary standards used by courts adequately accommodate the type of scientific evidence that FAR estimates represent. In highlighting the need to strike this balance in rules of scientific evidence, the present paper contributes to an emerging literature that assesses the potential of attribution science for solving the problem of causation in climate litigation (Lusk 2017, Marjanac and Patton 2018).

The second contribution of the paper is to apply the first contribution to a specific proposal for how to accommodate scientific evidence in evidentiary standards by modifying an existing set of admissibility criteria. We use as the object of this application the *Daubert* standard. This standard offers a set of specific criteria for admissibility and applies to the United States, a jurisdiction in which climate liability suits already were launched. We find that the five *Daubert* criteria would, unmodified, exclude all FAR estimates on the basis of one criterion ('*testability*') and be inapplicable to FAR estimates on another ('*error rate*'). A simple elimination of these two criteria would ensure that FAR estimates are principally admissible, but would also allow parties to introduce unreliable FAR estimates. We argue, however, that a modified set of criteria, including criteria directly aiming at the reliability of the FAR estimates, would be capable of leading to both admissible and reliable FAR estimates.

The third article **"A Model of Solar Geoengineering Liability"** investigates the incentives that liability regimes provide in the context of solar geoengineering (SG). Although liability has been widely discussed as a potential instrument of SG governance (Horton et al. 2014, Saxler et. al 2015, Reynolds 2015), an analytical examination of SG liability is missing. In the light of the unique SG incentive structure, such an analysis fills an important research gap.

The article's first contribution is to provide an analytical, game-theoretic model of SG liability. The key feature setting SG apart from more traditional domains of liability is that it constitutes, following Weitzman's terminology (Weitzman 2015), a public good-or-bad. A public good-or-bad is 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 metrics, mean temperature and mean precipitation, suggest that moderate amounts of SG would benefit most regions of the world (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015) and that SG would only start to be detrimental to

those regions' welfare if provided beyond those moderate amounts. 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 SG 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 SG has to be compensated for. Consequently, a liability regime in the model consists of a definition of harm and a liability standard.

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 SG 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 SG is not provided beyond the victim's optimum. I call this the *marginal definition of harm*. In contrast, in traditional liability settings the two definitions coincide. I consider the two standard types of liability rules, *strict liability* and *negligence*. *Strict liability* can be defined as in standard liability setting. In contrast, *negligence* 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. 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 SG 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 identify three ways of trading off the costs and benefits in the SG setting, giving rise to the *benefit-harm negligence* standard, the *aggregate harm negligence* standard and the *individual harm negligence* standard.

The performance of a liability regime is determined by its ability to account for both the negative externality (the public bad aspect) and the positive externality (the public good aspect) arising from SG provision. I find that only the combination of the *marginal definition of harm* with the *benefit-harm negligence* standard is able to do so. This liability regime is able to implement the socially optimal amount of SG. Liability regimes arising from other combinations exhibit biases. These biases may be towards SG levels too high, too low or of ambiguous direction, depending on the liability regime.

The second contribution of the article is derived from a numerical implementation of the liability model. In estimating welfare in both the free-driver outcome (Weitzman 2015) in absence of governance and in the presence of liability regimes, the article contributes to the literature on the regional differences of SG impacts. For the implementation, I use the Residual Climate Response (RCR) model (Moreno-Cruz et al. 2012). The climate model data stems from the Geoengineering Model Intercomparison Project (Kravitz et al. 2011). I find that the extent of the free-driver problem depends on the metric chosen: For a metric of mean temperature there is only moderate SG overprovison in the free-driver outcome and SG is capable of reducing regional climate damages effectively even in the absence of liability regimes. In contrast, there is drastic overprovision for a metric

of mean precipitation in the free-driver outcome, implying a substantially lower climate welfare than in the absence of any SG. However, when governed by any liability regime, equilibrium SG is also welfare-enhancing in case of the mean precipitation metric. I furthermore find that the choice of the definition of harm is more consequential than the choice of the liability standard for the performance of a liability regime with the *marginal definition of harm* outperforming the *absolute definition of harm*.

The fourth article **"Diverging Regional Climate Preferences and the Assessment of Solar Geoengineering"** introduces the possibility of regions having diverging preferences from historic climate conditions into the assessment of regional differences in SG impacts. The article is motivated by a shortcoming of the regional SG impact literature. The literature assumes that regions have climate preferences corresponding to some historic climate, e.g. 1990 or preindustrial climate conditions. Heyen et al. (2015) criticize that assumption and show by an illustrative example that diverging climate preferences have the potential to substantially alter the assessment of regional SG impacts. The article is related to article 3 in that it helps to overcome a shortcoming in the assessment literature that affects the contribution of the numerical implementation in article 3.

The first contribution of this article is to formally extend the RCR model (Moreno-Cruz et al. 2012) by allowing regions to have temperature preferences diverging from historic climate conditions. The article delivers the theoretical insight that the impact of diverging preferences can be split into two components. The first component changes the optimal, welfare maximizing, SG level, but does not affect regional disagreement over the SG level. The second component leaves the optimal SG level unaffected, but changes regional disagreement over SG. This decomposition helps in understanding how specific diverging preferences affect globally optimal SG and the disagreement over SG by different regions. I propose three different aspects of SG performance for evaluation. The first is relative effectiveness in damage compensation. This aspect measures in percent by how much optimal SG can reduce climate damages on the regional level. The second aspect measures the maximum climate welfare in terms of the minimum damage level that SG can implement. The third one measures the gross value of SG in terms of the maximum damage reduction it can achieve.

A numerical implementation in which high-latitude regions prefer higher temperatures and low-latitude regions prefer lower temperatures yields two main results. Firstly, it shows that the performance of optimal SG relative to the absence of diverging preferences depends on the aspect of SG performance one is interested in. The presence of diverging preferences may change SG performance in either direction and the direction generally depends on which of the three aspects of SG performance is considered. Secondly, the results from the implementation suggest two welfare implications. The first is that the maximum climate welfare that SG can implement may increase with the introduction of diverging temperature preferences. However, such a positive change in

maximum climate welfare only occurs when diverging temperature preferences are small in magnitude compared to the average climate change induced warming. The second implication is that the maximum climate welfare is often higher than climate welfare in the historic climate, i.e. in the climate before CO_2 driven temperature changes set in. These implications demonstrate that diverging climate preferences do not necessarily lower SG performance, at least when optimal SG can be implemented. I argue that these implications are likely to emerge more generally in scenarios in which high-latitude regions prefer higher temperatures and low-latitude regions prefer lower temperatures.

Outlook

Since the four articles contained in this dissertation draw separate conclusions regarding their specific research questions, I do not include a separate concluding section at the end of this dissertation. In lieu I will briefly summarize the four articles' contribution in the following and provide a short outlook regarding potential future research questions emerging from this dissertation.

The first part of the dissertation examines the relationship between liability rules and novel technologies from a general perspective. Specifically, it investigates the interaction between public learning about novel risks at the post-market stage and liability rules. Its core result is that the positive information externality emerging from public post-market learning can justify exemptions from liability for harm caused by novel and uncertain risks, thereby providing an economic rationale for existing exemptions for such risks.

The second part of this dissertation has its focus on one specific novel technology, solar geoengineering (SG). The three articles in the second part either examine incentive structures relevant to SG liability or provide valuable input for understanding strategic aspects of SG. The second article provides an analysis of the challenge of causation in the specific context of climate litigation. By examining the incentive structure of evidence production, it draws conclusions about the characteristics that legal standards of evidence production need to have in order to enable attribution science to play a role in solving the problem of causation in climate litigation. The third article delivers an analytical model of SG liability and illuminates how the unique SG incentive structure translates into settings of SG liability. The fourth article contributes to understanding the structure of regional SG disagreement by extending the Residual Climate Response model (Moreno-Cruz et al. 2012) to settings in which regions have diverging climate preferences. In that better knowledge regarding the structure of regional SG disagreement helps to better understand the incentives that individual agents or regions have in strategic SG settings, it also contributes to understanding strategic aspects of SG and SG incentive structures.

This dissertation gives rise to potential future research along several dimensions. Firstly, integrating the possibility of diverging climate preferences into strategic SG settings promises further insights both into the structure of the strategic aspects in these settings, as well as into the ensuing consequences for climate welfare. At least three relevant settings come to mind in which the presence of diverging preferences could have a discernible effect. The first and simplest is the free-driver incentive structure in absence of governance. In the second type of settings, some form of governance is present, e.g. liability. A third type of setting is 'coalition games' (Ricke et al. 2013). In these settings, there is no governance, but a minimum amount of international power is required for deploying SG.

The other two dimensions revolve around issues of learning: Secondly, the central theme of the first part, learning about novel risks, is clearly of great importance in the context of SG as well. There is literature on the intertemporal aspects of SG technology development (Goeschl et al. 2013) and on the intertemporal aspects of optimal decision-making regarding research into potential SG side effects (Quaas et al. 2017) in the face of uncertain climate sensitivity and uncertain future abatement choices. This literature focuses on intergenerational aspects, thus abstracting from intragenerational heterogeneity in SG impacts and preferences. However, there may well be conflict about the development of SG stemming from intragenerational heterogeneity (Heyen 2016) and the same may apply to SG risk research. Introducing intragenerational heterogeneity into that literature could lead to far-reaching insights into the intertemporal aspects of SG incentive structures and SG governance, as well as into the strategic aspects of intertemporal decision-making on SG.

Lastly, learning from risk research is a complement to the post-market learning setting considered in the first article of this dissertation. Learning about risks from novel technologies usually takes place in part by pre-market learning from testing (Shavell 1992, Ben-Shahar 1998) and in part by post-market learning from experience. Bringing together these two stages of learning about novels risks is likely to reveal new insights. In particular, the question of the optimal point in time for switching from pre-market to post-market learning and its interaction with liability regulation, as well as with instruments of direct regulation, is an important and exciting topic for further research.

Chapter 2

Torts, Experimentation, and the Value of Information*

*Co-authored by Timo Goeschl. We want to thank Ezra Friedman, Tim Friehe, Roland Kirstein, Alon Klement, seminar participants at the University of Heidelberg, ZEW Mannheim, the Graduate Institute in Geneva, as well as participants at the EAERE 2015 in Helsinki, the AREA 2015 in San Diego, the EALE meetings in Bologna and Vienna, SELE 2015 in Chicago, the workshop on liability and innovation in Heidelberg 2016 and the ALEA 2017 in New Haven for helpful comments. We gratefully acknowledge funding by the German Research Foundation DFG under grant number GO1604-3.

2.1 Introduction

Nanotechnology, biotechnology, novel chemicals – novel technologies and their applications shape and transform life in modern society. While promising significant benefits to society, their introduction also creates novel and uncertain risks to consumers, workers, third parties or the environment. Liability regimes aim at providing adequate incentives for optimally balancing expected harm, precaution expenditures, and benefits from risky activities. In contrast to well-introduced products and technologies with abundant experiential information, society by definition still has to learn about the nature and scope of novel and uncertain risks. Beyond their usual static incentive effects, liability regimes influence whether and how such novel products and technologies are introduced, thereby entailing dynamic learning effects regarding the novel and uncertain risks in question.

The injurer's information about a risk that has materialized into harm occupies a preeminent role in tort law: Questions of foreseeability, of how to deal with unknowable risks and whether to grant state-of-the-art defenses[1] for such risks are ubiquitous. In the aftermath of the introduction of strict products liability in the Restatement (Second) of Torts in the US, various legal commentators strongly argued for state-of-the-art defenses and against liability for unknowable risks (Connolly 1965, Byrne 1973, Robb 1982, Murray 1982). Subsequently, several states enacted state-of-the-art provisions and courts rejected strict products liability for unknowable risks (Owen 2010).[2] This trend culminated in the Restatement (Third) of Torts: Products Liability, limiting injurers' liability to "foreseeable risks of harm". In Europe, both the European Products Liability Directive 85/374/EEC and the European Environmental Liability Directive 2004/35/CE provide state-of-the-art defenses. The state-of-the-art defense is "in principle available in all but a very small number of jurisdictions" in the world (Reimann 2003). In contrast, the classic analysis of liability rules in the context of uncertain risks finds that strict liability implements the social optimum (Shavell 1992). Why is it then that most of modern law shields injurers, at least to some extent, from liability for novel and uncertain risks?

In this paper we provide an economic rationale for this phenomenon. Our theory rests on two assumptions. The first assumption is that experiential learning about uncertain risks at the post-market stage is essential. This sort of learning is public and creates information spillovers to other potential injurers. The assumption is contrary to the

[1] By a state-of-the-art defense, we mean an affirmative defense which refers to the scientific knowledge at the time when an activity was carried out or a product marketed, not an industry standard for care as it is sometimes understood.

[2] A turning point was the New Jersey Supreme Court decisions taken in *Beshada v. Johns-Manville Products Corp.* (an asbestos case) and *Feldman v. Lederle Laboratories* (a drug side effects case) in 1982 and 1984, respectively. The New Jersey Supreme Court rejected the defendants motion to strike a state-of-the-art defense in the former as irrelevant in a strict liability case. Only two years later, the same court held in the latter case that there is only "a duty to warn [...] based on reasonably obtainable or available knowledge". Owen (2010) covers the fascinating story of the rise and fall of strict products liability in the US in depth.

classic literature on uncertain risks (Shavell 1992, Ben-Shahar 1998). There, injurers can ex-ante privately acquire perfect information about the risk. In such a setting, learning does not affect other potential injurers and strict liability is always first best. Our notion of learning has more in common with models in which the injurer's care choice in period 1 determines learning about the true care-harm technology for a second period (Baumann and Friehe 2016). Here, strict liability can be inferior to a negligence rule that employs a dynamically optimal due care level, given the presence of other potential injurers. In our setting, injurers are firms that market products and learning occurs of the risk properties of a novel technology.

The second assumption is that learning depends on the cumulative market experience with the novel technology. This assumption connects learning about novel risks at the post-market stage to existing exemptions from liability for such risks. We contend that the role of these exemptions is to offer firms that market products based on novel technologies a form of discount on the expected harm. This 'experimentation discount' induces marketing beyond the static optimum, thereby supporting experimentation with the novel technology. Experimentation produces public and socially valuable information on the risk characteristics of the technology. The cost of information production via experimentation is excessive marketing, implying excessive exposure of other parties to the potential risk, from a static perspective. In this, our characterization of the learning environment differs from the existing literature on uncertain risks (Shavell 1992, Ben-Shahar 1998, Baumann and Friehe 2016) that focuses on the care level and does not consider the relationship between the marketing-decision and learning.

The absence of formal models that capture the importance of actual market experience for learning about novel and uncertain risks contrasts with the legal literature that informally acknowledged this link already at the time of the Restatement (Second) of Torts: Kessler (1967) submitted that "[s]ufficient user experience is indispensable to research and making the supplier a guarantor of safety without such research may be regarded as too burdensome.", a view that was adopted and cited by the US Court of Appeals for the Second Circuit in case on side effects of the drug 'Aralen' (*Lydia Basko v. Sterling Drug, Inc., and Winthrop Laboratories, 416 F.2d 417 (2d Cir. 1969)*). A similar stance has been taken by the US Court of Appeals of California in a warning case on the same drug (*Christofferson v. Kaiser Foundation Hospitals, Civ. No. 27014, Court of Appeals of California, 1971*).

The contribution of this paper is threefold: Firstly, we deliver an economic rationale for the exemptions tort law provides for novel and uncertain risks. Secondly, by doing so, this paper is the first to formalize the post-market learning mechanism in a tort context. Thirdly, we provide a framework in which the broader issues of regulation can be explored when learning at the post-market about uncertain risks is relevant and which can be connected to the literature on private pre-market learning about uncertain risks.

The framework that this paper provides has three main building blocks, an information structure, a market structure, and a tort law structure. The information structure in our model builds on Shavell's (1992) two-period model of private learning: There is a novel technology which is hazardous with probability p and safe with probability $1-p$ (Shavell 1992). Firms can market products and applications of this novel technology. The technology's risk is generic: Either all applications are hazardous or none is. In case the novel technology is hazardous, there is a standard harm function $h(x)$ which depends on care x (which can be thought of as pre-market testing and post-market monitoring) and is the same for all applications. For reasons of tractability, we first-order approximate the relationship between cumulative market experience in period 1 and the societal value of information from learning for period 2 is chosen with a piecewise linear function: The value of information linearly increases with the cumulative market experience until a threshold is reached at which information is perfect. The slope of the value of information, i.e. the rate at which marketing the technology produces information, determines the full information threshold, which marks the passage from additional marketing being informationally productive to it being informationally unproductive.

The market structure in the model is purposefully simplified and willfully disregards strategic learning between firms (d'Aspremont and Jacquemin 1988, Kamien et al. 1992, Amir 1996): Each firm has developed exactly one application of the novel technology and faces a binary marketing-decision in each period. Different applications belong to different markets and do not interact. The private benefits from marketing are heterogeneous across firms and are net of any direct costs. A firm's marketing-decision depends on the relative size of its private benefits compared to care costs and expected liability payments.

The tort law structure is captured in a single parameter, the experimentation discount, that we formally introduce at the beginning of section 2.2. The concept follows Landes and Posner (1985): Under a strict liability regime, the experimentation discount is zero and the firm has to bear the full total accident costs. We contrast this with experimentation regimes for which the experimentation discount is positive and a firm's expected exposure to liability is therefore smaller than expected harm, leading to experimentation with the novel technology. The largest experimentation discount possible is afforded to by standard negligence. Possible experimentation discounts form a continuum, reflecting differences in existing regimes at least along two dimensions – the scope of the exemption granted and the burden of proof. An example for a high experimentation discount is product liability in the Restatement (Third) of Torts (all unforeseeable harm excluded, burden of proof lies on plaintiff). An example for a low experimentation discount is the European Products Liability Directive (state of scientific knowledge must have precluded discovering the defect, burden of proof on defendant) (Howells and Mildred 1997).[3]

[3]For example, provisions in US states range from placing the burden of proof that the harm was foreseeable, through rebuttable presumptions of nondefectiveness for products confirming to the state of the art, to affirmative state-of-the-art defenses which place the burden of proof on the defendant (Owen

Nanotechnology provides an illustration of our setting: One prominent class of nanotechnology are carbon nanotubes (CNT), other classes are nanowires and semiconductor nanocrystals. CNTs correspond to the novel technology in our model. CNTs have potential applications in areas as different as cancer diagnosis and treatment, rechargeable batteries, automotive parts, microelectronics and sporting goods (Ji et al. 2010, De Volder et al. 2013). These potential applications offer very different private benefits to firms developing and marketing them. At the same time they share common but uncertain risks, for example asbestos-like pathogenicity (Poland et al. 2008, Kostarelos 2008).

Strict liability implements the static optimum and is first-best if and only if optimal experimentation is zero. In case optimal experimentation is strictly positive, but not larger than experimentation implemented by standard negligence, there is an experimentation discount implementing optimal experimentation. When liability regimes can be tailored to the specific novel technology at hand, the social optimum can then be implemented. However, the exemptions provided for uncertain risks are general rules and not technology-specific. Courts have to adhere to those general rules, such that the first-best does not seem to be viable in the real world.

When the experimentation discount can not be tailored to specific novel technologies, the choice of the discount has to trade off too little and too much experimentation across different novel technologies. Comparatively narrow experimentation regimes, like the European Products Liability Directive, choose differently than regimes with higher experimentation discounts, like the one given by the Restatement (Third) of Torts: Products Liability. For a given technology, there is a (possibly empty) range of experimentation discounts which are superior to strict liability. In case the discount is too large, too many firms are incentivized to market and the costs of experimentation are larger than its value. In case the discount is too small, too little firms market – in the absence of marketing in the static optimum potentially even none.

Three factors determine the model: the prior probability that the technology is hazardous, the information rate from marketing applications of the novel technology and the difference between the sizes of the largest private benefits and the total accident costs, the 'initial net benefits'. Positive initial net benefits imply positive marketing in the static optimum, negative ones set the initial marginal costs of experimentation. We make three main observations regarding how these factors: Firstly, optimal experimentation is positive in case static marketing is positive, but not necessarily if static marketing is zero. The reason is that the initial marginal costs of experimentation are zero in the former case, but not in the latter. Secondly, optimal experimentation is the most valuable and the largest when the initial net benefits are zero, since the marginal

2010). Similarly, there are also increments in the potential scope of the exemptions, e.g. the European Liability Directive demands that harm arising from an activity must have been "not considered likely to cause environmental damage according to the state of scientific and technical knowledge" in order to shield an injurer from liability.

costs of experimentation are lowest and value of information maximal. In contrast, the maximum experimentation discount for an experimentation regime to be superior to strict liability is in general not largest when the initial net benefits are zero. The reason is that the initial net benefits not only determine the costs of experimentation, but also the amount of experimentation in case the experimentation discount is not a choice variable. Thirdly, a larger information rate increases the maximum experimentation discount in case static marketing is zero, but may increase or decrease the maximum experimentation discount in case static marketing is positive: The reason is that a higher information rate not only increases the marginal value of experimentation, but also increases the share of the value of information realized by a positive amount of statically optimal marketing.

Besides the literature on tort law and learning (Shavell 1992, Ben-Shahar 1998, Baumann and Friehe 2016), this paper is tightly linked to the literature on tort law and innovation. Immordino et al. (2011) provided an explanation for the claim that strict tort law is detrimental to product innovation (Burk and Boczar 1993, Viscusi and Moore 1993, Finkelstein 2004). In their model a single innovator is perfectly and privately informed about her innovation's risk characteristics after innovation, but before marketing (which takes place for every innovation). In this setting harsher regulation, be it ex-ante or ex-post, reduces innovation by a general deterrence effect, because the innovator only learns the product's risk characteristics once R&D is over. We provide an alternative mechanism. In our model the firm is itself not perfectly informed about the product's risk characteristics before marketing. The decisive step here is not the innovation, but actually bringing the product to the market. Experimentation regimes lead to experimentative marketing of products. This experimentation is informationally productive and leads to other products with the same risk characteristics being marketed later on, in case the underlying novel technology proves to be safe. Our mechanism gives an explanation for why specifically liability exemptions referring to the informational status like state-of-the-art defenses (Connolly 1965, O'Reilly 1987, Fondazione Rosselli 2004) are discussed as advancing product innovation.

The paper proceeds as follows. In the next section, we introduce the model. In section 2.3 the concepts of the value of information and the cost of experimentation are introduced. On this basis we then examine the socially optimal behavior. Section 2.4 analyzes the relative and absolute performance of strict liability and experimentation regimes. In section 2.5 we discuss extensions and the the impact of our information modeling choices on our results. Section 2.6 concludes.

2.2 The Model

A novel technology with multiple potential applications (e.g. carbon nanotubes) is at the center of our model. Different firms can independently market the novel technology's potential applications with each firm having developed exactly one application. These applications sell on different markets and are neither substitutes nor complements. The technology might be hazardous. From prior scientific laboratory research, a common and public societal belief regarding the likelihood of the technology being hazardous has emerged, but further information cannot be deduced from research. Only experience with the technology at the post-market stage can eliminate the uncertainty regarding the risk. With probability p the technology is hazardous, with probability $1-p$ it is safe. All agents share this prior.

The technology's potential hazard is generic: In case the technology is actually hazardous, marketing a specific application is associated with expected harm $h(x)$ to other parties. The harm function is the same for all applications. Expected harm can be reduced by investment into precautions, the care level x. We employ the usual assumptions $h'(x) < 0$ and $h''(x) > 0$. There is a unit continuum of firms and firms' private benefits from bringing their application to the market are heterogeneous: Firm n derives a private benefit $b(n) = 1 - n$ from marketing their application.

There are two periods. Firms decide whether to market their application in period 1 and, if so, choose the amount of care. In period 2 firms can revisit their decisions: They can e.g. market their application if they had not done so in period 1 or they can retract their application. The marketing of applications in period 1 implies experience with the technology which leads to public learning about whether the technology is hazardous. If society learns about the true state of the world, firms can base their decisions in period 2 on this knowledge. The total value from this advantage in decision-making to society is the value of perfect information (VOPI). This VOPI is an upper bound on the public value of information derived from experience with the technology at the post-market stage.

For reasons of tractability, we linearly approximate the analytical relationship between the amount of applications marketed in period 1 and the public value of information produced.[4] The parameter characterizing the relationship is the full information threshold n_{info}: If at least n_{info} firms market their application in period 1, uncertainty is resolved and society learns about the true state of the world for sure. If $n < n_{info}$ firms market their application, uncertainty is resolved with probability $\frac{n}{n_{info}}$ and with probability $1 - \frac{n}{n_{info}}$ no learning takes place. The expected value of information is therefore a piecewise linear function with positive slope up to n_{info} and a zero slope beyond n_{info}.

[4] For further discussion, we refer the reader to the penultimate section of this paper.

We characterize liability regimes by the concept of the experimentation discount. Depending on the liability regime, firms marketing an application involving uncertain risks do not have to bear all of the expected harm. The experimentation discount D summarizes, in a single parameter, the reduction in liability-related costs a non-negligent firm is afforded to under a specific regime. Given the experimentation discount, a non-negligent firm's total accident payments (care costs plus expected liability) are

$$(1 - D) \cdot (x^* + ph(x^*)).$$

The experimentation discount identifies a liability regime and lies between zero and one, where a discount of zero is equivalent to strict liability.[5] We will refer to liability regimes with a strictly positive experimentation discount as experimentation regimes. We restrict our attention to the non-trivial situation in which technologies known to be hazardous ($p = 1$) incur higher accident costs $x + h(x)$ than the benefits of any of their application $b(n)$ generate.

In our model, both the initial firm's private benefits from marketing and the number of firms (and applications) are normalized to one. We will show in section 2.5 that an extension of the model to allowing for different levels of private benefits from marketing to firms or to allowing for different numbers of firms and applications does not extend the model in substance. This implies that the findings in our model easily generalize to different levels of private benefits and different numbers of applications of firms: It will turn out that all costs and benefits can be interpreted as a fraction of the initial firm's private benefits and that the number of firms can be interpreted as a share of firms.

2.3 Social Optimum

We first examine socially optimal behavior in the second period. Afterwards, we introduce the value of information and the costs of experimentation in order to discuss socially optimal behavior in the first period.

[5] On a deeper level, two components jointly determine the experimentation discount. The first is the liability regime in force, which allows non-negligent firms to escape liability in a share d of cases. The second is the share f of expected harm in total accident costs in equilibrium $f = \frac{ph(x^*)}{x^* + ph(x^*)}$. Using d and f, we can rewrite the total accident payments as

$$x^* + (1 - d) \cdot ph(x^*) = (1 - d \cdot f) \cdot (x^* + ph(x^*)).$$

The experimentation discount D is then nothing else but the product $D = d \cdot f$. This implies that standard negligence ($d = 1$) corresponds to to an experimentation discount of f. Obviously, f is endogenously determined by expected harm $ph(x)$ and determines the maximum experimentation discount feasible for a given novel technology.

2.3.1 The Static Benchmark and Social Optimum in Period 2

We first discuss the static benchmark in which only expected payoffs for the current period are of interest and potential future gains from learning are disregarded. The results of this discussion correspond to those of the standard liability model (Shavell 1980), but show how these known results translate into our setting of a continuum of firms.

Total expected accident costs for each firm are

$$x + p \cdot h(x).$$

The statically optimal care level x_S^* is the same for all firms and fulfills

$$1 + p \cdot h'(x_S) = 0.$$

The socially optimal marketing-decision from a static point of view depends on the firms' private benefits $b(n)$ and the total expected accident costs. The private benefits curve $b(n)$ less the total accident costs $x_S^* + p \cdot h(x_S^*)$ gives the static net benefits curve

$$B(n) = b(n) - x_S^* - p \cdot h(x_S^*) = 1 - x_S^* - p \cdot h(x_S^*) - n.$$

Firms' applications should be marketed from a static perspective where the static net benefits curve is positive. The initial firm's static net benefits from marketing determine whether any marketing at all is optimal from a static point of view, and if so, how much of it. From now on we denote

$$B_{SOC} \equiv 1 - x_S^* - p \cdot h(x_S^*)$$

and refer, slightly inaccurate but succinctly, to B_{SOC} as the 'initial net benefits' from marketing. The initial net benefits are the initial firm's private benefits less the total accident costs. Since the maximum private benefits are normalized to one, the initial net benefits can be expressed in terms of the total accident costs and vice versa:

$$1 - B_{SOC} = x_S^* + ph(x_S^*).$$

In case the initial net benefits are negative, even the initial firm's private benefits are smaller than the total accident costs and no firm's application should be marketed from a static perspective. In case the initial net benefits are positive, the first B_{SOC} firms should market from a static perspective, since those firms' private benefits exceed the

total accident costs. The amount of marketing in the social optimum is therefore

$$n_S^* = \begin{cases} B_{SOC} & \text{if } B_{SOC} \geq 0 \\ 0 & \text{if } B_{SOC} < 0 \end{cases}$$

If the novel technology is known to be hazardous ($p = 1$), the initial net benefits are negative and firms should not market, since we assume $b(n) < x + h(x)$ for all firms. In contrast, if the technology is known to be safe ($p = 0$), there is no negative externality and all firms should market.

There is no learning opportunity in period 2. The static benchmark therefore fully describes socially optimal behavior in the second period. Note, that socially optimal behavior in period 2 depends on the state of information p inherited from period 1.

Lemma 1.

The social optimum in period 2 is characterized by statically optimal behavior x_S^* and n_S^*.

2.3.2 The Value and the Cost of Experimentation

The rationale for society to undertake experimentation is the value of information from learning the real state of the world. Experimentation is the difference between actual marketing and statically optimal marketing.

$$n_{exp} = n - n_S^*$$

The value of information captures the informational benefit from all marketing, including marketing which would have been undertaken for static reasons alone. The value of experimentation is restricted to the informational benefits of pure experimentation. We now introduce the value of information and then derive on this basis the value of experimentation. We then turn towards the costs of experimentation.

Since the choice of care does not influence whether information is revealed or not in our model, dynamically optimal care is always identical to statically optimal care. The social optimum in period 1 is therefore fully described by the optimal amount of experimentation, since experimentation determines the total amount of marketing.

2.3.2.1 The Value of Information

Society's value to learning p's true value, the value of perfect information (VOPI), equals the expected welfare difference between optimal behavior under resolved and unresolved

uncertainty in period 2. It is derived from better decision-making in period 2. This value is therefore created in period 1, but materializes in period 2. Since the social planner can always induce socially optimal behavior in period 2 by implementing strict liability (see section 2.4), this value is simply the expected difference in welfare between socially optimal behavior under perfect information and socially optimal behavior under uncertainty.

After information has been revealed, there are two possible states in period 2: Either, the technology is hazardous and no firm should market their application or the technology is safe and all firms should market. In the first case welfare is zero, in the second case welfare is $\frac{1}{2}$. Under uncertainty, welfare given socially optimal behavior is $\frac{1}{2}\max[0, B_{SOC}]^2$. Taking into account the respective ex-ante probabilities of the technology being hazardous or safe, the expected value of perfect information is

$$\text{VOPI} = \left[p \cdot 0 + (1-p) \cdot \frac{1}{2}\right] - \frac{1}{2}\max\left[0, B_{SOC}\right]^2 = \begin{cases} \frac{1}{2}(1-p) - \frac{1}{2}B_{SOC}^2 & \text{if } B_{SOC} \geq 0 \\ \frac{1}{2}(1-p) & \text{if } B_{SOC} < 0 \end{cases}$$

The VOPI arises from better decisions regarding whether to market applications and the choice of care levels compared to making these decisions under uncertainty. It is the expected difference between welfare after information has been revealed and welfare in the no information baseline. All else being equal, the VOPI gets smaller as welfare in the no information baseline gets larger, i.e. the more positive the initial net benefits from marketing become. Any change in the non-positive domain of the initial net benefits does not change the VOPI, since it does not change welfare in the no information baseline.

The smaller the prior probability of the technology being hazardous, the larger the prior expectation of receiving favorable information. This effect increases the VOPI. However, a change in the prior implies an indirect change in expected harm and the optimal care level. This changes the initial net benefits – the smaller p, the larger B_{SOC}. For non-negative initial net benefits, a smaller prior therefore not necessarily increases the VOPI. Whether this is the case or not depends on the exact harm function in place and is in general ambiguous. In discussing the effects of the prior, we will from now on only consider the direct effect. The direction of the indirect effect follows from the discussion of the initial net benefits. When both effects point into the same direction the net effect is evident, if they point into opposing directions, the net effect is ambiguous and depends on the harm function.

The probability of learning the true state of the world increases linearly in the amount of marketing. The marginal value of information from marketing the novel technology's applications is therefore constant and with at least n_{info} firms marketing the full value

of perfect information is appropriated. The value of information is

$$\text{VOI} = \begin{cases} \dfrac{n}{n_{info}} \cdot \text{VOPI} & \text{if } n < n_{info} \\ \text{VOPI} & \text{if } n \geq n_{info} \end{cases}$$

The marginal value of information is the product of the VOPI and the information rate, i.e. the inverse of the full information threshold:

$$r_{info} = \frac{1}{n_{info}}.$$

The higher the VOPI, the larger the expected value of what can be learned. The higher the information rate, the faster learning occurs and the faster the full VOPI is appropriated, so that further marketing then does not deliver any further information.

2.3.2.2 The Value of Experimentation

The value of experimentation is the value of the information generated by experimentation on top of the information which would have been generated in the static optimum. In case there is not any marketing in the static optimum there is no difference between the value of information and the value of experimentation. However, in case there is marketing in the static optimum, the value of information generated by the first n_S^* firms has to be subtracted from the value of information in order to obtain the value of experimentation.

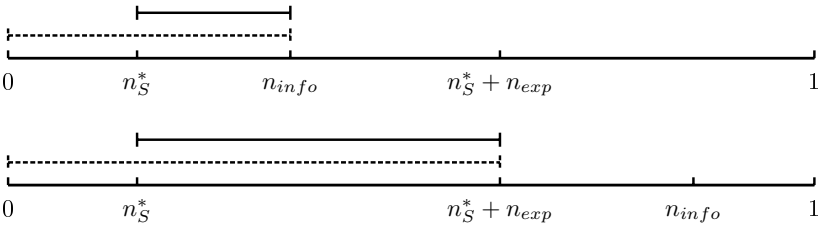

Figure 2.1: Comparison of the value of information and the value of experimentation for positive marketing in the static optimum. The dashed line represents the value of information generated by all marketing, the solid line the value of experimentation. The difference is the value of information which would have been generated in the static optimum.

The value of experimentation is linear in the amount of marketing as long as the total amount of marketing is below the full information threshold. In case the total amount of marketing is above the full information threshold, some of the experimentation is superfluous and does not yield any informational benefit. There are then three options. If even marketing in the static optimum is larger than the full information threshold, there is no value of experimentation. If marketing in the static optimum is zero the value

of experimentation is the entire VOPI. And, lastly, if static marketing is in between, the value of experimentation is the fraction of the VOPI which is realized by experimentation – as opposed to the marketing which would have taken place in the static optimum anyways.

If the total amount of marketing is above the full information threshold, all informational value which can be potentially derived from experimentation is realized. The value of experimentation then equals the maximum value of experimentation (MVOE) in this setting.

Lemma 2.

1. *The relationship between the value of experimentation and the value of information is*
$$\text{VOE} = \max\left[0, \left(1 - \frac{\max[B_{SOC}, 0]}{n_{info}}\right)\right] \cdot \text{VOI}$$

2. *i) If total marketing is smaller than the full information threshold ($n < n_{info}$), the value of experimentation is*
$$\text{VOE} = \frac{n_{exp}}{n_{info}} \cdot \text{VOPI}$$

   ii) If total marketing is larger than the full information threshold ($n \geq n_{info}$), the value of experimentation is
$$\text{VOE} = \text{VOPI} \cdot \begin{cases} 1 & \text{if } B_{SOC} < 0 \\ 1 - \frac{B_{SOC}}{n_{info}} & \text{if } 0 \leq B_{SOC} < n_{info} \\ 0 & \text{if } B_{SOC} \geq n_{info} \end{cases}$$

   This constitutes the maximum value of experimentation.

The marginal value of experimentation equals the marginal value of information: It is the information rate times the VOPI, if total marketing is below the full information threshold and zero if total marketing is above the full information threshold. The maximum value of experimentation depends on the VOPI and the fraction of marketing below the full information threshold which is not covered by marketing in the static optimum. This means that an increase in the information rate reduces the value of experimentation, if marketing in the static optimum is positive and total marketing is larger than the full information threshold. The reason is that in this case more of what can be learned is already learned under static marketing, leaving less information to be acquired by experimentation.

2.3.2.3 The Cost of Experimentation

The cost of experimentation is the welfare society has to give up in order to create additional information from experimentation. This is the expected static loss in period 1 implied by the deviation from statically optimal marketing.

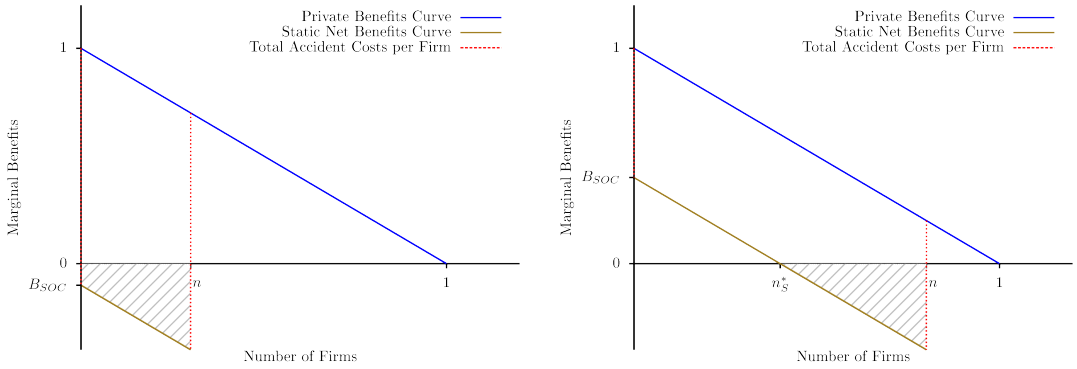

Figure 2.2: The costs of experimentation. The expected total accident costs per firm marketing is the difference between the private benefits and the static net benefits. The cost of experimentation from $n > n_S^*$ firms marketing is the static loss from doing so (the hatched area). Left: $B_{SOC} \leq 0$, therefore $n_S^* = 0$. The marginal costs of experimentation can be decomposed into the constant part $-B_{SOC}$ and a linearly increasing part. Right: $B_{SOC} > 0$, therefore $n_S^* = B_{SOC}$. The marginal costs of experimentation solely consists of the linearly increasing part.

In case the statically optimal number of firms is positive, the difference between static societal benefits and costs from the marginal firm marketing their application is zero. Therefore, the marginal loss related to the first firm experimenting is zero. The marginal loss increases linearly, since the subsequent firms' private benefit decreases linearly. If the statically optimal number of firms is zero, even the expected societal value from the first firm marketing is negative. Therefore, the first unit of experimentation involves a strictly positive static loss, which then also increases linearly.

Lemma 3. *The cost of experimentation is*

$$\text{COE} = \begin{cases} \frac{1}{2} \cdot n_{exp}^2 & \text{if } B_{SOC} \geq 0 \\ -B_{SOC} \cdot n_{exp} + \frac{1}{2} \cdot n_{exp}^2 & \text{if } B_{SOC} < 0 \end{cases}$$

The costs of experimentation to society are convex in the amount of marketing, since firms are heterogeneous in their private benefits. The initial net benefits B_{SOC} determine the marginal costs of experimentation: If they are non-negative, the initial marginal costs of experimentation are zero and the marginal costs of experimentation curves are the same for all non-negative initial net benefits. However, if they are negative, the initial marginal costs of experimentation are positive, shifting the marginal costs of experimentation curve upwards. The more negative the initial net benefits, the larger the (initial) marginal costs of experimentation.

2.3.3 Dynamic Social Optimum in Period 1

Dynamic considerations may render some experimentation optimal in period 1. Some experimentation is optimal if the initial marginal benefits of experimentation are larger than its initial marginal costs. This can be expressed in a minimum information rate for some experimentation to be optimal. However, if marketing in the static optimum is positive, there also exists a maximum information rate for experimentation to be optimal, since all information may be gathered due to marketing in the static optimum.

If the initial net benefit are non-negative, the initial marginal costs of experimentation are zero. Except for the case that marketing in the static optimum exceeds the full information threshold, the marginal benefits from the first unit of experimentation are strictly positive and some experimentation is then always optimal. If the initial net benefits are negative, the initial marginal costs of experimentation are strictly positive. Some experimentation is then optimal if and only if the initial marginal benefit of experimentation, the information rate times the VOPI, is larger than the initial marginal costs of experimentation, i.e. the initial marginal static loss $-B_{SOC}$. For negative initial net benefits, an increase in the initial net benefits therefore lowers the minimum information rate, while an increase results in a decrease in the maximum information rate if the initial marginal benefits are positive. A decrease in the prior probability p that the technology is hazardous decreases the minimum information rate for negative initial net benefits: Such a decrease increases the chances of receiving favorable information about the true state of the world, thereby increasing the VOPI.

The optimal amount of experimentation is determined by the intersection of its marginal benefits and costs. However, it is constrained from above by the full information threshold n_{info}. This may lead to corner solutions.

Proposition 1.

1. *The socially optimal care level in period 1 equals the statically optimal care level* x_S^*.
2. *If marketing in the static optimum is at least as large as the full information threshold, experimentation is never optimal. Otherwise, if the initial net benefits are non-negative, some experimentation is always optimal. If the initial net benefits are negative, some experimentation is optimal if and only if the information rate times the VOPI is larger than its initial marginal costs. This can be expressed in terms of minimum and maximum information rates for some experimentation to be optimal:*

$$r_{info}^{min,opt} = \begin{cases} 0 & \text{if } B_{SOC} \geq 0 \\ -\frac{B_{SOC}}{\text{VOPI}} & \text{if } B_{SOC} < 0 \end{cases}$$

$$r_{info}^{max,opt} = \begin{cases} \frac{1}{B_{SOC}} & \text{if } B_{SOC} \geq 0 \\ \infty & \text{if } B_{SOC} < 0 \end{cases}$$

3. The optimal amount of experimentation is

$$n_{exp}^{*} = \begin{cases} \min\left[r_{info} \cdot \text{VOPI}, \max\left[0, n_{info} - B_{SOC}\right]\right] & \text{if } B_{SOC} \geq 0 \\ \min\left[\max\left[0, B_{SOC} + r_{info} \cdot \text{VOPI}\right], n_{info}\right] & \text{if } B_{SOC} < 0 \end{cases}$$

The marginal value of experimentation is the largest when the initial net benefits are non-positive, since the VOPI is then the largest. The marginal costs of experimentation are the smallest when the initial net benefits are non-negative. Therefore, the optimal amount of experimentation and the net value of experimentation to society are largest if the initial net benefits are zero. Higher information rates increase the initial marginal value of experimentation, but at the same time tighten the constraint on the maximum productive amount of experimentation. Higher information rates therefore initially increase the optimal amount of experimentation, but only until the constraint becomes binding and the optimal amount of experimentation then decreases with further increases in the information rate: The faster learning initially occurs, the higher the value of marketing, but the faster everything of value is learned. A lower prior probability that the technology is hazardous increases the VOPI. It therefore increases the marginal value of experimentation, hence the optimal amount of experimentation.

2.4 Liability Regulation

We will now examine the incentives liability regimes provide for different firms regarding their marketing decisions and the ensuing welfare consequences. If the risk is uncertain at the time of the marketing decision, the injurer receives an experimentation discount on the expected harm if she exerts due care x_S^*. An experimentation discount of zero corresponds to strict liability. We refer to all liability regimes with strictly positive experimentation discount as experimentation regimes. The experimentation discount determines how much less a firm's total accident payments are compared to the total accident costs. Given the experimentation discount D, an injurer's expected total accident payments are

$$(1 - D) \cdot (x_S^* + ph(x_S^*)).$$

The experimentation discount determines, in relative terms, how much of the negative externality is not internalized in equilibrium by firms marketing their applications.

In general, the maximum experimentation discount feasible is one. However, for a given novel technology the maximum experimentation discount is

$$f = \frac{ph(x^*)}{x^* + ph(x^*)},$$

corresponding to standard negligence. In our analysis, we will allow for all experimentation discounts between zero and one. However, for a specific technology, only experimentation discounts up to f are feasible. Where relevant, we will explicitly discuss the implications.

2.4.1 Behavior under Liability Regimes

Since every single firm's marketing-decision is non-pivotal for information acquisition, firms do not have any strategic considerations regarding the learning effects of their marketing-decisions. Firms' incentives to market their application are therefore determined by their expected static pay-off. Given the same state of information, a given liability regime leads to the same behavior in both periods. If uncertainty about the risk is resolved, experimentation regimes do not offer any experimentation discount. In that case, experimentation regimes are not different from strict liability.

Since any experimentation discount is conditional on the firm exerting due care, firms will choose the socially optimal care level in equilibrium. Strict liability ($D = 0$) makes the injurer internalize the total accident costs and therefore implements the static optimum. Experimentation regimes ($D > 0$) make it worthwhile for a firm to market if its net benefits from marketing plus the absolute experimentation discount the firm receives (the experimentation discount times the total accident costs) are larger than zero:

$$[b(n) - (x_S^* + ph(x_S^*))] + D \cdot (x_S^* + ph(x_S^*)) \geq 0 \ .$$

Since the total accident costs are determined by the initial net benefits,

$$1 - B_{SOC} = x_S^* + ph(x_S^*),$$

experimentation under such a regime can be expressed in terms of the initial net benefits and the experimentation discount and firm n will market if

$$n \leq B_{SOC} + D \cdot (x_S^* + ph(x_S^*)) = B_{SOC} + D \cdot (1 - B_{SOC}) \ .$$

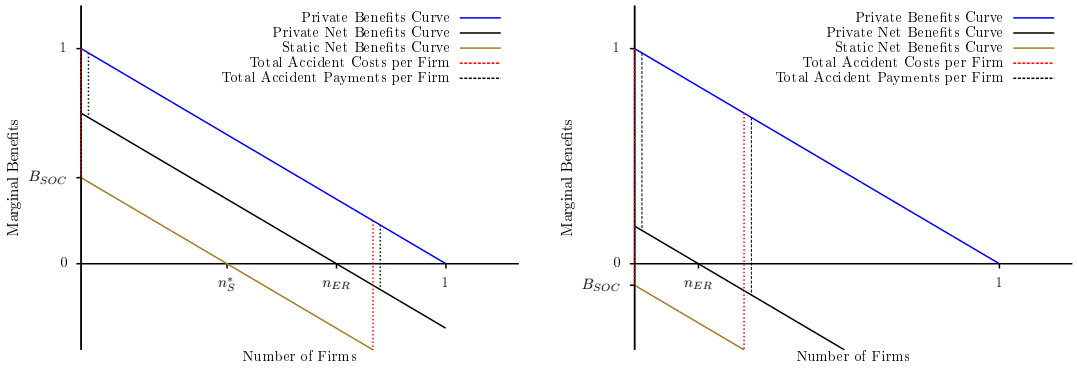

Figure 2.3: Marketing under experimentation regimes. Due to the experimentation discount the total accident payments are smaller than the total accident costs and the private net benefits curve is located above the static net benefit curve. The number of firms marketing under a experimentation regime is therefore (weakly) larger compared to the social optimum: Experimentation is always positive if statically optimal marketing is positive. In case statically optimal marketing is zero, experimentation is only positive if the absolute experimentation discount is larger than the initial net benefits.

Lemma 4. *All liability regimes implement the optimal care level* x_S^*.

1. *Strict liability implements the statically optimal number of firms:*

$$n_{SL} = \begin{cases} 0 & \text{if } B_{SOC} < 0 \\ B_{SOC} & \text{if } B_{SOC} \geq 0 \end{cases}$$

2. *Experimentation regimes implement the statically optimal number of firms if uncertainty about the risk is resolved. If uncertainty about the risk is unresolved, experimentation regimes implement a weakly larger number of firms than the static optimum:*

$$n_{ER} = \begin{cases} 0 & \text{if } B_{SOC} + D \cdot (1 - B_{SOC}) \leq 0 \\ B_{SOC} + D \cdot (1 - B_{SOC}) & \text{if } B_{SOC} + D \cdot (1 - B_{SOC}) > 0 \end{cases}$$

The substance of lemma 4 is well-known (Shavell 1980, Polinsky 1980). The only cases in which experimentation regimes do not implement a higher number of firms than the static optimum is if either uncertainty about the risk is resolved, or if the experimentation discount is too small to induce any firm to market its application. The latter is the case if

$$D \leq \frac{-B_{SOC}}{1 - B_{SOC}}$$

and can only occur if there is zero marketing in the static optimum.

Since the static optimum is always optimal in period 2, strict liability is first-best in period 2 and experimentation regimes in general implement too many firms in period 2 (with the exception of the two cases outlined above).

2.4.2 Strict Liability in Period 1

In period 1 uncertainty about the risk is always unresolved. Strict liability implements the static optimum and it is therefore first-best whenever optimal experimentation is zero. There are two such cases. In the first one the static net benefits from marketing are high enough for uncertainty to be resolved in the static optimum. In the second case the initial marginal costs of experimentation are larger than the marginal value of experimentation. The second scenario can only occur if statically optimal marketing is zero. In all other cases some experimentation is optimal and strict liability therefore not first-best.

Proposition 2. *Strict liability always implements zero experimentation. In period 1 it is first-best if either*

$$B_{SOC} \geq n_{info} \quad \text{or} \quad B_{SOC} + \frac{1}{2}\frac{1}{n_{info}}(1-p) \leq 0.$$

In all other cases optimal experimentation is positive and too few firms market under strict liability.

2.4.3 Experimentation Regimes in Period 1

Experimentation regimes are in general inferior from a static perspective, but may have a dynamic edge. The amount of experimentation implemented by a regime with experimentation discount D can be deduced from lemma 4:

Lemma 5.

1. *Experimentation under a regime with experimentation discount D is*

$$n_{exp}^{ED} = \begin{cases} D(1-B_{SOC}) & \text{if } B_{SOC} \geq 0 \\ \max\left[0, B_{SOC} + D(1-B_{SOC})\right] & \text{if } B_{SOC} < 0 \end{cases}$$

2. *In case the initial net benefits are negative, experimentation is zero for some experimentation discounts. Experimentation is then only positive if*

$$D > \frac{-B_{SOC}}{1-B_{SOC}}.$$

Since liability regimes implement the socially optimal care level, they are first-best if they also implement the socially optimal amount of marketing. Given optimal experimentation is zero, all regimes implementing zero experimentation are first-best. Given optimal experimentation for a specific novel technology is positive, there is an experimentation discount which implements just the right amount of experimentation for that specific

novel technology. Whether a given experimentation discount can be implemented by a liability regimes for a given novel technology, depends on whether D is larger than the experimentation discount f implemented by standard negligence for the given technology. In case D is larger than f, the experimentation discount is not feasible, otherwise it is.

Proposition 3. *An experimentation regime is first-best if and only if*

$$n_{exp}^{ED} = n_{exp}^{*}.$$

If marketing in the static optimum is smaller than the full information threshold, there is always an experimentation discount such that an associated regime is first-best:

i) If optimal experimentation n_{exp}^{*} for a novel technology is larger than zero this experimentation discount is

$$D = \begin{cases} \frac{n_{exp}^{*} - B_{SOC}}{1 - B_{SOC}} & \text{if } B_{SOC} < 0 \\ \frac{n_{exp}^{*}}{1 - B_{SOC}} & \text{if } B_{SOC} \geq 0 \end{cases}.$$

ii) If optimal experimentation n_{exp}^{*} for a novel technology is zero, these experimentation discounts are

$$D \leq \frac{-B_{SOC}}{1 - B_{SOC}}.$$

The results for strict liability and experimentation regimes imply that, in case liability regimes can be tailored to a specific novel technology at hand, the first-best outcome can be implemented if optimal experimentation is not too large, i.e. unless $D > f$ and optimal experimentation therefore exceeds experimentation implemented by standard negligence. However, whether liability regimes can actually be fine-tuned to specific novel technologies is questionable. Firstly, it is far from clear whether courts possess adequate knowledge about the public information gains from bringing the technology to the market. Secondly, the legal framework giving rise to the experimentation discount – mainly the scope of the exemptions and the burden of proof – in real-world liability regimes is not technology-specific, but applies to all novel technologies in the same jurisdiction in the same way.

2.4.4 Non-Technology-Specific Liability Regimes

We will now assume that liability regimes are fixed across novel technologies and investigate the trade-offs of choosing the liability regime, as well as the relative ranking of strict liability and experimentation regimes. For a given novel technology, a liability regime will almost always implement too much or too little experimentation: The amount of experimentation implemented by a given liability regime is determined by the

initial net benefits from marketing and the absolute experimentation discount. It does not depend on the information rate and the prior probability of the technology being hazardous. In case no experimentation is optimal, strict liability clearly is superior. In case some experimentation is optimal and experimentation implemented by an experimentation regime is too small compared to the social optimum (but still positive), the experimentation regime is still superior to strict liability, since costs of experimentation are convex and the value of information is concave. If experimentation is too large, this is not necessarily the case.

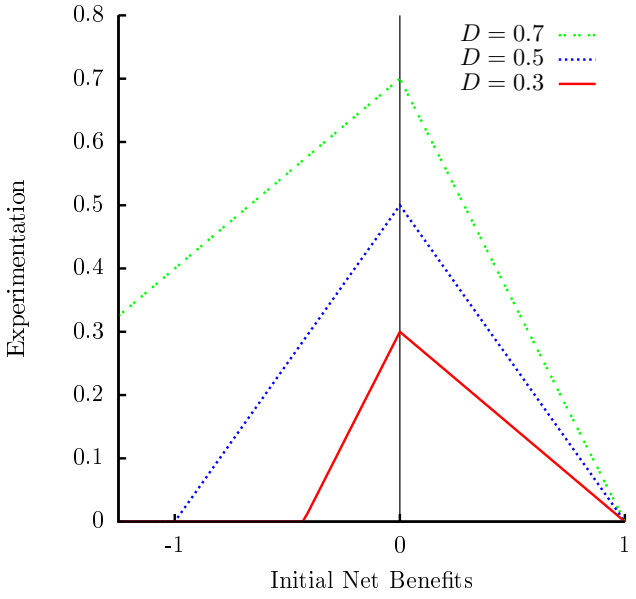

Figure 2.4: Experimentation under regimes with different experimentation discounts. Experimentation is fully determined by B_{SOC} and D, but it does not depend on the information rate and the the prior probability of the technology being hazardous. For any experimentation discount experimentation is largest for $B_{SOC} = 0$ and is linearly decreasing in B_{SOC} in either direction.

We will identify liability regimes with their associated experimentation discounts, i.e. strict liability is simply a liability regime with an experimentation discount of zero. Experimentation is the larger the larger the experimentation discount is. The initial net benefits from marketing influence the amount of experimentation as well: Experimentation is determined by the absolute experimentation discount firms receive minus, in case zero marketing is statically optimal, the static loss in welfare from the first firm marketing. Higher initial net benefits are equivalent to lower total accident costs. Larger initial net benefits of one unit are therefore associated with a smaller absolute experimentation discount of D units. If the initial net benefits are non-negative, this is the only effect present and experimentation is decreasing in B_{SOC} at the rate D. However, if the initial net benefits are negative, the initial static loss from marketing decreases by one unit and experimentation is therefore increasing by $1 - D$ if the initial net benefits increase by one unit.

Lemma 6. *Experimentation is increasing in the experimentation discount. It is increasing in the initial social benefits if $B_{SOC} < 0$ and decreasing if $B_{SOC} \geq 0$.*

$$\frac{d}{dB_{SOC}} n_{exp}^{ED} = \begin{cases} -D & \text{if } B_{SOC} \geq 0 \\ 1 - D & \text{if } B_{SOC} < 0 \end{cases}$$

The amount of experimentation cannot be adjusted to a novel technology under a non-technology-specific liability regime. The absolute cost and value of the experimentation implemented are therefore of interest for comparing strict liability to an experimentation regime, rather than the marginal quantities. Equivalently, one can compare the average cost and value of experimentation: In case total marketing is smaller than the full information threshold, both costs and value of experimentation can be easily decomposed into average costs and average value times experimentation. It then suffices to compare these average quantities. However, in case total marketing is larger than the full information threshold, the absolute value of experimentation cannot be decomposed in the same way and it is easier to compare the absolute cost and value of experimentation.

We measure the welfare effect of experimentation regimes relative to strict liability: The costs and value of experimentation are measured relative to the strict liability outcome to which we assign a welfare level of zero. The value and costs of experimentation always refer to a liability regime implementing the respective amount of experimentation, i.e. to a specific experimentation discount which may be positive or zero.

The Absolute and Average Value of Experimentation

Calculating the absolute and average value of experimentation in the relevant cases is straightforward: If total marketing is smaller than the full information threshold, the value of experimentation equals the marginal value of experimentation. Otherwise the absolute value of experimentation is of interest, which then equals the maximum value of experimentation.

Lemma 7.

1. *If total marketing is smaller than the full information threshold ($n_S^* + n_{exp}^{ED} < n_{info}$), the average value of experimentation is*

$$\text{VOE}_{ED} = \frac{\text{VOPI}}{n_{info}} \cdot \begin{cases} (B_{SOC} + D(1 - B_{SOC})) & \text{if } B_{SOC} < 0 \\ D(1 - B_{SOC}) & \text{if } B_{SOC} \geq 0 \end{cases}$$

2. *If total marketing is larger than the full information threshold ($n_S^* + n_{exp}^{ED} \geq n_{info}$), the absolute value of experimentation is*

$$\text{VOE}_{ED} = \text{VOPI} \cdot \begin{cases} 1 & \text{if } B_{SOC} < 0 \\ \frac{n_{info} - B_{SOC}}{n_{info}} & \text{if } 0 \leq B_{SOC} < n_{info} \\ 0 & \text{if } B_{SOC} \geq n_{info} \end{cases}$$

The Absolute and Average Costs of Experimentation

The costs of experimentation under a given liability regime can always be decomposed into the average costs of experimentation times the amount of experimentation implemented:

$$\text{COE}_{ED} = \bar{c}_{ED} \cdot n_{exp}^{ED}.$$

In case the initial net benefits are non-negative, the average costs of experimentation $\bar{c}_{ED}$ are simply half the amount of experimentation

$$\frac{1}{2} D(1 - B_{SOC}),$$

since the marginal costs for the unit of experimentation are zero and the marginal costs of experimentation increase linearly. Larger experimentation discounts increase experimentation and therefore also the average costs of experimentation. Similarly, higher initial net benefits correspond to smaller total accident costs, therefore to a smaller absolute experimentation discount, less experimentation and smaller average costs of experimentation.

In case the initial net benefits are negative, the marginal costs of experimentation are positive even for the first firm. Marginal costs of experimentation still increase linearly, resulting in average costs consisting of two parts. The first is again half the amount of experimentation, the second one the initial marginal costs $-B_{SOC}$:

$$\frac{1}{2}(B_{SOC} + D(1 - B_{SOC})) - B_{SOC}.$$

Larger experimentation discounts again imply larger average costs of experimentation. Here, higher initial net benefits increase experimentation, but at the same time decrease the initial marginal costs of experimentation. The latter effect outweighs the former one and the average costs of experimentation decrease with higher initial net benefits. A larger experimentation discount implies more experimentation, which in turn implies larger average costs of experimentation. Therefore, the absolute costs of experimentation are increasing in the experimentation discount.

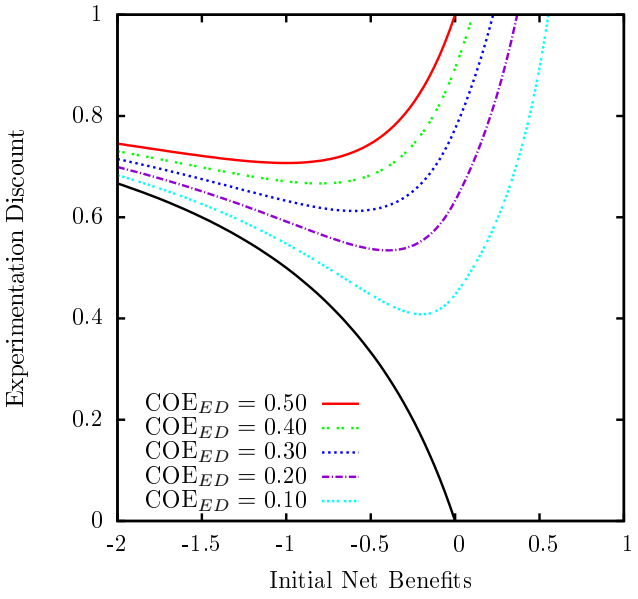

Figure 2.5: Experimentation isocost curves. In the area below the solid black line experimentation is zero. Increasing D shifts to a higher isocost line. Increasing B_{SOC} shifts to a higher isocost line for $B_{SOC} < B^*_{SOC}$ and shifts to a lower isocost line for $B_{SOC} > B^*_{SOC}$ (compare lemma 8). Since the VOPI is bounded from above by 0.5, experimentation regimes for parameter combinations above the 0.5 isocost line are always inferior to strict liability.

In case the initial net benefits are non-negative, the absolute costs of experimentation are decreasing in the initial net benefits, because the amount of experimentation is then solely determined by the absolute experimentation discount. Larger initial net benefits (or equivalently smaller total accident costs) imply a smaller amount of experimentation and therefore also smaller average costs of experimentation. In case the initial net benefits are negative, the absolute costs of experimentation are increasing in the initial net benefits up to B^*_{SOC} (compare lemma 8), which depends on the experimentation discount in place. The absolute costs are decreasing beyond this point. The reason is that the absolute costs of experimentation are average costs times experimentation: While the average costs of experimentation are decreasing in the initial net benefits, experimentation is increasing in the initial net benefits. For details we refer the reader to the proof of lemma 8.3. These characteristics of the absolute costs of experimentation are reflected in Figure 2.5, which shows several experimentation isocost curves.

Lemma 8.

1. *The costs of experimentation for a given experimentation discount can be decomposed into average costs of experimentation times the amount of experimentation:*

$$\text{COE}_{ED} = \bar{c}_{ED} \cdot n^{ED}_{exp}$$

The average costs of experimentation are

$$\bar{c}_{ED} = \begin{cases} \frac{1}{2}D(1 - B_{SOC}) & \text{if } B_{SOC} \geq 0 \\ \frac{1}{2}D(1 - B_{SOC}) - \frac{1}{2}B_{SOC} & \text{if } B_{SOC} < 0 \end{cases}$$

The average costs of experimentation are decreasing in the initial net benefits B_{SOC} and in the experimentation discount D.

2. *The absolute costs of experimentation are increasing in the experimentation discount D.*

3. *For a given experimentation discount D, there exists*

$$B^*_{SOC} = -\frac{D^2}{(1 - D^2)} < 0,$$

such that the absolute costs of experimentation are increasing in the initial net benefits for $B_{SOC} < B^*_{SOC}$ and decreasing in the initial net benefits for $B_{SOC} > B^*_{SOC}$.

Relative Ranking of Strict Liability and Experimentation Regimes

An experimentation regime is superior to strict liability if the absolute value of experimentation outweighs the absolute costs of experimentation:

$$\text{VOE}_{ED} > \text{COE}_{ED}.$$

In case total marketing is smaller than the full information threshold, this is equivalent to the marginal value of experimentation being larger than the average costs of experimentation:

$$r_{info} \cdot \text{VOPI} > \bar{c}_{ED}. \tag{1}$$

In case total marketing exceeds the full information threshold, the full value of experimentation is realized, which is capped at the maximum value of experimentation. Experimentation regimes are then superior if the maximum value of experimentation is larger than the absolute costs of experimentation:

$$\text{MVOE} - \text{COE}_{ED} > 0.$$

This means that the fraction of the VOPI which is not realized under marketing in the static optimum, i.e. under strict liability, must be larger than the costs of the amount of experimentation implemented:

$$\left(1 - \frac{\max[0, B_{SOC}]}{n_{info}}\right) \cdot \text{VOPI} > \text{COE}_{ED}. \tag{2}$$

In fact, irrespective of whether total marketing exceeds the full information threshold, an experimentation regime is superior to strict liability if and only if both conditions hold: On the one hand it is necessary that the maximum value of experimentation must be large enough to exceed the actual absolute costs of experimentation. On the other hand it is necessary that the marginal value of experimentation is larger than the actual average costs of experimentation. Since each of the conditions is sufficient for either of the cases of total marketing exceeding or not exceeding the full information threshold, the conditions are jointly necessary and sufficient.

The experimentation discount only enters through the cost side in both conditions and a larger experimentation discount implies in both cases higher (average and absolute) costs. Both conditions therefore imply a maximum experimentation discount. For an experimentation regime to be superior to strict liability, the actual experimentation discount has to be below both of the maximum experimentation discounts implied by either condition, since both conditions have to hold. At the same time experimentation has actually to be positive. This means that the experimentation discount also has to be larger than some minimum experimentation discount in case the initial net benefits are negative.

Proposition 4.

1. *An experimentation regime is superior to strict liability if and only if the following conditions are fulfilled:*

   i)
   $$r_{info} \cdot \text{VOPI} > \bar{c}_{ED}$$

   ii)
   $$(1 - \max[0, B_{SOC}] \cdot r_{info}) \cdot \text{VOPI} > \text{COE}_{ED}$$

   iii)
   $$\frac{-B_{SOC}}{1 - B_{SOC}} < D$$

2. *Let the initial net benefits be negative. An experimentation regime is superior to strict liability if it holds for the experimentation discount D that*
   $$\frac{-B_{SOC}}{1 - B_{SOC}} < D < \min\left[\frac{\sqrt{2 \cdot \text{VOPI} + (B_{SOC})^2}}{1 - B_{SOC}}, \frac{2 \cdot \text{VOPI} \cdot r_{info} + B_{SOC}}{1 - B_{SOC}}\right]$$

3. *Let the initial net benefits be non-negative. An experimentation regime is superior to strict liability if it holds for the experimentation discount D that*
   $$D < \min\left[\frac{\sqrt{2 \cdot \text{VOPI} \cdot (1 - r_{info} \cdot B_{SOC})}}{1 - B_{SOC}}, \frac{2 \cdot \text{VOPI} \cdot r_{info}}{1 - B_{SOC}}\right]$$

The first condition states that the marginal value of experimentation has to be larger than the average costs of experimentation for an experimentation regime to outperform strict liability. The condition is fulfilled for more experimentation discounts, the higher the information rate is, since the information rate increases the marginal value of experimentation. Increasing the initial net benefits has two effects: It decreases the average costs of experimentation and decreases the value of perfect information by increasing welfare in the no-information baseline. The latter effect only exists if the initial net benefits are larger than zero, but reduces the value of perfect information eventually to zero. The maximum information discount for which the first condition is fulfilled is therefore inversely U-shaped. The maximum is attained for initial net benefits strictly larger than zero, since the latter effect is quadratic and hence initially zero.

The second condition states that the maximum value of experimentation has to be larger than the absolute costs of experimentation. A higher information rate decreases the maximum experimentation discount fulfilling this condition, since it decreases the maximum value of experimentation by increasing the fraction of the value of perfect information realized under strict liability. This effect is only present if the initial net benefits are larger than zero. Increasing the initial net benefits increases the absolute costs of experimentation below B^*_{SOC} and decreases the absolute costs of experimentation beyond B^*_{SOC} (compare lemma 8). It has no effect on the maximum value of experimentation for non-negative initial net benefits and decreases the maximum value of experimentation for positive initial net benefits, since it decreases the value of perfect information and increases the fraction of it which is realized under strict liability. An increase in the initial net benefits therefore lowers the maximum experimentation discount such that the second condition is fulfilled for initial net benefits below B^*_{SOC} and increases it between B^*_{SOC} and zero. Above zero the net effect depends on the the relative size of the opposing effects.[6]

The third condition only depends on the initial net benefits and states that experimentation has actually to be positive: It demands that the absolute experimentation discount is larger than the social costs of the first unit of marketing. For non-negative initial net benefits this condition is therefore fulfilled for all positive experimentation discounts. For negative initial net benefits increasing the initial net benefits leads to more experimentation discounts implementing at least some experimentation.

Whether condition (1) or condition (2) is the binding one for the maximum experimentation discount crucially depends on the information rate. Condition (1) is fulfilled for large information rates and not for small ones. For positive initial net benefits, condition (2) is not fulfilled for large information rates and is fulfilled for small information rates if the VOPI is larger than the absolute costs of experimentation. For non-positive initial

[6] It can be easily shown that for information rates below two, the maximum experimentation discount increases at zero before eventually dropping to zero as the maximum value of experimentation goes to zero. For information rates above two, the maximum experimentation discount directly decreases in the initial net benefits from zero on.

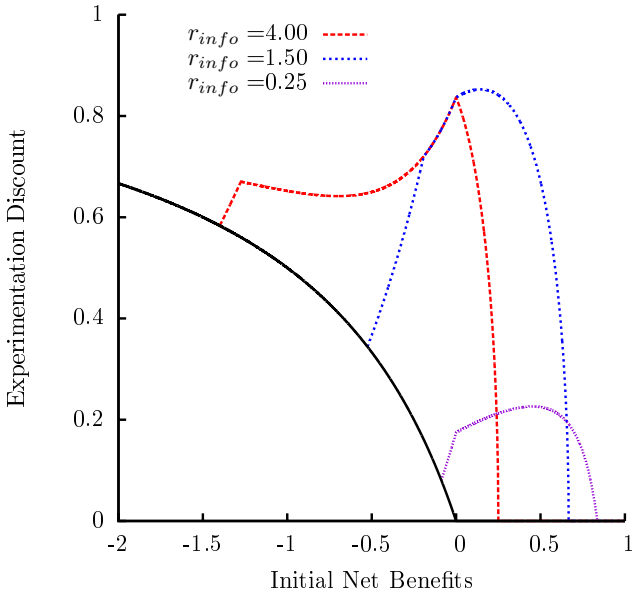

Figure 2.6: Maximum and minimum experimentation discount for an experimentation regime to be superior to strict liability. The minimum experimentation discount is shown as the black solid line. The other curves represent the maximum experimentation discount for varying information rates. In case the initial net benefits are negative, an increase in the initial net benefits increases the maximum experimentation discount if the information rate is low enough such that condition (1) is binding or the initial net benefits are larger than B^*_{SOC}. In case the initial net benefits are non-negative, the maximum experimentation discount directly begins to fall for the highest information rate, whereas it initially rises and later falls for lower information rates. Higher information rates imply a higher experimentation discount for non-positive initial net benefits, but do not necessarily so for positive initial net benefits.

net benefits condition (2) is independent from the information rate and fulfilled if the VOPI is larger than the absolute costs of experimentation. Therefore, for large enough information rates, condition (2) always becomes the binding one. Condition (1) becomes the binding one if the VOPI is larger than the absolute costs of experimentation and, in case the initial net benefits are positive, the information rate is small.

The preceding discussion enables us to deduce the relationships between the determinants of our model and the maximum experimentation discount. An increase in the information rate increases the maximum experimentation discount as long as condition (1) is the binding condition. If condition (2) is binding, it decreases the maximum experimentation discount in case the initial net benefits are positive and does not change the maximum experimentation discount in case the initial net benefits are non-positive.

In case the initial net benefits are negative, an increase in the initial net benefits increases the maximum experimentation discount, except for the case that condition (2) is the binding one and that the initial net benefits are smaller than B^*_{SOC}. In case the initial net benefits are non-negative, the effect of an increase in the initial net benefits depends both on which of the conditions is binding and on the exact values of the initial net benefits and the information rate. In contrast, in the social optimum, both the amount

of experimentation and the net value of experimentation are largest when the initial net benefits are zero. The discrepancy is due to the fact that the initial net benefits not only determine the initial marginal costs of experimentation and co-determine the VOPI, but also determine the amount of experimentation given a fixed experimentation discount. At larger initial net benefits than zero less experimentation is implemented compared to initial net benefits of zero, since total accident costs are smaller, implying smaller average and absolute costs of experimentation. Since a maximum experimentation discount by definition implements too much experimentation compared to the social optimum, decreasing experimentation by higher initial net benefits increases the maximum experimentation discount, if the decrease in costs outweighs the accompanying decrease in the value of experimentation. In case total marketing exceeds the full information threshold, the maximum experimentation discount may also be larger for initial net benefits smaller than B^*_{SOC} compared to initial net benefits of zero, since experimentation is increasing in the initial net benefits.

In every case the only direct effect of a decrease in the prior is an increase in the value of experimentation, thereby increasing the experimentation discounts fulfilling conditions (1) and (2) and increasing the maximum experimentation discount.

Trade-offs in Choosing the Experimentation Discount

The preceding result stated which experimentation regimes are superior to strict liability in a given setting. For clarifying the trade-offs in choosing an experimentation discount it is useful to look at a fixed experimentation discount's impact on different settings and compare the outcomes with the social optimum for each setting. We will do so by deriving minimum and maximum information rates for experimentation regimes to be superior to strict liability.

Condition (1) corresponds to a minimum information rate: The larger the information rate, the larger the marginal value of information. Condition (2) corresponds to a maximum information rate for positive initial net benefits, since a larger information rate decreases the maximum value of experimentation in that case. In case of non-positive initial net benefits, condition (2) is independent from the information rate and, together with the requirement of positive experimentation, constitutes a precondition for an experimentation regime to be superior unrelated to the information rate.

Proposition 5.

1. *Let the initial net benefits be non-negative. An experimentation regime with experimentation discount D is superior to strict liability if and only if*

   i)

$$\frac{-B_{SOC}}{1-B_{SOC}} < D < \frac{\sqrt{2 \cdot \mathrm{VOPI} + (B_{SOC})^2}}{1-B_{SOC}}.$$

ii)

$$r_{info} > r_{info}^{min,ED} = \frac{\bar{c}_{ED}}{\text{VOPI}} > r_{info}^{min,opt}$$

2. *Let the initial net benefits be negative. An experimentation regime with experimentation discount D is superior to strict liability if and only if*

i)

$$r_{info} > r_{info}^{min,ED} = \frac{\bar{c}_{ED}}{\text{VOPI}} > r_{info}^{min,opt}$$

ii)

$$r_{info} < r_{info}^{max,ED} = \frac{\text{VOPI} - \text{COE}_{ED}}{\text{VOPI} \cdot B_{SOC}} < r_{info}^{max,opt}$$

For a positive experimentation discount, the minimum information rate for the experimentation regime to be superior to strict liability is larger than the minimum information rate in for at least some experimentation to be socially desirable. The reason is that, given a non-technology-specific experimentation regime, the marginal value of experimentation has to exceed the average costs of experimentation for the experimentation implemented to be welfare-enhancing. In contrast, in the social optimum the marginal value only has to exceed the initial marginal costs of experimentation for some experimentation to be optimal. The former is larger than the latter, since the costs of experimentation are convex.

The maximum information rate, in case the initial net benefits are positive, for an experimentation regime to be superior to strict liability is smaller than the corresponding maximum information rate for some experimentation to be socially desirable. In the social optimum, the maximum information rate is simply determined by the full information threshold. For a non-technology-specific experimentation regime, the maximum value of experimentation must exceed the costs of the experimentation actually implemented. For negative initial net benefits there is no maximum information rate in either the social optimum or for a non-technology-specific experimentation regime. However, for the experimentation regime there is a condition ensuring that the amount of experimentation is actually positive and that the VOPI exceeds the actual costs of experimentation. Again, these requirements arise because the experimentation discount is fixed and experimentation is no choice variable.

These differences in the conditions under which some experimentation is socially optimal and the condition under which non-technology-specific experimentation regimes implement experimentation in a welfare-enhancing way, highlight the inevitable shortcomings of a non-technology-specific experimentation regime for individual novel technologies. For some novel technologies experimentation is too large. This may happen if experimentation is not optimal in the first place or if actual experimentation is larger than the socially optimal one. In the latter case, the experimentation regime may even be inferior to strict liability although some experimentation is socially desirable. The larger the experimentation discount, the more often this is the case. For other novel technologies

experimentation is too low. As long as experimentation is still positive, the experimentation regime outperforms strict liability. However, in case the initial net benefits are negative, experimentation becomes zero for too small experimentation discounts, rendering the the experimentation regime equivalent to strict liability. The smaller the experimentation discount, the more often this is the case.

This implies the presence of two related trade-offs. Firstly, the experimentation discount trades off too much and too little experimentation for different novel technologies. Secondly, the experimentation discount trades off how often the experimentation regime is inferior to strict liability despite some experimentation being optimal and how often the experimentation regime is equivalent to strict liability despite some experimentation being optimal, i.e. how often it implements experimentation in a welfare-reducing way and how often it fails to implement socially desirable experimentation. Comparatively narrow experimentation regimes, like the European Products Liability Directive, favor not implementing any experimentation more often over implementing experimentation in a welfere-reducing way more often compared to regimes with higher experimentation discounts, like the one given by the Restatement (Third) of Torts: Products Liability. In order to know the globally optimal experimentation discount, one would need to know the probability density $f(r_{info}, B_{SOC}, p)$ across novel technologies.

Impact on Expected Product Innovation

In our setting product innovation happens when an existing innovation, i.e. an application of the novel technology, is actually marketed. Experimentation increases product innovation by definition and unambiguously in period 1. Given there is some experimentation in period 1, the probability of learning whether the novel technology is hazardous or safe between periods is increased. The increased probability can lead to society learning that the novel technology is safe, that the novel technology is hazardous or to not learning the true state of the world. In the first case, product innovation is increased compared to the no-experimentation baseline and it is not affected in the third case relative to the baseline. In the second case, product innovation is smaller compared to the baseline if marketing in the static optimum would have been positive and unaffected compared to the baseline if it would have been zero. Therefore, expected product innovation in period 2 is unambiguously larger under experimentation in case marketing in the static optimum is zero. In case marketing in the static optimum is positive, the sign of expected product innovation depends on the prior and the amount of marketing in the static optimum.

Proposition 6. *In case there is experimentation in period 1, compared to the static optimum product innovation is*

1. *higher in period 1 compared to the static optimum,*

2. *higher in period 2 if marketing in the static optimum is zero,*

3. *higher in period 2 if marketing in the static optimum is positive if* $1 - p > n_S^*$ *and smaller otherwise.*

The only possibility for expected product innovation to be smaller in case there is experimentation, is when the prior probability of the novel technology being safe is smaller then the number of firms marketing in the static optimum. In this case the instances in which the novel technology turns out to be hazardous times the applications marketed in the static optimum is larger than the instances in which the novel technology turns out to be safe times the applications not marketed in the static optimum. For this to be the cases the total accident costs in case the technology is hazardous need to be small. In all other cases product innovation is higher in case there is experimentation, in particular when no application of the novel technology would have been marketed at all.

2.5 Extensions and Discussion

2.5.1 Extensions

We discuss two extensions. The first is allowing for different initial private benefits. The second is allowing for different amounts of firms and applications of the novel technologies.

Private Benefits

We now allow for private benefit curves of the form

$$b_S(n) = S \cdot (1 - n),$$

reflecting different stakes S. The following lemma shows that allowing for different stakes does not enrich the model in substance.

Lemma 9. *Let setting 1 be defined by* $b_S(n)$, p, $h(x) = H(x)$, n_{info} *and setting 2 be defined by* $b_1(n)$, p, $h(x) = \frac{H(S \cdot x)}{S}$, n_{info}.

1. *In settings 1 and 2, the following quantities coincide:* n_S^*, n_{exp}^*, n_{exp}^{SL}, n_{exp}^{ED}.

2. *The following quantities and functions are larger in setting 1 than in setting 2 by a factor of* S*:* x_S^*, $p \cdot h(x_S^*)$, VOE, COE.

Lemma 9 says that in settings 1 and 2 the same number of firms marketing are optimal and that the same liability regimes implement the same number of firms. The only

difference is in the stakes: Care costs, expected harm and private benefits are larger in setting 1 by a factor of S compared to setting 2. This is reflected in that the value and costs of experimentation also only differ by a factor of S across the two settings and implies that the same liability regimes do not only lead to the same outcomes, but also entail the same welfare levels up to the common factor of S. Therefore setting 1 and setting 2 are equivalent with respect to their positive and normative implications and all costs and benefits in the original model can be interpreted as fractions of the initial firm's private benefits. Since setting 2 is within the bounds of the original model, allowing for different stakes does not enrich the model in substance.

Lemma 9 tells us that for the model outcome only the relative size between the initial firm's private benefits and the total accident costs is important. Therefore, nothing is lost if we normalize S to one. In the model both the size of the private benefits from marketing and the size of the total accident costs are captured in the initial net benefits. An increase in the private benefits relative to the total accident costs corresponds to an increase in the initial net benefits in the model, with the consequences for experimentation in the social optimum and the performance of experimentation regimes relative to strict liability discussed in the preceding two chapters.

Number of Applications and Firms

We now allow for private benefit curves of the form

$$b_B(n) = (1 - \frac{n}{B}),$$

reflecting different numbers of applications of the technology. We call the number of applications the technology's 'breadth'. The following lemma shows that allowing for different breadths does not enrich the model in substance.

Lemma 10. *Let setting 1 be defined by $b_B(n)$, p, $h(x)$, n_{info} and setting 2 be defined by $b_1(n)$, p, $h(x)$, $\frac{n_{info}}{B}$.*

1. *In settings 1 and 2 the following quantities coincide: x_S^*, $p \cdot h(x_S^*)$.*
2. *The following quantities are larger in setting 1 than in setting 2 by a factor of B: n_S^*, n_{exp}^*, n_{exp}^{SL}, n_{exp}^{ED}.*
3. *The following functions are larger in setting 1 than in setting 2 by a factor of B when evaluated at B times the number of firms compared to setting 2:* VOE, COE.

Lemma 10 says that in settings 1 and 2 the same share of firms marketing are optimal and that the same liability regimes implement the same share of firms. The value and costs of experimentation differ by a factor of B across the two settings when evaluated

for the same share of firms. Therefore setting 1 and 2 are equivalent with respect to their positive and normative implications for the same share of firms and the number of firms in the original model can be interpreted as share of a population of firms. Since setting 2 is within the bounds of the original model and the outcome of setting 1 can directly be deduced from setting 2, allowing for different breadths of technologies does not enrich the model in substance.

Lemma 10 tells us that for the model outcome only the relative size between the number of firms and the information rate is important. Therefore, nothing is lost if we normalize B to one. An increase in the novel technology's breadth and an increase in the information rate per firm are both equivalent to an increase in the information rate in the model, with the consequences for experimentation in the social optimum and the performance of experimentation regimes relative to strict liability discussed in the preceding two chapters.

2.5.2 Discussion

We kept the model employed in this paper simple for reasons of clarity and tractability. We now discuss the implications of our model choices. We modeled the value of information as a piecewise linear function. This is clearly a simplification, but does not have substantial repercussions on our results compared to a strictly concave and strictly increasing function[7] with identical initial information rate. The only qualitative differences are that, given such a function, in case of positive marketing in the static optimum some experimentation is always optimal (not only for $n_S^* < n_{info}$) and that there are only interior solutions for positive amounts of socially optimal experimentation. Quantitatively, the socially optimal amount of experimentation would go down compared to our model in cases in which there is an interior solution in our model and go up in cases in which there is a strictly positive corner solution. The ability to implement the optimal amount of experimentation in case liability regimes can be tailored to the specific technology at hand does obviously not depend functional form of the value of information.

The transition at the full information threshold from positive slope to zero slope is an approximation of a continuously decreasing information rate. We identified two conditions for an experimentation regime to be superior to strict liability when liability regimes are not technology-specific. Condition (1) concerns settings in which total marketing is

[7]The type of function we have in mind here is of the following form:

$$\frac{n}{n + n_{info}} \cdot \text{VOPI}.$$

The initial information rate here is also $\frac{1}{n_{info}}$. It determines the how fast the rate of information decreases with marketing. The value of information asymptotically approaches the value of perfect information.

below the full information threshold and condition (2) concerns settings in which it is above. These conditions would merge into a single condition for a strictly concave value of information. However, this would not change the model outcomes qualitatively: For the results concerning the maximum experimentation discount one would simply have to replace the information rate by the average information rate. Quantitatively, this would imply a reduction of the maximum experimentation discount.

Two characteristics concerning the comparative statics directly depend on the information rate: Firstly, the maximum experimentation discount decreases with increasing information rate in case marketing in the static optimum is positive and total marketing exceeds the full information threshold. Secondly, the maximum experimentation discount can decrease with increasing initial net benefits in case marketing in the static optimum is zero and total marketing exceeds the full information threshold. Both of these characteristics derive from condition (2) and are still present in case of a strictly concave functional form for a very high initial information rate.

We employed a two-stage model with perfect information diffusion and perfectly correlated risk characteristics. Compared to our model, we expect that a model employing continuous learning in time would make the initial units of marketing relatively more important and valuable. Compared to our model, experimentation would then be particularly valuable in case of zero marketing in the static optimum. We expect that less than perfect information diffusion and imperfectly correlated risk characteristics would lower the value of experimentation. Allowing for different priors among agents would necessitate a complete reconceptualization of the information structure with uncertain consequences.

In our model, all applications of the novel technologies are marketed by different firms and applications belong to different markets. When single firms have control over a significant portion of applications, strategic considerations enter the marketing decision. Such firms would generally undertake some, but too little, experimentation, knowing that they can partially reap the informational benefits from experimentation in the future. Considering firms which compete with their applications on the same market would open up new strategic incentives with uncertain overall consequences.

2.6 Conclusion

How should and does tort law deal with uncertain risks emerging from novel technologies? Exemptions for harm materializing from novel and uncertain risks are ubiquitous in tort law. (Reimann 2003, Owen 2010). However, the classical analysis of tort law for uncertain risks (Shavell 1992) comes to the conclusion that strict liability implements the social optimum. We provide an economic rationale for the exemptions for novel and uncertain risks resting on the assumptions that post-market experience is essential in

learning about novel technologies' risk characteristics and that learning is a function of the cumulative market experience with the novel technology. The information generated by post-market experience is public. Marketing the novel technology therefore involves a positive information externality on top of the negative risk externality. Experimentation, i.e. marketing beyond the static optimum, can then be optimal due to the public value of information. Our interpretation of the exemptions prevalent in tort law is that they act as an experimentation discount on the expected harm for firms bringing the novel technology to the market. Thereby, regimes employing such exemptions can implement experimentation, whereas strict liability does not provide such a discount and does therefore not implement any experimentation. With this rationale, we offer a new mechanism to support the longstanding claim that strict tort regimes can cause innovative and novel products to be withheld from the market (Burk and Boczar 1993, Viscusi and Moore 1993, Finkelstein 2004) and the first one to explain this claim with specific reference to state-of-the-art defenses and unknowable risks (Connolly 1965, O'Reilly 1987, Fondazione Rosselli 2004).

When liability regimes can be tailored to the specific novel technology at hand, the first-best can be implemented as long as optimal experimentation does not exceed the amount of experimentation implemented by standard negligence. However, the exemptions provided for uncertain risks are general rules and not technology-specific. Courts have to adhere to those general rules, such that the first-best does not seem to be viable in the real world. Non-technology-specific liability regimes have to trade off implementing too much and too little experimentation across novel technologies. In particular, they trade off for how many novel technologies an experimentation regime is inferior to strict liability due to excessive experimentation and for how many novel technologies an experimentation regime does not implement any experimentation although some experimentation would be socially desirable. The European Products Liability Directive and the Restatement (Third) of Torts: Products Liability can serve as examples of regimes striking these trade-offs differently. Granting only comparatively narrow exemptions for uncertain risks, the former is an example for a relatively low experimentation discount, while the latter is closer to a standard negligence regime and is therefore an example for a relatively large experimentation discount.

The analysis presented in this paper offers a variety of avenues for further work and generalization. On the one hand there are generalizations possible within our framework. In order to identify the first-order effects of public post-market stage learning, we employed a two-stage model with discrete marketing choices in quantity and time, perfect and instantaneous information diffusion, perfectly correlated risks across the applications of the novel technology, applications of the technology which are neither substitutes nor complements and with each firm only having control over one application of the technology. Relaxing those assumptions would lead to more refined understanding of the relationship between learning about novel risks at the post-market stage and tort law.

On the other hand the framework we developed in this paper serves as an ideal starting point for thinking about the regulation of novel and uncertain risks more generally. The insights on post-market learning could and should be connected and related to issues of pre-market testing, post-market monitoring and direct regulation of novel substances such as enacted in the REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) regulation in the European Union.

Appendix

Proof of Lemma 3. Statically optimal marketing is $n_S^* = \max[0, B_{SOC}]$. The static social value of the initial firm marketing is B_{SOC}. If $n_S^* = 0$, all marketing is experimentation and the static social value from the initial firm marketing B_{SOC} is non-positive and is the initial marginal costs of experimentation. Since the private benefits of marketing decrease with slope one, the marginal costs of experimentation in that case are $-B_{SOC} + n_{exp}$. If $n_S^* > 0$, the initial marginal costs of experimentation are zero and the the marginal costs of experimentation in that case are n_{exp}. □

Proof of Lemma 8. Only the third statement remains to be shown.

$$\text{COE}_{ED} = \bar{c}_{ED} \cdot n_{exp}^{ED} = (\frac{1}{2}D(1 - B_{SOC}) - \frac{1}{2}B_{SOC}) \cdot (B + D(1 - B_{SOC}))$$

The statement follows from

$$\frac{d}{dB_{SOC}} = -\frac{1}{2}(d+1)(B_{SOC} + D(1 - B_{SOC})) + (1 - D)(D(1 - B_{SOC}) - B_{SOC})\frac{1}{2} = 0$$

The rationale behind this result is the following: The amount of experimentation is increasing in the initial net benefits. The average costs of experimentation are decreasing in the initial net benefits. For substantially negative initial net benefits, the amount of experimentation is close to zero, while the average costs of experimentation are relatively high. The change in the amount of experimentation therefore dominates the change in the costs of experimentation, which is a direct implication of the product rule. If the initial net benefits are close enough to zero, the situation is reversed. The larger the experimentation discount, the faster the amount of experimentation increases and the slower the average cost of experimentation decreases in the initial net benefits. Therefore, the larger the experimentation discount, the larger is the interval in which the costs of experimentation are increasing in B_{SOC}. □

Proof of Lemma 9. The care optimization problem is

$$\text{Setting 1:} \quad \min[x + H(x)], \quad \text{Setting 2:} \quad \min[x + \frac{H(S \cdot x)}{S}]$$

Results in setting 1 are referred to by a superscript '1', in setting 2 by a '2'. The optimization yields

$$x_S^{*,1} = S \cdot x_S^{*,2} \quad \text{and} \quad p \cdot H(x_S^{*,1}) = S \cdot \left(\frac{H(x_S^{*,1})}{S}\right) = S \cdot \left(\frac{H(S \cdot x_S^{*,2})}{S}\right)$$

It follows that

$$B_{SOC}^1 = B_{SOC}^2 \quad \text{and} \quad n_S^{*,1} = n_S^{*,2}.$$

This implies that static behavior is identical while static payouts are larger by a factor of S in setting 1. Therefore, VOI, VOE and COE are larger by a factor of S in setting 1, implying

$$n_{exp}^{*,1} = n_{exp}^{*,2}.$$

Lastly, for a given experimentation discount, behavior is determined by

$$b_S(n) - D \cdot (x_S^{*,1} + pH(x_S^{*,1})) = 0 \text{ and } b_1(n) - D \cdot \frac{(x_S^{*,1} + pH(x_S^{*,1}))}{S} = 0,$$

leading to identical outcomes on both settings for all liability regimes. □

Proof of Lemma 10. The claims about x_S^*, $p \cdot h(x_S^*)$, n_S^* and COE are evident. Payoffs at the same share of firms are identical. Since static marketing and the maximum number of firms marketing in setting 1 is larger by a factor of B compared to setting 2, we have

$$\text{VOPI}^1 = \text{VOPI}^2.$$

Since there are B times more firms in setting 1 at the same share of firms than in setting 2, but there is a B times higher information rate in setting 2, this implies the statement about VOE and therefore also the statement about n_S^*. The statements about n_{exp}^{SL} and n_{exp}^{ED} directly follow from the definition of the experimentation discount. □

Proof of Proposition 1. Only part three remains to be shown. In case we have an interior solution, equating marginal costs and marginal value of experimentation yields $\max[0, B_{SOC}] + r_{info} \cdot \text{VOPI}$. In case we have a corner solution the possible values are zero and $\max[0, B_{SOC}]$ if $B_{SOC} < 0$. and $\max[0, B_{SOC} - n_{info}]$ if $B_{SOC} \geq 0$. □

Proof of Proposition 4. Only the second and the third statement remain to be shown. Only the inequalities concerning the maximum experimentation discount habe to be shown. $r_{info} \cdot \text{VOPI} > \bar{c}_{ED}$ means

$$r_{info} \cdot \text{VOPI} > \frac{1}{2} D(1 - B_{SOC}) - \frac{1}{2} B_{SOC}$$

in case the initial net benefits are negative and

$$r_{info} \cdot \text{VOPI} > \frac{1}{2} D(1 - B_{SOC})$$

in case they are non-negative. The respective latter inequalities follow from these statements. $(1 - \max[0, B_{SOC}] \cdot r_{info}) \cdot \text{VOPI} > \text{COE}_{ED}$ means

$$\text{VOPI} > (\frac{1}{2} D(1 - B_{SOC}) - \frac{1}{2} B_{SOC}) \cdot (B + D(1 - B_{SOC}))$$

in case the initial net benefits are negative and

$$(1 - B_{SOC} \cdot r_{info}) \cdot \text{VOPI} > (\frac{1}{2} D(1 - B_{SOC})) \cdot (D(1 - B_{SOC}))$$

in case they are non-negative. The respective former inequalities follow from these statements. □

Proof of Proposition 5. Only $r_{info}^{min,ED}(D) > r_{info}^{min,opt}$ and $r_{info}^{max,ED}(D) > r_{info}^{max,opt}$ remain to be shown. The former follows from the convexity of the costs of experimentation. The latter follows since $r_{info}^{max,opt} = \frac{1}{B_{SOC}}$. □

Proof of Proposition 6. Only the third part remains to be shown. Let q be the increase in probability that the true state of the world is learned. Expected product innovation relative to the static baseline is then

$$q \cdot [(1 - p)(1 - n_S^*) - p n_S^*].$$

Expected product innovation is therefore positive if and only if $1 - p > n_S^*$. □

Chapter 3

Establishing Causation in Climate Litigation: Admissibility and Reliability*

*Co-authored by Martin Carrier, Timo Goeschl, Johannes Lenhard, Henrike Martin, Ulrike Niemeier, Alexander Proelß, Hauke Schmidt. We want to thank audiences at the Workshop on Tort Law and Innovation in Heidelberg 2016 and at the Workshop on Climate Engineering Liability and Regulation in Kiel 2017 for helpful comments. Timo Goeschl and I gratefully acknowledge funding by the German Research Foundation DFG under grant number GO1604-3.

3.1 Introduction

Climate litigation has attracted renewed interest as a governance tool both in the context of climate change (Thornton and Covington 2016, Marjanac et al. 2017, McCormick et al. 2017) and in the context of potential solar geoengineering[1] (Horton et al. 2014, Reynolds 2015). A key challenge in climate litigation is to assess the factual basis of causation (Horton et al. 2014, Marjanac et al. 2017, McCormick et al. 2017; also, compare recent cases: decisions to investigate by the Human Rights Commission of the Philippines, 2017; Court of Appeal Hamm, Lliuya v. RWE, 2017). This challenge becomes particularly salient for extreme events, such as prolonged droughts or excessive precipitation: Due to the inherent stochasticity of the climate system, attributing a single extreme event to human intervention into the climate system with certainty is impossible (e.g. Allen et al. 2007). The traditional legal 'but for' test of deterministic necessary causation is, therefore, not suitable in this context, and attribution based on probabilistic notions appears to be the only possible option.

The Fraction of Attributable Risk (FAR) has been proposed as a method to tackle the problem of causation in climate litigation (Allen 2003, Allen et al. 2007, Horton et al. 2014, Marjanac et al. 2017). Furthermore, individual attribution studies explicitly refer to liability (Otto et al. 2017) and potential legal implications of event attribution (Hauser et al. 2017). FAR is a standard concept in attribution science, and Hannart et al. (2016) argue that it can be interpreted as the probabilistic counterpart of the traditional legal 'but for' test in the context of event attribution. Researchers have applied this concept to extreme events, including heat waves, cold spells, droughts and floods at an increasing rate (Herring et al. 2016a, Herring et al. 2018a), and by now an 'attribution community' has come into existence within climate science. FAR essentially quantifies the fraction of the total probability of an event which can be traced back to a climate alteration (e.g. greenhouse gas emissions or a deliberate climate intervention). In this paper, we assume FAR to be the relevant concept, although the discussion applies equally to the potential use of other concepts of attribution science in a trial.

Against this background, the paper makes two contributions. The first is to demonstrate that FAR can only be an effectual tool for resolving questions of causation if and to the extent to which evidentiary standards used by courts adequately accommodate the type of scientific evidence that FAR estimates represent. Evidence must be legally admissible in order to be considered at all by the court. Excluding FAR estimates from a trial on the basis of admissibility requirements denies them a role in resolving issues of causation. However, relaxing admissibility requirements affects the reliability of FAR estimates that parties bring to court: It gives parties that have material interests in the outcome of

[1] Solar geoengineering is a novel, untested, but potentially effective form of intervention in the global climate system with the purpose of counteracting global warming. Deployment, however, will entail the risk of undesirable side effects (Schäfer et al. 2015, Ocean Studies Board and National Research Council (2015), Niemeier and Tilmes 2017).

the trial greater latitude to choose those methodologies that produce FAR estimates most favorable to their position. Typically, such FAR estimates will not accurately reflect the true statistical relationship between climate alteration and extreme event. Thus, rules on scientific evidence need to strike a balance between, on the one hand, the admissibility of FAR evidence and, on the other, maintaining the epistemological quality, i.e. reliability, of FAR evidence. In highlighting the need to strike this balance in legal admissibility criteria, the present paper contributes to an emerging literature that assesses the potential of attribution science for solving the problem of causation in climate litigation (Lusk 2017, Marjanac and Patton 2018).

The second contribution of the paper is to apply the first contribution to a specific proposal for how to accommodate FAR estimates in evidentiary standards by modifying an existing set of admissibility criteria. We use as the object of this application the *Daubert* standard. This standard offers a set of specific criteria for admissibility. It applies to the United States, where climate liability suits have been launched, and is influential in other jurisdictions such as Canada and the UK. The *Daubert* standard consists of five criteria that, unmodified, would exclude every FAR estimate on the basis of one criterion ('*testability*') and be inapplicable to FAR estimates on another ('*error rate*'). A simple elimination of these two criteria would ensure that FAR estimates are principally admissible, but would also allow parties to introduce unreliable FAR estimates based on biased methodological choices. We argue, however, that a modified set of criteria, including criteria directly aiming at the reliability of the FAR estimates, would be capable of leading to both admissible and reliable FAR estimates. We furthermore discuss other relevant factors, including the type of the extreme event, the existence of a framework to determine the reliability of FAR estimates, and such a framework's accessibility to courts.

3.2 Poland vs. Australia: A Fictitious Tort Case

The challenges of establishing causation in climate litigation are significant, but can appear arcane to the novice. To provide tangibility, we introduce a fictitious case based on a solar geoengineering (SG) scenario, developed for the Geoengineering Model Intercomparison Project (Kravitz et al. 2011). It involves the calculation of FAR estimates based on two climate models and serves as illustration for the problem of causation discussed in this paper. We choose a case based on a SG scenario, since it can be framed with two clear parties and thus avoids specifying details of a climate change trial which are important (compare Otto et al. 2017), but not essential for the message of the present paper. Technical details supporting the specifics of the scenario can be found in the appendix.

The year is 2060. Australia has been deploying stratospheric aerosol injection for several years by now. Its goal is to stabilize radiative forcing at the levels present in the year 2020, thus offsetting the effect of the rise in atmospheric greenhouse gas concentration levels. This year, Poland, an area of significant grain production, experiences severe drought conditions. There is no precipitation for 40 days in a row. Plants wither; the grain harvest is severely reduced and can only be partially salvaged through expensive emergency irrigation. Estimates put the economic damages at 18€ billion.

Poland has experienced droughts in times before SG. The duration, intensity, and resultant damages of the 2060 drought are unusual, however. Some climate models give reason to suspect a link. According to these models, the region within which Poland is situated generally faces drier conditions under the current climate compared to a climate without SG activities. These results provides grounds for claiming that Australia's actions have made the drought more probable. Poland decides to appeal to an international court to hold Australia liable for the drought and to receive compensation.

In court, Poland presents evidence based on data from the climate model developed at the Max-Planck-Institute, the MPI-ESM. To quantify the relative contribution of Australia's SG deployment to the probability of the drought, the evidence employs the method of Fraction of Attributable Risk (FAR). The evidence reports a FAR of 0.83, meaning that the estimated probability that Australia deploying SG was necessary for the event to occur is 83%. Poland lays out on which modeling choices and assumptions its assessment rests: It presents the simulations based on the MPI-ESM, explains how it characterized the drought in terms of duration and spatial extent, that it used the climate index of Consecutive Dry Days (CDD) and which statistical techniques were employed to perform the FAR estimation.

Australia challenges Poland's claim in court. It presents a set of simulations based on the HadGEM model developed by UK Meteorological Office. Australia's quantification yields a FAR of only 0.18, supported by a methodological brief similar to Poland's. The court is thus confronted with diverging and complex scientific evidence that must be assessed in order to adjudicate the case. Both parties have provided evidence to the court and both have introduced FAR estimates to support their case. However, Poland's and Australia's FAR estimates differ substantially. Ultimately, the court needs to come to an assessment of the merits of the evidence and to make a judgment on existence or absence of a causal link between Australia undertaking SG and Poland's harm.

3.3 Admissibility, Reliability and Evidence Production

Behind the specifics of the fictitious case presented above, there lies a general problem of admissibility, reliability and causation in climate litigation. In abstract terms, it can be captured by a setting featuring one applicant ('victim') who claims before a court to have

suffered damage from an extreme event allegedly caused by one respondent ('injurer'). Each of the two parties brings forward evidence in the form of FAR estimates and is in control over the methodological choices underlying their respective estimate. It is from these choices that the contestability of FAR estimates derives: There exist various degrees of freedom for the methodological choices underpinning a FAR estimate. These choices can impact significantly on the magnitude of the FAR estimate (Stott et al. 2016, Stott et al. 2018) and different choices may lead to different conclusions (Hauser et al. 2017). In climate litigation, the parties have material interests in the outcome of the trial. Their control over the methodological choices that generate FAR estimates therefore calls for some evidentiary standard.

We conceptualize the court's assessment of the evidence as a two-step process.[2] The first step constitutes the evidentiary standard: The court determines whether a FAR estimate is admissible. Only admissible evidence can be taken into account by the court in the second step. The admissibility test does not evaluate the outcome of a FAR estimate, but its underlying methodology. In the second step, the court evaluates all the admissible evidence by weighting it in order to come to a judgment. In this second step, the court depends on the reliability[3] of the evidence. FAR evidence can only help the court to arrive at a legally correct judgment to the extent that it accurately reflects the true statistical relationship between climate alteration and extreme events. Therefore, FAR evidence has to be both admissible and reliable.

The more lenient the evidentiary standard is, the easier it is for FAR evidence to pass the legal hurdle of admissibility. However, the nature of the admissibility test has repercussions for the reliability of the FAR estimates that parties will produce. Applicant and respondent differ in their interest vis-à-vis the desired outcome of the trial. In a trial, therefore, the methodological choices underlying FAR estimates have to be conceptualized as strategic choices aimed at maximizing the likelihood of prevailing in court. The parties make these strategic choices in light of how the court assesses them. The more lenient the admissibility test, the greater the latitude to choose methodologies that produce FAR estimates most favorable to each party's position. However, the more a desired result of a FAR estimate determines its underlying methodology, the more biased and the less reliable is the FAR estimate.

[2]This a simplification of the legal reality in that the two steps are not formally separated in all jurisdictions. However, for the purposes of this paper, the simplification is innocuous. For details, we refer the reader to the legal appendix.

[3]We use the terms reliable and reliability in a strictly epistemological and not in a legal sense here, in accordance with the following definition: 'An object (a process, method, system, or what have you) is reliable if and only if (1) it is a sort of thing that tends to produce beliefs, and (2) the proportion of true beliefs among the beliefs it produces meets some threshold, or criterion, value'. (Goldman 1986, p. 26). In legal contexts, 'evidentiary reliability' of evidence is sometimes used as a criterion for admissibility. In social sciences other than law, 'reliability' is usually employed as a prerequisite of 'validity' and has he meaning of consistency or repeatability (Carmines and Zeller 1979). In order to avoid terminological confusion, we want to emphasize that we exclusively refer to the epistemological definition. For a more detailed discussion of these terms and definitions, see Haack (2008).

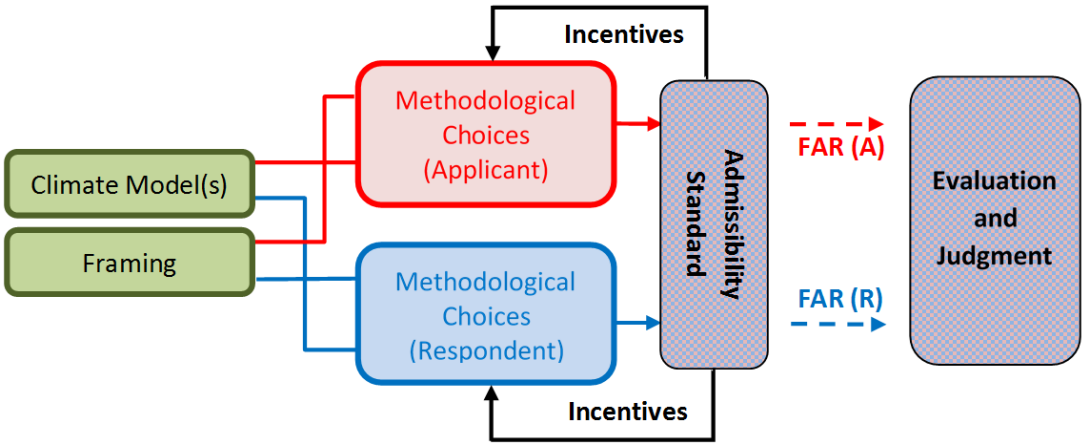

Figure 3.1: Model of Production and Assessment of FAR Evidence.

Normatively speaking, the admissibility test ought to ensure that reliable FAR estimates are admissible and that unreliable ones are not. However, FAR estimates cannot be perfectly reliable due to limits in the understanding of the climate system – their reliability is not absolute, but gradual. How well FAR estimates can help solve the issue of causation in climate litigation therefore crucially depends on how well rules on scientific evidence strike the balance between the admissibility of FAR evidence and incentivizing their reliability (compare Figure 3.1).

In the following, we will provide a tight definition of the FAR, introduce the *Daubert* standard, and discuss the nature of the methodological choices underlying FAR estimates in order to examine, in the next section, how well the *Daubert* criteria strike a balance between admissibility and reliability for FAR evidence.

The Fraction of Attributable Risk

FAR is a method to determine the relative contribution to the probability of an event occurring (e.g. the drought in Poland) by a specific condition (e.g. SG). If the probability of the drought is P_0 in absence of SG (the *counterfactual*) and the probability in the presence of SG (the *actual* climate) is P_1, SG increases the total probability of the event by $P_1 - P_0$. The fraction of the total probability of the event occurring, P_1, attributable to SG is then

$$\text{FAR} = \frac{P_1 - P_0}{P_1} = 1 - \frac{P_0}{P_1}$$

For example, a FAR of 0.5 means that the probability for a drought in Poland in the presence of SG is twice the probability in its absence – in other words half of the drought's risk is then attributable to SG. In order to estimate P_0 and P_1, model simulations are run for the *actual* climate and the *counterfactual* climate.

Methodological Choices: Climate Model(s) and Framing

We divide the methodological choices underlying a FAR estimate into two modules. The first module concerns the choice of one or more climate model(s) from the set of available models. In the fictional case, applicant and respondent choose different climate models (the HadGEM and the MPI-ESM, respectively). The second module, the framing of the event, consists of the choice of metrics and definitions used to characterize the event, including the statistical tools employed to process the climate model data for the metrics and definitions used. In the fictional case, the event is framed as the incident of 40 consecutive dry days, with an implicit set of definitions of what constitutes a 'dry' day, a certain spatial definition of the event, etcetera. For details on the framing choices employed in the fictional case, we refer the reader to the technical appendix.

The reliability of a FAR estimate hinges on both modules: Firstly, on the reliability of the climate model(s) employed for simulating the event as it has been framed, and secondly, on the framing of the event. The framing has to be valid, meaning that 'it measures what it purports to measure' (Carmines and Zeller 1979). The framing of an event is valid if it actually captures the event it purports to describe. The difference between reliability and validity is that reliability refers to a process that yields a numerical value, while validity refers to the suitability of certain choices for the description of a phenomenon or event. Both are not absolute concepts and to be understood in a gradual sense.

The Daubert Criteria

Courts have struggled for a long time with how to handle scientific evidence. Notably, in international law, which is the relevant regime for inter-state litigation, no rules or generally accepted criteria concerning the admissibility of scientific evidence exist. Criteria that have been developed within domestic jurisdictions vary mostly due to differences between common and continental legal traditions. However, at a certain stage of every single trial, a court will (implicitly or expressly) be called upon to decide whether it is willing to take into account the evidence provided by the parties. The most advanced criteria for this test have been developed in the US legal system in the shape of the *Daubert* standard. This standard consists of five criteria ('*Daubert* criteria'), which will be used here as an example of a procedural gateway that could also be referred to in inter-state disputes in future. The *Daubert* criteria are:

I. whether the theory or technique in question can be and has been tested ('*testability*'),

II. whether it has been subjected to peer review and publication ('*peer review*'),

III. its known or potential error rate ('*error rate*'),

IV. the existence and maintenance of standards controlling its operation ('*standards of control*'), and

V. whether it has attracted widespread acceptance within the relevant scientific community ('*general acceptance*').

3.4 Admissibility and Reliability of FAR Estimates under Daubert

3.4.1 Admissibility of FAR Estimates

It is useful to distinguish between two types of reasons leading to the inadmissibility of a FAR estimate. The first is that any way of using the method of FAR is as a matter of principle inadmissible. The second is that only certain methodological choices are inadmissible. The inadmissibility of FAR estimates on principle grounds does not seem to be optimal. In this, we rely on the judgment of a report of the *US National Academy of Sciences* (NAS 2016) on the attribution of extreme weather events. It concludes, "it is now often possible to make and defend quantitative statements about the extent to which human-induced climate change has [...] influenced either the magnitude or the probability of occurrence of specific types of event or event classes." Furthermore, this sentiment resonates with similar assessments in the two most recent of the annual reports on extreme events published in the *Bulletin of the American Meteorological Society* (BAMS).

We separate the *Daubert* criteria into two groups. The criteria *testability* (I) and *error rate* (III) directly refer to FAR estimates' scientific merit, and we refer to them as substantive criteria. They may render FAR estimates inadmissible on principal grounds. We refer to the criteria *peer review* (II), *standards of control* (IV) and *general acceptance* (V) as formal criteria. They relate to the status FAR estimates have within the scientific community. *Peer review* and *general acceptance* are arguably already fulfilled today for the method of FAR as such. FAR is introduced as the 'commonly accepted event attribution technique' in the latest annual report in the BAMS (Herring et al. 2018b), and there is a large literature employing the concept FAR for event attribution as documented by the annual BAMS reports. There are general recommendations for performing event attribution in NAS (2016), which can be interpreted as *standards of control* for the method of FAR as such. The more interesting role of the formal criteria is therefore in examining specific methodologies. We start out by discussing the substantive criteria and their potential to render FAR estimates inadmissibility on principal grounds.

Testability demands falsifiability (Popper 1959) of the theory underlying the evidence presented.[4] This requirement presents an insurmountable hurdle for FAR estimates due to at least two reasons, and its application would thus likely result in any FAR estimate being inadmissible: Firstly, any FAR estimate relies on climate model output. A climate model cannot be expected to exactly reproduce the climate system. It is a tool, which can be used to investigate and understand the climate system. Physical processes represented in climate models are, due to computational constraints, generally subject to approximations and parameterizations in climate modeling. In a strict sense, climate models are therefore known to be false, and the criterion of falsifiability is not appropriate (Otto 2012). Instead of looking for universal properties, it is more appropriate to ask whether a given model is "unusable in answering specific questions" (McAvaney 2001), and a model should be limited to particular domains, while accepting a particular level of inaccuracy (Petersen 2012). Secondly, a necessary part of a FAR estimation is estimating the probability of the event concerned happening in the *counterfactual*. By definition, the *counterfactual* cannot be observed in reality. Both the assumptions regarding the *counterfactual* and a climate model's reliability for the *counterfactual* cannot be compared to real world measurements.

Applying the third criterion (*error rate*) in a strict sense to FAR estimates is also problematic. The method of calculating a FAR has itself no error rate. Any such rate derives from errors in estimating the probabilities P_0 and P_1 and depends strongly on the specific circumstances of the estimation, such as the event in question and the methodology used. Definite numerical error rates are not suitable for assessing climate model performance. Furthermore, whether framing an event a certain way is valid can ultimately not be determined in a quantitative way. Lastly, error rates are not applicable to the *counterfactual*, since such error rates require experimentation in controlled environments.

The preceding assessment resonates with the legal literature on the *Daubert* standard, which comes to the conclusion that the *Daubert* criteria (I) and (III) are not suitable for the intricacy of methods and procedures which characterize current scientific practice more generally (Jasanoff 2005). Specifically, *testability* has been criticized for being ill-suited for the courtroom (Haack 2010). Both criteria exclude evidence from scientific fields in which classical experimentation is not possible as well as evidence from the modeling of complex systems. In those cases, the *Daubert* standard often produces 'evidence-narrowing' decisions (Heinzerling 2006), implying that plaintiffs frequently would lose cases they should win (McGarity 2004, Wagner 2005, Swinehart 2008).

From a naive point of view, one could argue that the *Daubert* criteria (I) and (III) should simply be abandoned, rendering FAR estimates admissible in general and allowing courts to include the information FAR estimates can provide into their evaluation. However, the incentive problem raised in the previous section then comes into play. In the absence

[4] Note that the US Supreme Court explicitly refers to Popper in its judgment in which it developed the *Daubert* criteria, and it explicitly equates *testability* with falsifiability.

of substantive criteria in the legal gate, the latitude of parties to choose, for example, a certain climate model would be too large. In absence of substantive criteria, the formal criteria can only be directly applied to individual methodological choices. However, this is unlikely to present a real hurdle to methodological choices. For example, all climate models frequently employed in scientific studies would arguably pass the criteria of *peer review* and *general acceptance*, and there exist at least some *standards of control* for individual climate models. However, not any such climate model is equally suitable for simulating all extreme events. Parties could take advantage of this situation, and courts would then receive considerably less reliable evidence from both parties than might be produced. In this light, the question of an appropriate modification of the *Daubert* standard and alternatives to criteria (I) and (III) arises. While the concept of *testability* is entirely unsuitable for examining FAR estimates, this does not mean one cannot test their performance in specific settings and the principal idea of assessing an error rate is sensible.

3.4.2 Reliability of FAR Evidence Produced

The purpose of introducing alternative substantive criteria suggests that such criteria should directly demand reliability. Reliability in itself is a quantitative concept. While the reliability of a FAR estimate cannot be directly measured or observed due to the same reasons as for an error rate, reliability can be and often is argued for. For example, climate model reliability for specific tasks is usually quantitatively (e.g. model validation) and qualitatively (e.g. validity of assumptions, process understanding) argued for. This distinguishes reliability from a purely numerical error rate, which has ultimately to be obtained by experimentation. The reliability of FAR estimates hinges on the climate model reliability and the validity of the framing. Instead of one criterion demanding the reliability of FAR estimates, separately demanding climate model reliability and framing validity helps clarify the different roles both concepts play in arriving at a reliable FAR estimate. While unreliable climate model data likely does not correctly represent the event as it has been framed, an invalid framing does not capture the actual event as such and even a reliable climate model then simulates a fictitious event. We remind the reader that, while reliability is a quantitative concept, validity is a qualitative one.

In order to be meaningful, climate model reliability and framing validity always have to refer to a specific event. Before formulating the alternative substantive criteria, we would like to emphasize that this point is indeed crucial. Climate model reliability for simulating extreme events depends on the type of the event, the region and the climate model. Christidis et al. (2013) find that the reliability of the HadGem3-A model for extreme event attribution varies with region and type of the event. They examine three different cases, concluding that model reliability is high regarding the attribution of the 2009/2010 UK cold winter and the July 2010 Moscow heat wave, but reliability is low

regarding the Pakistan floods in July 2010. NAS (2016) states that "[a]ttribution is more feasible for some events than for others' and 'the optimal choice of [...] model will depend on the question being addressed and the event under consideration" (compare also Stott et al. 2016). Sillmann et al. (2013) evaluate the performance of various climate models with respect to many different metrics of climate extremes, concluding that the performance is greatly dependent on both the metric and the climate model under consideration. Summing up, the question of climate model reliability hinges on the demonstration of reliability for the specific event and cannot be meaningfully answered for a climate model in isolation.

Key part of an event's framing is the definition of its spatial and temporal dimension, which metric or extreme index is used to describe the event, the threshold defining the event, whether a conditional or unconditional modeling approach is used, the statistical tools employed and, in the case of SG, how the intervention in question is modeled. The attribution science community emphasizes the importance of how an event is framed (NAS 2016, Stott et al. 2016, Stott et al. 2018) and that an "[a]ttribution result can depend strongly on the definition of the event" (Stott et al. 2016). For example, there are various potential choices of relevant extreme indices for a given extreme event: A warm or cold spell can be defined in terms of duration, intensity or a combination of both; droughts can be defined as meteorological, hydrological, socioeconomic and agricultural (Wilhite and Glantz 1985). The relative merit of different choices depends on the type of the event, the geographic and climatological characteristics of the relevant region and the specific aspects important to the legal claim. While there never is a single-best way to frame an event, but a range of valid and scientifically defensible framing, there are clearly also less valid and invalid framing choices for a given event.

Modifying the third *Daubert* criterion *error rate* into two criteria, demanding '*sufficient climate model reliability for the specific event*' (IIIa) and '*sufficient validity of the framing for the specific event*' (IIIb), would accommodate the need for suitable substantive criteria. It furthermore highlights the importance of taking into account the specifics of a given event in the assessment of climate model reliability and framing validity. The court needs to decide in each specific case how to strike the balance between admissibility and FAR estimate reliability. This balance may depend on, for example, how settled the science of attribution is for a specific type of event or the availability of other evidence and is reflected in the word 'sufficient'.

The formal criteria (II), (IV) and (V) can serve as a basis for carrying out the assessment of reliability and validity by the substantive criteria. Having established that methodologies can only be meaningfully examined with respect to a specific event, the formal criteria can only be meaningfully applied to the methods with which the reliability and validity of the chosen methodology is demonstrated or attacked, not to a methodology in and of itself. For example, it is not important whether a climate model is *generally accepted* for being used in attribution studies. However, it is important whether the

climate model in the specific context satisfies *generally accepted* methods of establishing climate model reliability. *Peer review*, *standards of control* and *general acceptance* may be used as different hurdles to clear, reflecting different applications of what is deemed 'sufficient' by a court in a given context. The requirements may differ between types of events, different parts of the methodology and may depend on the availability of other evidence. In some contexts, attribution science may be settled in a way such that there are actual *standards of control*, for others there may be *general acceptance* of a certain methodology and for some there may simply be single examples in the *peer reviewed* literature.

The court's assessment necessarily has to rely on the methods for assessing extreme event methodologies developed by the attribution science community. In that, a court is constrained in two ways. The first is that the court's assessment is limited by the best-possible assessment given the contemporary epistemology of extreme event attribution. To date, there is no settled epistemology of climate modeling. However, there is consensus on the basic principles for extreme event model evaluation in the attribution community: the model's ability to simulate the event, the understanding and representation of the processes driving the event and the quality of the observational record (Herring et al. 2016b, Stott et al. 2016, NAS 2016). The extent to which there is an agreed-upon framework or operationalization of these basic principles is not clear. For example, NAS (2016) cites three studies with differing approaches to model evaluation, concluding that "[s]uch evaluations are necessary, but they are not a sufficient demonstration of model quality", and reemphasizes the role for the mechanisms producing variability and extremes, as well as their representation in the model in the assessment of a climate model's reliability for a specific event. The extent to which conclusions can be drawn on a model's reliability for the *counterfactual* seems to date to be an open question (NAS 2016), and studies usually simply assume that the model's reliability is the same for the *factual* and the *counterfactual* (e.g. Christidis et al. 2013). In light of differing findings dependent on the climate model choice in attributing the 2015 European drought to climate change, Hauser et al. (2017) call for "multi-model and multi-method based event attribution". Sillmann et al. (2013) find that multi-model approaches generally outperform individual models in reproducing historic extreme index reanalysis data.

How different framing choices impact on the outcome of a FAR estimate is subject of an ongoing debate in the attribution community (Stott et al. 2018). Currently, there is a discussion regarding the merits of different choices of statistical tools and paradigms (Stott et al. 2017, Mann et al. 2017), and Hauser et al. (2017) obtain differing results for the attribution of the 2015 European drought to climate change, depending on framing and climate model choices. Furthermore, different metrics lead to different attribution assessments for the last California drought (Seager et al. 2015, Diffenbaugh et al. 2015, Williams et al. 2015). There is consensus that the strongest conclusions about an event

can be drawn when FAR estimates based on different framing choices of an event and on different climate model choices yield consistent results (NAS 2016, Hauser et al. 2017, Stott et al. 2018). While the attribution community clearly emphasizes and appreciates the importance of framing choices, by the attribution community's own account this is an area where progress can still be made.

The second way in which a court is constrained in the assessment of methodology is how well the contemporary epistemology of extreme event methodology is accessible to the court. If an implicitly agreed-upon framework (potentially in part) suitable for answering questions of reliability and validity exists within the attribution science community, but is not explicitly drafted or in any other way identifiable by the court, the framework's mere existence is of little help to a court. A court does not have the expertise to tell the scientific merits of competing arguments concerning model reliability by experts by itself. However, it may well be equipped to evaluate which line of argumentation is in line with certain well-defined and clearly framed standards or operationalizations of the basic principles of model evaluation and which is not.

Both constraints jointly determine how well a court can assess the reliability of a FAR estimate. The more accurately a court can distinguish between reliable and unreliable FAR estimates, the better it can include the former while keeping the latter out, and the finer it can strike the trade-off between admissibility and reliability. The literature on the Russian heat wave in 2010 might serve as illustration here. Dole et al. (2011) concluded that the event was largely natural, while Rahmstorf and Coumou (2011) concluded that the anthropogenic influence was significant. This apparent contradiction could be resolved by understanding that Dole et al. (2011) focused on the magnitude of the event, while Rahmstorf and Coumou (2011) focused on the occurrence frequency of the event (Otto et al. 2012). It is easily imaginable that a court could be confused about the validity of the two approaches for resolving the question of causation without the clarification in the literature. The difference between 'frequency' and 'magnitude' has, by now, also been explicitly discussed in NAS (2016), making it much less likely that a court would overlook the contribution, misunderstand the point, would be deceived about the point, or be doubtful about the scientific consensus in this respect. Using the 'frequency' approach for answering questions of causation might be an example which courts identify as a *standard of control* and exclude any 'magnitude' approach.

A different example are high temperature extreme events. The latest annual report in the BAMS states that the "majority of heat papers now use a widely established and accepted methodology" (Herring et al. 2018b). Here, courts might ascribe *general acceptance* to such methodologies, at least to the extent that they can identify the details of the, according to Herring et al. (2018b), "established and accepted methodology". In contrast, assessing a bias correction undertaken on climate model output might present difficulties to a court. Stott et al. (2016) warn that "bias corrections methods need to be applied" in some cases, but "should be applied with caution", while NAS (2016)

simply states that "some bias correction will almost certainly be required" and "the validity of this must be established". Shiogama et al. (2013) note that their results were sensitive to bias correction. A court today would probably not be able to assess how a given bias correction would affect the reliability of climate model data – even if it might be blatantly obvious in that specific instance to a neutral attribution scientist. A court might therefore only be able to check whether *peer reviewed* studies have employed bias corrections in similar circumstances or not.

Irrespective of the legal institutions, the contemporary state of the art and the fundamental ability of climate models to perform event attribution set an upper limit on climate model reliability (Trenberth et al. 2015, Otto 2016). While future advances in climate modeling may well further improve this state of the art, we refrain from discussing its potential future evolution. However, the extent to which future advances actually improve the reliability of FAR estimates in a trial depends on how issues we raised in this paper are resolved.

3.5 Conclusion

Causation is a key challenge in climate litigation and attribution science has increasingly been brought up as a potential means to resolve that challenge. Usually, the admissibility of climate model based evidence, such as a Fraction of Attributable Risk (FAR), is at the center of the discussion. We argued that such evidence is only effectual if it is both admissible and reliable. While inadmissible evidence is excluded from trials, unreliable evidence does not help in arriving at the right conclusion. The parties of the trial have a material interest in the outcome of the trial. Lowering standards of admissibility impacts on the reliability of evidence, since it gives the parties more leeway in choosing methodologies leading to FAR estimates favorable to their case. FAR can therefore only be an effectual tool for resolving questions of causation to the extent that relevant legal rules on scientific evidence strike the right balance between the admissibility of FAR evidence and the reliability of FAR evidence.

We provide a specific proposal for how to accommodate FAR estimates in evidentiary standards by modifying an existing set of admissibility criteria, the *Daubert* standard. This standard offers a set of five specific criteria for admissibility and applies within United States, where climate liability suits have been launched, and is influential in other jurisdictions such as Canada and the UK. Two of these criteria are substantive in that they address the scientific merit of evidence. One of these two criteria ('*testability*') would exclude every FAR estimate. The other ('*error rate*') is inapplicable to FAR estimates, since numerical error rates are ultimately dependent on experimentation, which is not feasible in event attribution. We argued that dropping the first one, and modifying the second one into two criteria directly addressing the reliability of FAR

estimates, is capable of leading to admissible and reliable FAR estimates. FAR estimate reliability depends on both climate model reliability and framing validity. Accordingly, the alternative criteria demand '*sufficient climate model reliability for the event in question*' and '*sufficient framing validity for the event in question*'. The other three, formal, criteria *peer review*, *standards of control* and *general acceptance* can serve as a basis for carrying out the assessment of reliability and validity through the substantive criteria. The word 'sufficient' reflects the trade-off necessary between admissibility and reliability. The optimal trade-off is likely to differ between different types of extreme events, since attribution science is more advanced for some types of events, like temperature-related ones, than for others. This implies that how well the problem of causation can be solved at best depends on the type of the event.

In assessing the reliability of FAR estimates, a court is dependent on the contemporary epistemology of extreme event modeling and the accessibility of this knowledge to the court. The combination of the two sets an upper limit on how well courts can assess the reliability of FAR estimates, which determines how well a court can strike the trade-off between admissibility and reliability. NAS (2016) and the annual special reports on event attribution in the BAMS are important steps in making the relevant knowledge accessible and identifiable to courts in potential future cases. Especially with ongoing and future advances in attribution science in mind, we can only encourage the compilation of similar reports in the future.

Technical Appendix: Fictitious Tort Case

We base the fictitious court case on simulations performed for the Geoengineering Model Intercomparison Project GeoMIP (Kravitz 2011). The purpose of the case is not to perform a credible FAR analysis. Its purpose is, firstly, to demonstrate the variety of choices to be made even in an analysis only going through the basic steps of a FAR estimation. Secondly, it shows that, at least when naively executed, FAR estimates can widely differ when different climate models are chosen, while all other choices are identical.

We chose scenario G3 from GeoMIP as the relevant scenario. In G3, solar geoengineering (SG) is deployed by injecting sulfur into the stratosphere. The basis of G3 is the RCP4.5 scenario (e.g. Meinshausen 2011) of the fifth phase of the Coupled Model Intercomparison Project (CMIP5). Sulfur injections in G3 start in year 2020 and go on for 50 years. The injections are designed to keep the top of the atmosphere radiative forcing constant at the level of 2020, despite the increasing greenhouse gas concentrations given in the RCP4.5 scenario. For the analysis of both the G3 and RCP4.5 simulations of both climate models (see below) three ensemble members have been available.

The climate models we use in the fictitious court case are in all likelihood different from climate models in any potential future climate litigation. The way we use the climate model output to arrive at FAR estimates constitutes a radical simplification of the state of the art in event attribution. Our FAR estimates therefore do not remotely reflect the reliability of FAR estimates in the literature or in a potential future trial and it is a futile task to estimate today how reliable future simulations of extreme events may be. However, as noted above, the purpose of the case is not to perform a credible FAR analysis and our aspiration is not to speculate about the future development of attribution science. A further difference to a potential future case is that we compare two fictitious future climate states, while in an actual case one of the simulations would need to describe the factual climate of that time. However, also for an actual and likely transient climate state the evaluation of simulated low-frequency extreme events is inherently difficult.

We let the opposing parties in the fictitious court case choose the two models described below just because of opportunity. Other models that simulated the G3 scenario had either no ensemble available, the warming in the stratosphere seemed to be unrealistic or coupling of the aerosols to radiation was not available or erroneous. The latter is necessary to simulate a realistic radiative forcing of the sulfate aerosols.

The choice of Poland and Australia has no political background. However, it is not completely arbitrary. An analysis of precipitation by Aswathy et al. (2015) and a metric for surface dryness in G3 simulations indicated that Australia could gain from sulfur injections by increased winter precipitations in many areas. Poland is part of an area for which at least one of the models analyzed by us indicates a potential impact of SG. In addition, the choice of a drought case is somewhat arbitrary, but informed by the well-known result that SG affects the water cycle (e.g. Tilmes et al. 2013). Some extremes like hurricanes or flash floods in mountainous regions are small-scale phenomena and need a much finer grid resolutions than used in most present-day climate models to be simulated explicitly. Droughts, however, are often related to large-scale stationary weather patterns that may be more realistically represented in these models.

Model description

The claims of Poland base on the Max-Planck-Institute Earth System Model (MPI-ESM, (Giorgetta et al. 2013) while Australia bases its analysis on the HadGEM2-ES, the UK Met Office Earth System Model (Collins et al. 2011). Both models are state of the art climate models used in particular for the CMIP5 simulations.

The atmospheric part of the HadGEM2-ES is also used as weather forecast model. Further, the model includes components for the simulation of tropospheric chemistry, aerosols, land surface and hydrology, terrestrial carbon cycle, ocean sea ice and ocean

biogeochemical processes. In the model configuration used for the G3 simulations an aerosol microphysical model is included that explicitly simulates the evolution of the sulfur from its injection in the form of SO2 to sulfate aerosols.

The MPI-ESM is a state of the art coupled three-dimensional atmosphere-ocean-land surface model, with a well-represented stratosphere and an interactive carbon cycle. It consists of an atmosphere and ocean component and includes submodels for land processes, vegetation, and ocean biogeochemistry. Within the MPI-ESM, in contrast to the HadGEM2-ES simulations, the stratospheric aerosol layer is prescribed via its optical properties (Niemeier et al. 2013). It was pre-calculated in simulations with ECHAM5-HAM (Stier et al. 2005), a general circulation model coupled interactively to an aerosol microphysical model. Simulations with different injection rates were performed (Niemeier et al. 2011) and the resulting aerosol optical properties interpolated to monthly values for use in the MPI-ESM.

Metric used

Consecutive dry days (CDD) is a standard extreme index used in climate modeling. It is part of the list of ten extreme indices in Frich et al. (2002), emerging from the meeting of the World Meteorological Organisation (WMO) Commission for Climatology (CCl)/Climate Variability (CLIVAR) Working Group on Climate Change Detection in September 1998. However, it does not exhaustively describe the phenomenon 'drought' and other extreme indices could be chosen as well. Furthermore, droughts can be defined in various categories, like meteorological, hydrological, socionomic, and agricultural droughts (Wilhite and Glantz 1985). Besides relatively simple extreme indices like CDD, there exist more complex metrics for describing droughts, e.g. the Palmer drought index and standardized precipitation index (Guttman 1998), which may be more or less suitable depending on the category of drought one has in mind.

Analysis of extreme values

RCP4.5 is a transient future scenario with steadily increasing greenhouse gases. As the scenario requires to balance only additional post-2020 radiative greenhouse-gas forcing by stratospheric aerosol, the sulfur injections in the first years after 2020 are very small with the consequence of a small signal-to-noise ratio. Therefore, our analysis concentrates on the last 20 years (2050-2069) of the simulation. From the three ensemble members available from each model for each scenario, we created fictitious time-series of 60 years assumed representative for the two climate states with and without application of SG.

Figure 3.2 shows the difference of the maximum number of consecutive dry days (CDD) calculated for the RCP4.5 and the G3 scenarios. A day is considered 'dry' if precipitation

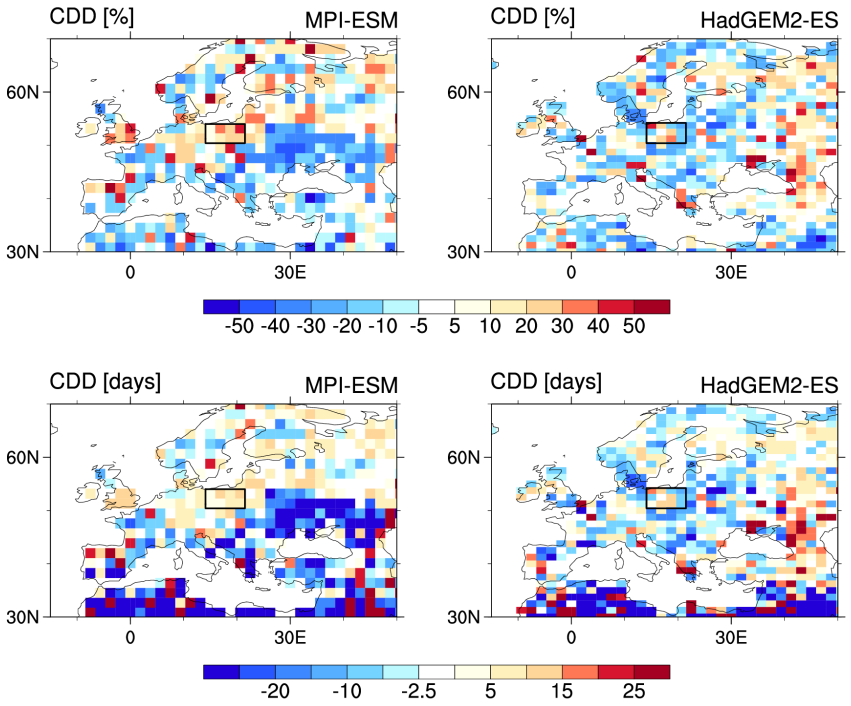

Figure 3.2: Anomalies of maximum numbers of consecutive dry days (CDD) simulated with the climate models MPI-ESM (left) and HadGEM2-ES (right). Plotted are the differences between the G3 and RCP4.5 scenarios in percent (top) and days (bottom). The maximum CDD values have been analyzed from fictitious 60-year time-series representing the climate state between years 2050 and 2069, which was created out of three realizations. The boxes mark the region defined as 'Poland'.

is less than 0.1 mm. The analysis for this paper focuses on the area of 'Poland'. Results may depend on how one exactly defines the spatial boundaries of the area. We chose the spatial boundaries as 14°E to 22°E and 50°N to 54°N in order to include all model grid-boxes which cover the area of Poland. The area average gives an increase of CDD of one week for Poland in the MPI-ESM and an increase of two days for Poland in the HadGEM2-ES. Figure 3.2 shows that the signal of CDD extremes is very noisy and indicates how difficult it may be to unambiguously attribute specific events in a certain region to the applied forcing.

FAR estimation

In order to estimate probability density functions (PDFs) for a CDD event in a given year, we calculated the maximum CDD (XCDD) event of each single year in the time series, both for the factual (G3) and the counterfactual (RCP4.5) climate states. We did so, by first taking the area average and calculating the XCDD on that basis. For each climate model and each scenario, this allows the calculation of a PDF based on these 60 values. In order to obtain the PDFs, we fitted the 60 XCDDs to a Gumbel distribution and a Fréchet distribution. The Gumbel distribution shows better results than the Fréchet distribution and is often used in the literature for the statistics of

droughts (Vicente-Serrano and Beguería-Portugués 2003). We therefore decided to use the PDFs obtained from the Gumbel fits (compare Figure 3.3).

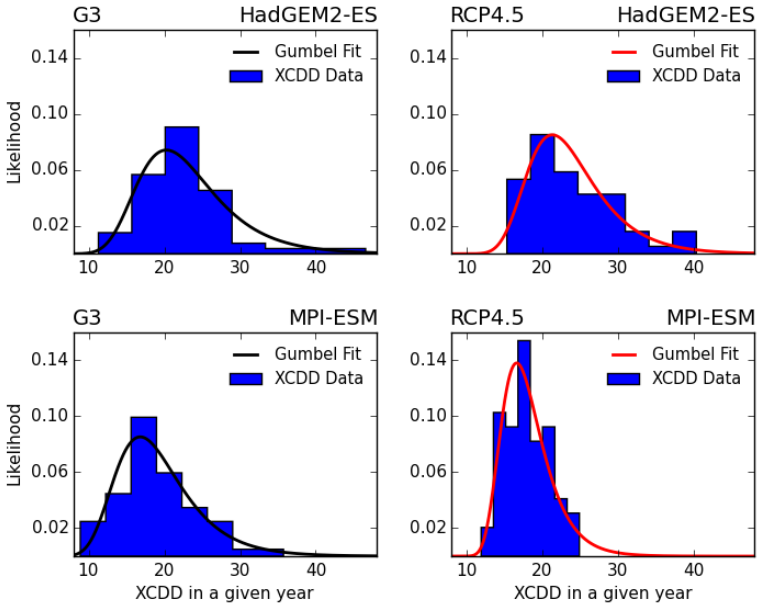

Figure 3.3: Histograms of maximum number of consecutive dry days (XCDD) and the probability density functions of XCDD in a given year resulting from a Gumbel fit. The histograms show the XCDD of each single year from the fictitious 60-year time series for results of the MPI-ESM (top) and the HadGEM2-ES (bottom), both for the G3 scenario (left) and the RCP4.5 scenario (right), respectively.

The FAR for climate model and given event XCDD is:

$$\mathrm{FAR}_i(\mathrm{XCDD}) = 1 - \frac{G_i^{RCP}(\mathrm{XCDD})}{G_i^{G3}(\mathrm{XCDD})}$$

Here, $G_i^{scenario}(\mathrm{XCDD})$ is the likelihood of the event, according to climate model i, in the counterfactual and factual, respectively. Given the event of 40 CDD we assumed, the FAR estimate based on the HadGEM2-ES model yields 0.18, while the FAR estimate based on the MPI-ESM model yields 0.94 (compare Figure 3.4).

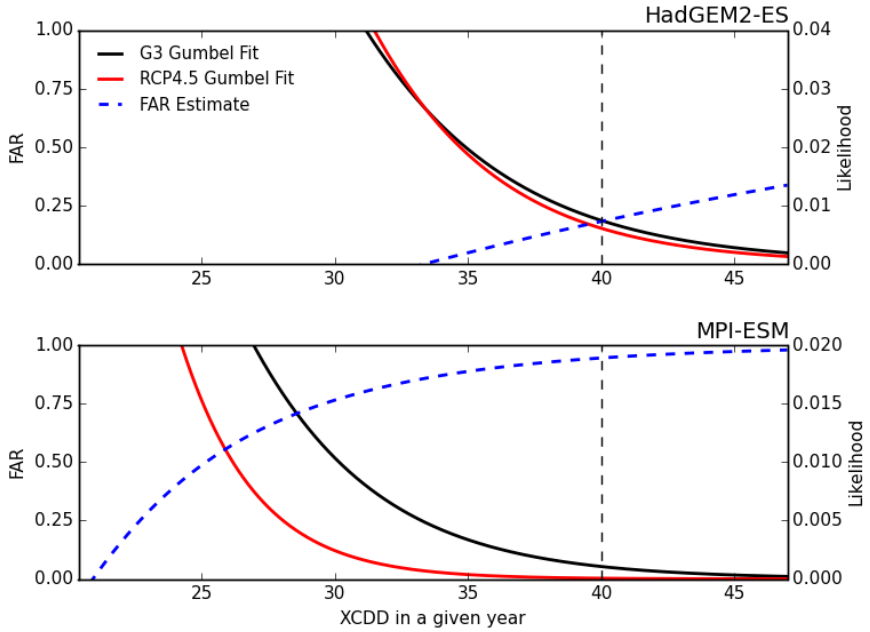

Figure 3.4: Probability density functions (PDFs) of yearly maximum consecutive dry days (XCDD) resulting from a Gumbel fit and FAR estimates for given XCDD. Results of the MPI-ESM (top) and HadGEM2-ES (bottom). FAR estimates based on the MPI-ESM are higher than those based on the HadGEM2-ES. The vertical dotted lines mark results for 40 XCDD, the event chosen in the fictitious court case.

Legal Appendix

The distinction between admissibility of evidence on the one hand and its evaluation on the other is not uniformly incorporated and applied in different domestic and international rules of procedure, and neither is the chronology between admissibility and evaluation. For example, in the U.S., as in other common law systems, the question of how to evaluate the evidence partly arises prior to the actual trial, namely when the judge assesses the admissibility of the evidence. The main reason for this admissibility test is that the jury should not be influenced by evidence which later crystallizes as being invalid, i.e. lacking scientific standards. The admissibility test is thus (at least implicitly) done prior to the hearing by the judge, who follows a set of rules of evidence (i.e. the *Daubert* criteria) that must be fulfilled for a scientific testimony to be brought before the court. In contrast, in civil law systems, admissibility only refers to relevance (van Rhee 2016) and certain prohibitions (Nunner-Krautgasser and Anzensberger 2016), since here the irreversible influence on a jury in trial does not exist. As far as international courts are concerned, in absence of a jury, all evidence has so far usually been considered as being admissible (for the ICJ: Devanay 2016, Tomka and Proulx 2015, Ridell and Plant 2009, for EU Courts: Barbier de la Serre and Sibony 2008), and detailed considerations of the provided evidence are generally missing. In substance, however, the differences between these approaches are less clear as one might expect at first sight: The assessment of what is referred to in common law systems as 'reliability'

within the admissibility test forms, at least to some extent, a 'functional equivalent' to the independent consideration and evaluation of the evidence in civil law procedure (compare Sladic and Uzelac 2016, Tomka and Proulx 2015, Barbier de la Serre and Sibony 2008). The fact that in many instances courts do not provide detailed information in their judgments on how they have evaluated the evidence that they have considered admissible does not mean that such evaluations have not taken place. Rather, with the exception of trials governed by the principle of official investigation, it is inevitable for every judge to at least implicitly assess the admissibility of and evaluate the evidence provided by the parties in order to be able to come to a decision. This two-step exercise is particularly challenging in situations where the factual basis of causation between a certain act and its alleged consequences is subject to scientific uncertainty. Climate intervention scenarios, with regard to which evidence can only be provided by way of climate models, are particularly striking examples of such situations.

Chapter 4

A Model of Solar Geoengineering Liability*

*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, Juan Moreno-Cruz 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 by the German Research Foundation DFG under grant number GO1604-3.

4.1 Introduction

Due to slow progress of climate change mitigation, techniques to increase the earth's reflection of solar radiation, so-called solar geoengineering (SG), have received increasing attention as potential means to reduce climate change risks. SG 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, SG 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 SG 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 SG has incentives to deploy SG 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 SG 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 SG governance have gained wide attention, with a focus on historical precedents, the applicability of existing international law to SG, 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 SG 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.

SG has a key feature which sets it apart from more traditional domains of liability like car accidents or pollution problems. Following Weitzman's terminology, SG 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 metrics, mean temperature and mean precipitation, suggest that moderate amounts of SG would benefit most regions of the world (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015) and that SG would only start to be detrimental to those regions' welfare if provided beyond those moderate amounts.[2] The public good-or-bad characteristic

[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.

[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.

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 SG 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 SG has to be compensated for.

In this paper I focus on the implications of SG'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 SG levels. The agent with the strongest preferences for SG is assumed to be the sole SG 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 SG 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 SG 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 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 SG 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 SG, 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 SG at all. In contrast, the *marginal definition of harm* reflects all increases in the respective victim's damage levels due to increases in SG provision. The former definition is therefore biased towards too high SG provision levels, while the latter is unbiased. This result is of importance for SG 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 SG 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 SG provision level in equilibrium and are biased to too low SG 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 SG 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 SG effects. I do so for two reasons. Firstly, in order to obtain an estimate of how severe the SG 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 SG 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 SG 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 SG governance problem might be substantial: Using an integrated assessment model approach for quantifying the free-driver outcome, Emmerling and Tavoni (2017) find SG 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 SG overprovison in the free-driver outcome in which SG 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 SG. These findings confirm earlier results that regional differences in SG 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 SG provision is according to the strongest preferences for SG.

In the presence of liability regimes, SG 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 4.2 lays out the general SG liability model. Section 4.3 discusses the definitions of harm, while section 4.4 the liability standards. Section 4.5 examines the performance of the various liability regimes. In section 4.6 the SG liability model is implemented into the RCR model. Section 4.7 concludes.

4.2 The SG Liability Model

I model SG 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 SG provision level x and that SG 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 SG 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 SG level x_i. In line with SG'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 SG at the margin and that agent n is the sole SG provider.[3]

There is a liability regime in place which makes the SG 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 SG provider as to pay given her behavior. The SG 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 SG provider has to make liability payments to other parties. Harm determines the amount of compensation a party receives, in case the SG provider has to compensate the party according to the prevailing liability standard.

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

4.2.1 Definition of Harm

There are two salient reference points for measuring SG harm: The first one is the potential victim's condition without any SG 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 SG 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 SG if her damage level is above her damage level in the complete absence of SG. The reference point here is the damage level at zero SG 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 SG 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 SG 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.

4.2.2 Liability Standards

There are two traditional types of liability standards, *strict liability* and *negligence standards*. Under *strict liability*, the SG 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 SG 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 SG 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 SG, the costs of refraining from increasing the SG level are the forgone benefits in form of reduced damages to the SG provider and, potentially, other agents. The costs of preventing harm are weighted against the prevented harm from not increasing the SG level. Since there is SG overprovision in absence of governance, the behavioral standard is conceptualized as a maximum level of SG provision in this model. Liability payments then depend on the SG 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 SG 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 SG, 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 SG 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 SG 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

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

Under *benefit-harm negligence*, the SG 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.

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

Interpretations of negligence focusing on the parties harmed and the SG 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 SG 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 SG provider's damage reduction from increasing SG 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 SG 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.

4.3 Assessment of the Definitions of Harm

For a liability regime to induce the socially optimal SG provision level, it must make the SG provider internalize the negative and the positive externalities on the other $n-1$ agents. Harm determines how large the compensation is which the SG 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 SG 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 SG 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 SG at all. Since her damages, given the provision level x_i, are even smaller than those in the complete absence of SG, a marginal increase in the provision level cannot render her damages larger than those in complete absence of SG. 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 SG, 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: $V(x) = \{i \mid x_i < x\}$.

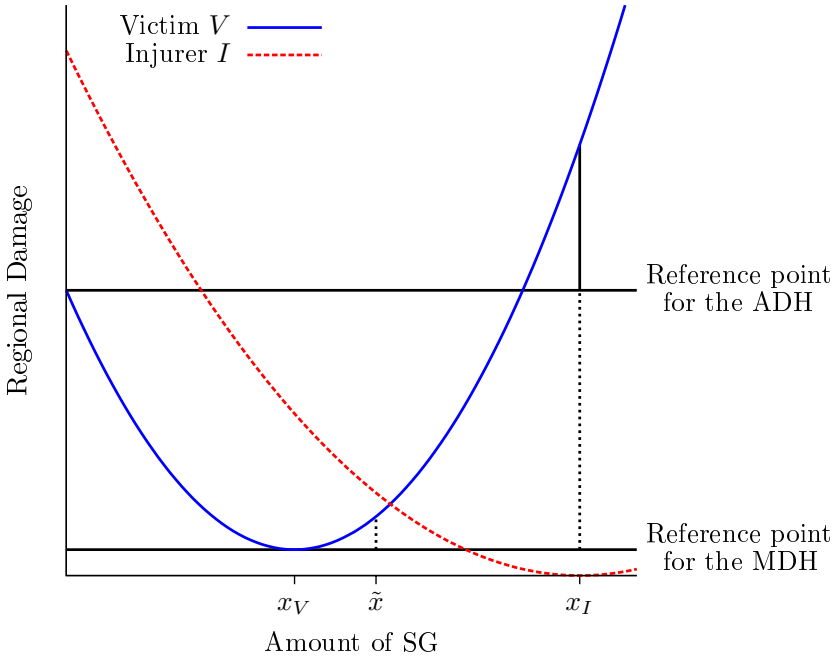

Figure 4.1: Reference points for the *marginal definition of harm* (MDH) and the *absolute definition of harm* (ADH). Victim V's damage function is in blue (solid curve), the injurer I's damage function is in red (dashed curve). The *marginal definition of harm* is represented by the combination of the 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.

I have just argued that for all agents with $x_i > 0$, there is a provision interval $[x_i, x_i^A]$ in which the agent i's marginal damage from SG 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 SG 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 SG provision levels x in equilibrium.

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

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

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 SG 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 SG 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 SG provision x. For all x, the sum of marginal damages for agents*

[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 SG, 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.

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 SG 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 SG 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 SG 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 SG compensation regimes more generally. Any compensation regime (e.g. insurance provided by the SG 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.

[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.

4.4 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 SG 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 SG provider internalize the negative externality to the extent that it is reflected in the definition of harm $h_i(x)$. They can therefore only internalize harm, but not the negative externality as such.

4.4.1 No Liability

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

4.4.2 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 SG 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 SG 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 SG provision levels. A liability regime employing *strict liability* can therefore in general not implement the socially optimal outcome x^*.

4.4.3 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 SG provision

level x_N. The SG 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 SG 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 SG 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 SG 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 SG 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 SG provider faces under *strict liability*, we have $x_{ALN} = x^*_{SL}$. It directly follows that the SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG. 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.

4.5 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 SG 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 SG 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 SG 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 SG provision level in equilibrium.*

2. *In combination with the marginal definition of harm, strict liability and the aggregate harm negligence standard implement too low SG 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 SG 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 SG 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 4.1: 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 SG 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).

4.6 Numerical Implementation

I numerically implement the SG liability model for two main reasons: Firstly, various studies have numerically examined the regional effects of SG (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 SG provision levels, finding that the socially optimal provision of SG 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 SG 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 SG 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 SG 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 SG 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 SG, 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 SG 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.

4.6.1 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 SG ($\Delta\mathcal{M}^i_{SG}(x)$). Both of these deviations are normalized by preindustrial regional interannual variability $\sigma^i_{M,pre}$. The SG provision level's impact is assumed to be linear[7]:

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

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

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

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

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

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

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

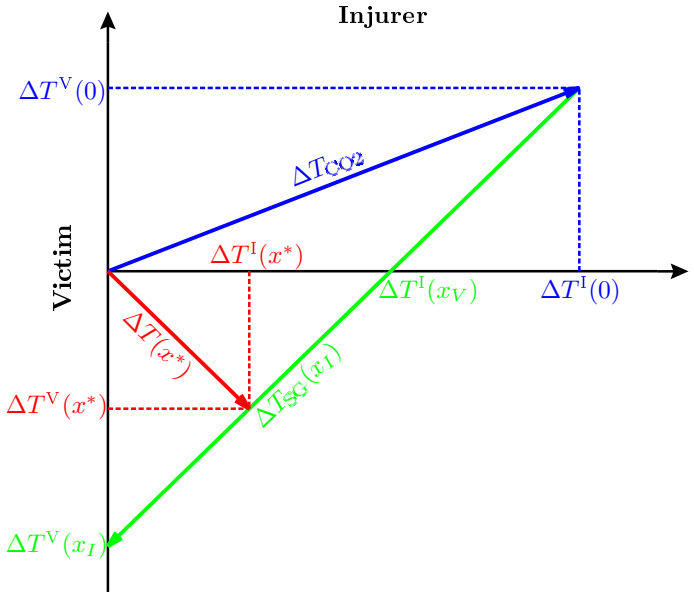

Figure 4.2: Residual Climate Response model. The horizontal axis shows changes in temperature for the injurer I, the vertical axis shows changes in temperature for the victim V. The blue CO_2 vector represents the temperature change due to climate change. The green SG vector represents the temperature change due to SG 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 SG. 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 SG 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 SG 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 SG) 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 SG is deployed on top of the elevated CO2 climate in order to restore the preindustrial global mean temperature. This implies that a SG provision level of $x = 1$ in the model corresponds to restoring preindustrial global mean temperature. M^i_{pre}, M^i_{CO2}, M^i_{SG} 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_{SG} and $\sigma^i_{M,pre}$. I used the average across climate models, since the free-driver scenario reflects the strongest preferences for SG and is therefore prone to outliers.

4.6.2 Results from the RCR Model

I report the equilibrium SG 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 SG level of 0.99 for the temperature metric and of 0.80 for the precipitation metric. These optimal SG levels entail residual damages of 0.2% and 5.1% of unmitigated climate change damages. Pareto-optimal SG 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]

[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_{SG} and $\sigma^i_{M,pre}$. For the median climate model, Kravitz et al. (2014) report a Pareto-optimal SG level of 0.91 for the temperature metric and of zero for the precipitation metric.

Concurrent research comes to the conclusion that the SG governance problem might be substantial: Using an integrated assessment model approach for quantifying the free-driver outcome, Emmerling and Tavoni (2017) find SG 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 SG 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 SG 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 SG is devastating if mean precipitation is the relevant metric.

These results reflect the findings from earlier studies that regional differences in SG 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 SG is even in the free-driver outcome very effective at reducing unmitigated climate change damages. However, SG 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 SG determines the level of SG. In this scenario the larger regional SG disparities for precipitation have a much stronger effect on welfare, compared to scenarios in which the socially optimal or the Pareto-optimal SG 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 99.9% of the possible reduction in residual damages.

Panel A: Temperature Metric

	Marginal Definition of Harm		Absolute Definition of Harm	
	SG Level	Damages	SG 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	
	SG Level	Damages	SG 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 4.2: SG levels are given as a fraction of the SG level which restores global mean temperature to preindustrial. Residual damages in percent of unmitigated climate change damages. No SG therefore corresponds to residual damages of 100%.

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 SG 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 SG 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 SG provision levels. Since the *marginal definition of harm* has no bias, the standards' bias drives the SG equilibrium provision level away from the social optimum. However, the *absolute definition of harm* is biased towards too high SG 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 performance 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.

4.7 Conclusion

SG is a set of techniques which has received increasing attention as potential means to offset climate change. Governance is a key issue for SG, 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 SG has incentives to deploy SG 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 SG liability regimes provide. Furthermore, I implemented the model numerically in order to obtain a first-order estimate of the extent of the SG governance problem and the capability of liability regimes to solve it in a simplified setting.

SG 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 SG 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 SG impacts have to be compensated for. The second one concerns the liability standards, which are about the circumstances under which harm from SG has to be compensated for. The liability model of SG 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 SG provider from SG'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* standard, 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 SG 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 SG 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 SG, 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 SG governance problem might be substantial: Using a more sophisticated integrated assessment model approach for quantifying the free-driver outcome, Emmerling and Tavoni (2017) find SG 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 SG overprovison in the free-driver outcome and SG 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 SG 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 SG 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 SG 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 SG then determines the SG provision level.

Liability regimes lead to a welfare-enhancing implementation of SG for the precipitation 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 SG liability regime.

This paper has focused on the specific incentives liability regimes provide in light of SG's good-or-bad characteristic. I therefore abstracted from various other SG aspects, which are important, but do not lie at the heart of the SG-specific incentives liability regimes provide. These aspects include uncertainty about SG impacts, other potential SG 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 SG (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 SG

liability regime. Currently, there is no dedicated SG 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 SG liability regime, customary international law may provide a legal basis for SG 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 SG could decide to form a coalition (Ricke et al. 2013) in order to provide SG jointly, while sharing the expected liability payments, thereby partly internalizing the positive externalities from SG 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 SG impacts in general and as a result also for the assessment of the performance of SG 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 SG 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 SG 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 SG 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 SG 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 SG 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 SG provision level of 5, which is below the socially optimal level $x^* = 6$. □

Chapter 5

Diverging Regional Climate Preferences and the Assessment of Solar Geoengineering*

*I want to thank Ben Kravitz for providing me with the climate model data used in this study. Furthermore, I want to thank Florian Diekert and Daniel Heyen for helpful comments. I gratefully acknowledge funding by the German Research Foundation DFG under grant number GO1604-3.

5.1 Introduction

Techniques to increase the earth's reflection of solar radiation, so-called solar geoengineering (SG), have received increasing attention as potential means to reduce climate change risks. While SG may well be able to compensate for increased temperatures on a global level, SG has regionally heterogeneous impacts (Lunt et al. 2008, Robock et al. 2008, Irvine et al. 2010, Ricke et al. 2010). These heterogeneous impacts are widely regarded as a source of substantial SG governance problems and as a potential source of conflict (Robock 2008, Shepherd 2009, Weitzman 2015, Heyen 2016, Pasztor 2017).

Regional differences in SG impacts have been the focus of a number of studies (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015, Pfrommer 2018). Generally, the results indicate that regional temperature disparities from SG may not be as severe as previously thought. However, this literature employs the assumption that regional climate preferences derive from a common baseline climate, e.g. 1990 or preindustrial climate conditions. This 'change-is-bad-assumption' has been criticized (Heyen et al. 2015) and research has provided empirical evidence that certain temperature levels may generally be more conducive to economic activity than others, irrespective of historic regional climate conditions. Schlenker and Roberts (2009) provide evidence for a non-linear, inversely U-shaped, relationship between crop yield and daily temperature. Graff and Neidell (2014) find a similar relationship between labor supply and daily temperature. Burke et al. (2015) estimate a globally generalizable, inversely U-shaped relationship between overall economic productivity and regional mean temperature. Additionally, there are region-specific reasons why a deviation from regional baseline climate conditions may be beneficial (Heyen et al. 2015), e.g. that more natural resources in high northern latitudes may become available as temperatures rise.

The main purpose of the paper is to gain a conceptual understanding of the impact that regionally diverging temperature preferences have on SG outcomes. To this aim, I extend the Residual Climate Response (RCR) model developed by Moreno-Cruz et al. (2012). The RCR model is the main framework for examining SG impacts on the regional level in the literature. The extension allows for regions to have temperature preferences diverging from the baseline climate and builds on an illustrative example by Heyen et al. (2015). They demonstrate by that example that regional temperature disparities may be substantially higher when regions have diverging temperature preferences. The extended model calculates the welfare maximizing SG level depending on the diverging temperature preferences that regions have. This optimal SG level determines each region's temperature. The difference between regional temperature and a region's desired temperature level is the region's residual temperature. Regional residual temperature determines regional damages and therefore regional climate welfare.

The key theoretical insight of the extended model is that the impact of diverging preferences can be split into two components. The first component changes the optimal SG

level, but does not affect the set of residual temperatures. This component therefore does not change regional disagreement over SG. The second component leaves the optimal SG level unaffected, but changes the set of residual temperatures. This decomposition helps in understanding how specific diverging preferences affect globally optimal SG and regional disagreement over SG.

At least three different aspects of SG performance are of interest in relation to diverging preferences. The first aspect is the relative effectiveness of SG in reducing damages. Relative effectiveness measures in percent the share of regional damages that optimal SG can compensate for. There are two potential damage baselines for measuring relative effectiveness. The first baseline is total damages (arising from the combination of CO_2 driven temperature changes[1] and diverging preferences), giving rise to the *total damage reduction metric M1* – the metric considered by Heyen et al. (2015). The second baseline is purely CO_2 driven damages, giving rise to the *CO_2 damage reduction metric M2*. While the *total damage reduction metric* measures how well optimal SG compensates for damages originating from the difference between regional temperatures in a high CO_2 climate and regionally preferred temperatures, the *CO_2 damage reduction metric* measures how well optimal SG compensates for damages caused by CO_2 induced temperature changes only. The two damage baselines – and the two metrics – are identical in the absence of diverging preferences.

Relative effectiveness measures SG performance with respect to one fixed set of diverging preferences. The other two aspects measure SG performance relative to SG performance in the absence of diverging preferences (the 'baseline scenario'). The second aspect is the change of the minimum climate damages, or equivalently, of the maximum climate welfare that SG can implement relative to the baseline scenario. It is captured by comparing residual damages (the sum of regional damages given optimal SG) for a given set of diverging preferences to residual damages in the baseline scenario. This second aspect gives rise to the *minimum climate damage metric M3*. By help of the *minimum climate damage metric*, one can compare the maximum climate welfare that SG can implement across different sets of diverging preferences. The last aspect is the change in the gross value of SG relative to the baseline scenario. It is captured by comparing the maximum reduction of damages that SG can achieve for a given set of diverging preferences to the maximum reduction in the baseline scenario and gives rise to the *gross value metric M4*. By help of the *gross value metric*, one can compare the gross value of SG across different sets of diverging preferences.

A set of diverging preferences can be expressed as a combination of a 'scenario' and of a 'preference strength'. A scenario describes regions' diverging temperatures preferences relative to each other, the preference strength determines the magnitude of the desired

[1] In the entire paper, CO_2 is intended to mean "CO_2 equivalent", i.e. CO_2 represents all greenhouse gases.

temperature deviations from the baseline climate. Scenarios in which high-latitude regions generally prefer higher temperatures than in the baseline climate and low-latitude regions generally prefer lower temperatures are of specific interest, since there is empirical evidence suggesting that such a pattern of diverging preferences generally holds (Burke et al. 2015). In order to develop a basic understanding of how such scenarios affect SG outcomes, I numerically implement two concrete realizations of such scenarios with data from the Geoengineering Model Intercomparision Project (Kravitz et al. 2011). For each scenario, I implement different strengths of diverging preferences.

The numerical implementation yields two main results. Firstly, it shows that the performance of optimal SG relative to the baseline scenario depends on the aspect of SG performance one is interested in. The presence of diverging preferences may change SG performance in either direction and the direction generally depends on which of the three aspects of SG performance is considered. The latter implies that the aspects and metrics developed are in fact independent and convey complementary information about how SG performance changes in the presence of diverging preferences.

Secondly, the numerical results suggest two welfare implications. The first is that optimal climate welfare (climate welfare given optimal SG) may increase relative to the absence of diverging preferences. However, such a positive change in optimal climate welfare only occurs when diverging temperature preferences are small in magnitude compared to average CO_2 induced warming. The second implication is that optimal climate welfare is often higher than climate welfare in the baseline climate, i.e. in the climate before CO_2 driven temperature changes set in. These implications demonstrate that diverging climate preferences do not necessarily lower SG performance, at least when optimal SG can be implemented. I argue that these welfare implications are likely to emerge more generally in scenarios in which high-latitude regions prefer higher temperatures and low-latitude regions prefer lower temperatures than in the baseline climate.

I proceed as follows. In section 5.2, I introduce the RCR model. Section 5.3 extends the RCR model. Construction and implementation of the scenarios follows in section 5.4. Section 5.5 concludes.

5.2 The Residual Climate Response Model

The Residual Climate Response (RCR) model (Moreno-Cruz et al. 2012) is a simple framework for evaluating regional effects of solar geoengineering (SG). The main purpose of the RCR model is to examine how well SG can compensate for CO_2 induced climate change on the regional level. Due to its simplicity, the model is intuitively accessible (compare Figure 5.1). Its strength lies in identifying first-order effects in the assessment of regional SG impacts and in providing a framework for conceptually thinking about regional disparities in SG outcomes.

The RCR model operates on a fixed number of regions. Regional climate damages derive from the difference in mean temperature between actual climate conditions and climate conditions in some baseline climate. The CO_2 vector T_{CO2} consists of the regional temperature increases due to CO_2 emissions relative to the baseline climate.[2] One unit of SG is defined as the amount of SG restoring global mean temperature to the level of the baseline climate. The SG vector T_{SG} consists of the regional mean temperature changes due to one unit of SG relative to the high CO_2 climate. Since the CO_2 and SG vectors are not congruent (i.e. the relative effects of CO_2 and SG on temperatures differ across regions), SG can only imperfectly compensate for CO_2 induced temperature increases on a regional level. The model assumes linearity in the effect of SG on regional temperature.[3] The SG level x is defined as a fraction of one unit of SG. Due to the linearity assumption, the residual vector T_{RES} of regional temperatures, given the SG level x, is

$$T_{RES}(x) = T_{CO2} + x \cdot T_{SG}.$$

The global welfare measure is residual damages, i.e. the sum of regional damages, and corresponds to the squared length of the residual vector T_{RES} for a given level of SG:

$$D(x) = \sum_i d^i(x) = |T_{RES}(x)|^2 \quad \text{with} \quad d^i(x) = (T^i_{CO2} + x \cdot T^i_{SG})^2.$$

The optimal SG level, minimizing residual damages, is

$$x^* = -\frac{\sum_i T^i_{SG} \cdot T^i_{CO2}}{|T_{SG}|^2} = -\frac{T_{SG} \bullet T_{CO2}}{|T_{SG}|^2} = \frac{|T_{CO2}| \cdot \cos(\varphi)}{|T_{SG}|},$$

where φ is the angle between the CO_2 and the SG vector and $(\bullet)$ the dot product between vectors.[4] The smaller φ, the more similar the CO_2 and the SG vector and the better SG can compensate for the temperature changes on a regional level caused by CO_2: In case the vectors are parallel ($\varphi = 0°$), compensation is perfect. Consequently, we then have $x^* = 1$ and $D(x^*) = 0$. In case the vectors are perpendicular ($\varphi = 90°$), compensation is not possible at all. Consequently, we then have $x^* = 0$ and $D(x^*) = |T_{CO2}|^2$. If not stated otherwise, the terms residual damages and residual vector from now on refer to the respective outcomes given the optimal SG level.

[2] While other metrics, like regional precipitation, are generally relevant and employed as well, I focus on the metric of regional temperature. Baseline climates employed in the literature are preindustrial and 1990 climate conditions (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015, Pfrommer 2018).

[3] Evidence for the reasonableness of this assumption is provided, among others, by Moreno-Cruz et al. (2012) and Kravitz et al. (2014).

[4] Technically, φ as depicted in Figure 5.1, is not the angle between the CO_2 and the SG vector, but between the CO_2 vector and the negative of the SG vector. φ is to be understood as defined by Figure 5.1. The reason for doing so is that I do not want to deviate from the definition in Moreno-Cruz et al. (2012). The only implication is that $\cos(\varphi)$ picks up a minus sign:

$$\cos(\varphi) = -\frac{T_{CO2} \bullet T_{SG}}{|T_{CO2}| \cdot |T_{SG}|}.$$

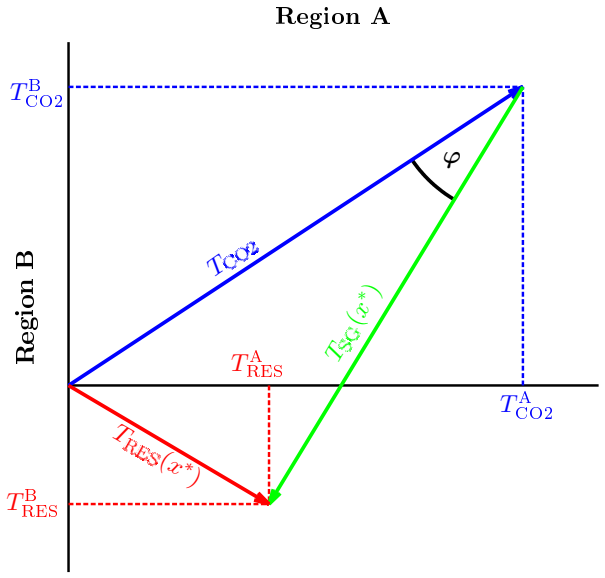

Figure 5.1: Two-region representation of the Residual Climate Response model. The horizontal axis shows changes in temperature for region A, the vertical axis shows changes in temperature for region B. The blue CO_2 vector represents regional temperature changes due to CO_2. The green SG vector represents regional temperature changes due to optimal SG. The red residual vector points to regional temperatures under optimal SG. Since regional damages are quadratic, the squared length of the residual vector is proportional to residual damages. The angel φ between CO_2 and SG vector represents the extent to which SG can compensate for regional temperature changes due to CO_2 and determines the socially optimal SG level as well as the metric M.

The metric M used for assessing SG is the relative effectiveness of optimal SG in compensating for CO_2 induced damages on the regional level.[5] Due to the linearity assumption, M does not depend on the length of the CO_2 vector, but only on φ. Since the dot product between the residual vector $T_{\mathrm{RES}}(x^*)$ and the SG vector T_{SG} is zero, the two are perpendicular (see Figure 5.1) and M evaluates to

$$\frac{D(0) - D(x^*)}{D(0)} = 1 - \frac{|T_{\mathrm{RES}}(x^*)|^2}{|T_{\mathrm{CO2}}|^2} = 1 - \sin^2(\varphi).$$

5.3 Extension of the Residual Climate Response Model

The following extension of the RCR model includes the possibility of diverging regional preferences from the baseline climate. The vector T_{DIV} of diverging temperature preferences consists of each region's preferred temperature relative to the baseline climate.

[5]The damages different regions experience may also be weighted. Such weights may, for example, reflect differences in population or economic output. However, the analysis does not fundamentally change in the presence of welfare weights and the possibility of welfare weights is therefore not further pursued in this paper.

From a societal perspective, the aim in deploying SG in the baseline model is to compensate for the CO_2 induced regional temperature deviation from the baseline climate as represented by the CO_2 vector. In the face of diverging preferences, the aim changes to compensating for the differences between regional temperatures in the high CO_2 climate and regions' preferred temperature levels. The residual vector in the extended model is therefore

$$T_{\mathrm{RES}}(x, T_{\mathrm{DIV}}) = (T_{\mathrm{CO2}} - T_{\mathrm{DIV}}) + x \cdot T_{\mathrm{SG}}$$

and the welfare goal is minimizing

$$D(x, T_{\mathrm{DIV}}) = |T_{\mathrm{RES}}(x, T_{\mathrm{DIV}})|^2.$$

When referring to the residual vector and the residual damages in the absence of diverging preferences, I will leave out the second argument.

Denote the angle between the SG vector and the diverging preferences vector by ϑ (see Figure 5.2). Furthermore, denote optimal SG in the baseline model as x^* and in the extended model as x^*_{DIV}. The optimal amount of SG in the presence of diverging preferences is then

$$x^*_{\mathrm{DIV}} = -\frac{T_{\mathrm{SG}} \bullet (T_{\mathrm{CO2}} - T_{\mathrm{DIV}})}{|T_{\mathrm{SG}}|^2} = x^* + \frac{|T_{\mathrm{DIV}}| \cdot \cos(\vartheta)}{|T_{\mathrm{SG}}|}.$$

The vector of diverging preferences can be decomposed into a component parallel and a component perpendicular to the SG vector. In case ϑ is larger than 90°, the parallel component points in the opposite direction of the SG vector. The diverging preferences vector then (partially) substitutes for the cooling the SG vector provides and less SG is optimal than in the absence of diverging preferences. The contrary is true in case ϑ is smaller than 90° and the parallel component points in the same direction as the SG vector. The decomposition of the diverging preferences vector into the parallel and perpendicular components is

$$T_{\mathrm{DIV}} = T^{\perp}_{\mathrm{DIV}} + T^{\parallel}_{\mathrm{DIV}},$$

$$\text{with} \quad T^{\parallel}_{\mathrm{DIV}} = \frac{T_{\mathrm{SG}} \bullet T_{\mathrm{DIV}}}{|T_{\mathrm{SG}}|^2} \cdot T_{\mathrm{SG}} \quad \text{and} \quad T^{\perp}_{\mathrm{DIV}} = T_{\mathrm{DIV}} - T^{\parallel}_{\mathrm{DIV}}.$$

The component parallel to the SG vector changes regional temperature preferences in the same proportions as SG changes regional temperatures. Increasing or decreasing SG relative to optimal SG in the baseline model can therefore perfectly compensate for the parallel component. This implies that the parallel component does not change the residual vector. Therefore, optimal SG changes, but regional damages and hence residual damages do not change relative to the absence of diverging preferences. In particular, when the diverging preferences vector is (anti)parallel to the SG vector, regions disagree about SG in exact the same way as in the baseline model.

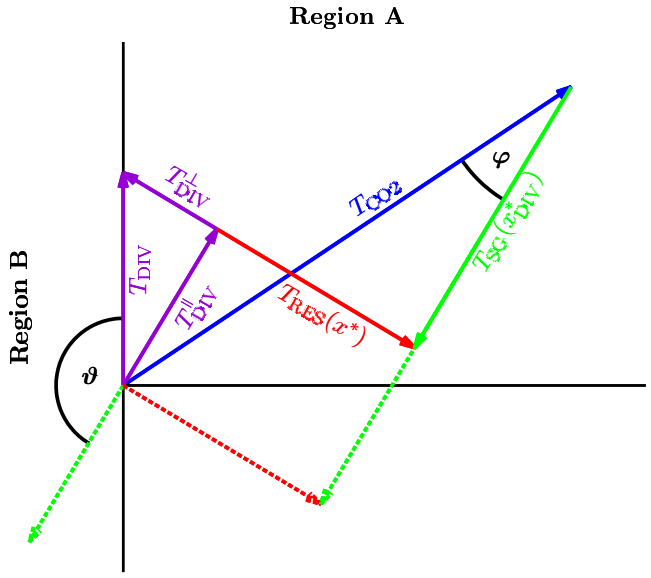

Figure 5.2: Extension of the Residual Climate Response model. The combination of the solid green vector and the dashed green vector represents regional temperature changes due to optimal SG in the baseline model, the dashed red vector represents the regional residual temperatures in the baseline model. The solid green vector represents regional temperature changes due to optimal SG in the presence of the purple diverging preferences vector. The latter can be decomposed into a component parallel and a component perpendicular to the SG vector. The parallel component changes the optimal SG level. The perpendicular component changes the residual vector. Since the angle ϑ between the diverging preferences vector and the SG vector is larger than 90°, optimal SG decreases relative to the baseline model. The residual vector in the presence of diverging preferences is the red baseline residual vector less the purple perpendicular component. Since the perpendicular component points into the opposite direction of the baseline residual vector, the residual vector in the presence of diverging preferences is longer than the baseline residual vector, i.e. residual damages increase.

In contrast, changing the SG level cannot compensate for the perpendicular component at all and the perpendicular component does therefore not affect optimal SG. It must then completely be taken up by the residual vector, thereby changing how regions disagree about SG relative to the baseline model. In particular, when the diverging preferences vector is perpendicular to the SG vector, optimal SG does not change compared to the baseline model.

The residual vector in the presence of diverging preferences may be longer or shorter than the residual vector in the baseline model – reflecting an increase or decrease in residual damages – depending on its perpendicular component's length and direction relative to the baseline residual vector. The angle between the perpendicular component and the baseline residual vector is from now on denoted by γ. The tighter γ and the shorter the perpendicular component relative to the baseline residual vector, the more likely it is that residual damages are smaller in the presence of diverging preferences. When γ is smaller than 90°, the vectors largely point into the same direction and partially cancel out. When γ is larger than 90°, both vectors point in opposing directions and the

perpendicular component necessarily increases residual damages. In Figure 5.2, we have $\gamma = 180°$, hence residual damages in the example depicted are larger in the presence of diverging preferences. When the perpendicular component is at least twice at large as the baseline residual vector, the perpendicular component necessarily overcompensates for the baseline residual vector and residual damages increase even if the vectors are parallel ($\gamma = 0°$).

Theoretical Result. *Let $T_{\mathrm{DIV}} = T^{\perp}_{DIV} + T^{\parallel}_{DIV}$ be the vector of diverging preferences and its decomposition into the perpendicular and the parallel component.*

1. *The optimal SG level is*

$$x^*_{\mathrm{DIV}} = x^* + \frac{|T_{DIV}| \cdot \cos(\vartheta)}{|T_{SG}|} = x^* + \frac{|T^{\parallel}_{\mathrm{DIV}}|}{|T_{\mathrm{SG}}|} \cdot \mathrm{sign}(\cos(\vartheta)).$$

2. *The residual vector is*

$$T_{\mathrm{RES}}(x^*_{\mathrm{DIV}}, T_{\mathrm{DIV}}) = T_{\mathrm{RES}}(x^*) - T^{\perp}_{\mathrm{DIV}}$$

3. *Residual damages in presence of diverging preferences are smaller than those in absence of diverging preferences if and only if*

$$\cos(\gamma) > \frac{1}{2} \frac{|T^{\perp}_{\mathrm{DIV}}|}{|T_{\mathrm{RES}}(x^*)|},$$

where γ is the angle between the perpendicular component $T^{\perp}_{\mathrm{DIV}}$ and the residual vector $T_{\mathrm{RES}}(x^*)$ in the baseline model.

Assessment Metrics

In the baseline model, the performance of optimal SG is captured by a single aspect, the relative effectiveness in damage compensation. In contrast, at least three different aspects of SG performance are of interest in relation to diverging preferences. Analogously to the baseline RCR model, the first aspect is the relative effectiveness of optimal SG in reducing damages for a given set of diverging preferences. Measuring relative effectiveness necessarily involves the definition of a damage baseline. In the baseline model, the obvious choice for the damage baseline is damages in absence of SG or, equivalently, damages caused by CO_2. However, damages in absence of SG and damages caused by CO_2 are different damage baselines in the presence of diverging preferences. These two damage baselines for measuring relative effectiveness give rise to two different metrics in the extended model. The first is the *total damage reduction metric M1*, using total damages (arising from the combination of CO_2 induced temperature changes and diverging preferences) as damage baseline. The *total damage reduction metric* is the one

used by Heyen et al. (2015) for illustrating the potential impact of diverging preferences on SG performance. The second damage baseline is the *CO_2 damage reduction metric M2*, using damages purely caused by CO_2 induced temperature changes (i.e. total damages less damages from the mere presence of diverging preferences) as damage baseline. The *total damage reduction metric* measures how well optimal SG compensates for damages arising from the differences between regional temperatures in the high CO_2 climate and regionally preferred temperatures. Since the *total damage reduction metric* measures SG effectiveness relative to the regional optima, its maximum value is 100%. The *CO_2 damage reduction metric* measures how well optimal SG compensates for damages arising from differences between regional temperatures in the high CO_2 climate and regional temperatures in the baseline climate. The relative effectiveness in compensating for damages purely caused by CO_2 induced temperature changes can therefore be higher than 100%. In those cases, optimal SG compensates for more damages than CO_2 causes, meaning that residual damages are lower than damages in the baseline climate. Analytical definitions of the metrics can be found in Table 5.1.

At least two different aspects of SG performance concerning the change in performance across sets of diverging preferences are of interest. The first aspect is the minimum climate damages, or, equivalently, the maximum climate welfare that SG can implement. This aspect is captured by residual damages, i.e. damages given optimal SG. In order to compare SG performance across different sets of diverging preferences, I normalize residual damages for a given set of preferences to baseline residual damages. This gives rise to the *minimum climate damage metric M3*, measuring how maximum climate welfare changes with the presence of diverging preferences. The second aspect is the gross value of SG, which is captured by the maximum damage reduction SG can achieve. For comparing the gross value across sets of diverging preferences, I normalize to the gross value of SG in the baseline model. This gives rise to the *gross value metric M4*, which measures how the gross value of SG changes with the presence of diverging preferences. Note that in the baseline model, or for any other fixed set of climate preferences, maximum climate welfare and the gross value of SG are redundant. However, they are independent across climate preferences, since total damages vary across different climate preferences.

Metrics in the Presence of Diverging Preferences

M1	*M2*	*M3*	*M4*
$\frac{\Delta D(T_{\mathrm{DIV}})}{D(0,T_{\mathrm{DIV}})}$	$\frac{\Delta D(T_{\mathrm{DIV}})}{D(0,T_{\mathrm{DIV}})-\lvert T_{\mathrm{DIV}}\rvert^2}$	$\frac{D(x^*_{\mathrm{DIV}},T_{\mathrm{DIV}})}{D(x^*)}$	$\frac{\Delta D(T_{\mathrm{DIV}})}{D(0)-D(x^*)}$

Table 5.1: $\Delta D(T_{\mathrm{DIV}})$ expresses the difference in damages between no SG and optimal SG for a given vector of diverging preferences: $\Delta D(T_{\mathrm{DIV}}) = D(0,T_{\mathrm{DIV}}) - D(x^*_{\mathrm{DIV}},T_{\mathrm{DIV}})$. The metrics are the *total damage reduction metric M1*, the *CO_2 damage reduction metric M2*, the *minimum climate damage metric M3* and the *gross value metric M4*.

The metric in the baseline model is determined by φ, whereas the metrics in the extended model additionally depend on ϑ and the relative length of the diverging preferences vector and the CO_2 vector.

5.4 Exemplary Implementation of Two Scenarios

The exemplary implementation of the extended RCR model delivers a closer examination of how a specific class of diverging preferences scenarios affects SG outcomes. I follow the relevant literature (Moreno-Cruz et al. 2012, Kravitz et al. 2014, Yu et al. 2015, Pfrommer 2018) in defining the regions of the RCR model according to Giorgi and Francisco (2000). I use data from the thirteen climate models which participated in the G1 experiment as defined in the Geoengineering Model Intercomparision Project (Kravitz et al. 2011) for the implementation. Each model performed a run simulating preindustrial climate conditions, a run simulating a climate with four times elevated CO_2 concentration levels, and a run in which SG is used to restore the global mean temperature to the preindustrial level. The regional mean temperatures, averaged over the thirteen models, of the three runs are used to calculate the CO_2 vector and the SG vector.

The SG literature usually normalizes regional temperatures to each region's preindustrial interannual variability when assessing SG (Heyen et al. 2015). However, potential sources for diverging regional preferences derive from phenomena related to absolute temperatures (Burke et al. 2015, Heyen et al. 2015). Therefore, I use absolute temperatures for the implementation. For comparison, I provide the results when using normalized temperatures in the appendix. The results I obtain for the baseline RCR model when using normalized temperatures are in line with the literature (Moreno-Cruz et al. 2012, Yu et al. 2015). The main difference in results between using normalized and absolute temperatures in the presence of diverging preferences is that optimal SG levels are substantially higher when using normalized temperatures. The difference is explained by low-latitude regions having on average a much smaller interannual variability in mean temperature than high-latitude regions. The upshot is that absolute temperatures should be used for the extended RCR model, unless one has evidence that the impacts one is interested in are better captured by normalized temperatures.

The Scenarios

I focus on scenarios which are based on the premise that, as a rule, high-latitude regions prefer a warmer climate and low-latitude regions prefer a cooler climate relative to the baseline climate. This general scenario structure is both plausible and supported by empirical evidence (Burke et al. 2015). I group regions into high-latitude, high-mid-latitude, low-mid-latitude and low-latitude bins (see Table 5.2). In scenario A, regions

pertaining to the high-latitude group desire a warmer climate than in the baseline climate and regions in the low-latitude group desire a colder one. All mid-latitude regions are content with the baseline climate. The difference in scenario B is that regions in the high-mid-latitude group also desire a warmer climate. Implementing two scenarios reduces the risk of picking up results which are idiosyncratic to a specific scenario and essentially serves as a robustness check. Regions' diverging preferences in the scenarios are always equally strong or not present: Each region's desired temperature is one unit above temperatures in the regional baseline climate, one unit below temperatures in the regional baseline climate or equals temperatures in the regional baseline climate.

Grouping of Regions and Scenarios

	Each Region's			Scenario	
	Avg. Lat.	Opt. SG Level	Res. Temp.	A	B
Greenland	67.5N	1.09	0.44	1.0	1.0
Alaska	66.0N	1.08	0.50	1.0	1.0
North. Europe	61.5N	1.09	0.33	1.0	1.0
North Asia	60.0N	1.12	0.66	1.0	1.0
WN America	45.0N	1.04	0.06	0.0	1.0
CN America	40.0N	1.07	0.23	0.0	1.0
Central Asia	40.0N	1.02	-0.08	0.0	1.0
Tibet	40.0N	0.99	-0.29	0.0	1.0
Mediterranean	39.0N	1.04	0.05	0.0	0.0
SS America	38.0S	1.02	-0.04	0.0	0.0
EN America	37.5N	1.03	0.01	0.0	0.0
South. Australia	37.5S	1.01	-0.09	0.0	0.0
East Asia	35.0N	1.03	0.02	0.0	0.0
Sahara	24.0N	0.99	-0.21	-1.0	-1.0
Central America	20.0N	0.97	-0.29	-1.0	-1.0
South Asia	17.5N	0.96	-0.35	-1.0	-1.0
Southern Africa	11.5S	0.97	-0.27	-1.0	-1.0
North. Australia	9.5S	0.95	-0.38	-1.0	-1.0
Southeast Asia	4.5N	0.93	-0.35	-1.0	-1.0
Amazon Basin	4.0S	0.99	-0.20	-1.0	-1.0
Western Africa	3.0N	0.97	-0.31	-1.0	-1.0
Eastern Africa	3.0N	0.97	-0.30	-1.0	-1.0

Table 5.2: Regions are ordered according to their average latitude and grouped into 'high-latitude', 'high-mid-latitude', 'low-mid-latitude' and 'low-latitude' bins. The left columns states each region's average latitude, each region's preferred SG level and each region's residual temperature under optimal SG. The right columns states regions' diverging preferences, relative to each other, from the baseline climate in the scenarios A and B.

The last part of the theoretical result implies non-linear effects in the relative length of the diverging preference vector on the change in SG outcomes. In order to capture these effects, I consider different strengths of diverging preferences (i.e. different sizes of what constitutes "one unit") for each scenario. In other words, a scenario defines regions' diverging preferences relative to each other (the direction of the diverging preference vector) and the preference strength defines the magnitude of the desired temperature deviations from the baseline climate (the vector's relative length). Jointly, the two determine a set of diverging preferences. I define preference strength in percent of the average regional temperature change caused by CO_2. Due to the RCR model's linearity, this relative definition leads to well-defined results. As an example, say CO_2 driven average regional warming is 4°C. A strength of 50% in diverging preferences then means that regions desiring a temperature change relative to the regional baseline climate do so by 2°C.

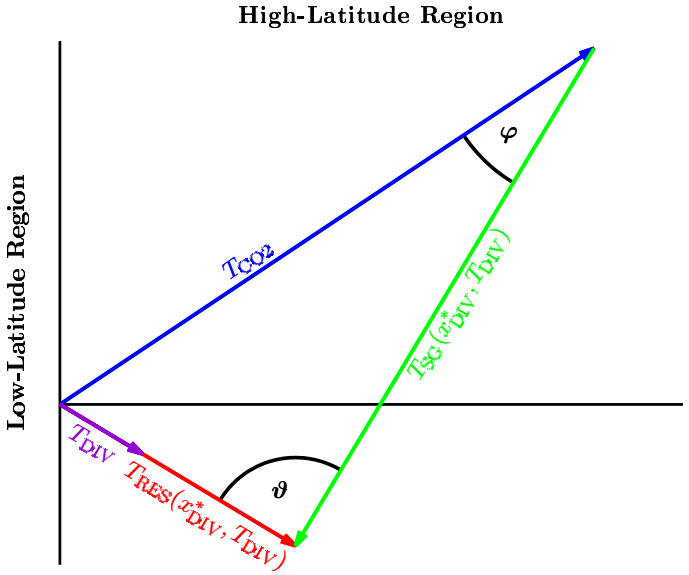

Figure 5.3: Two-region representation of the basic scenario structure. Given baseline climate preferences, optimal SG undercompensates for temperature in the high-latitude region, but overcompensates for temperature in the low-latitude region. The high-latitude region desires a warmer climate and the low-latitude region a colder climate compared to the baseline climate. In this specific example, the SG vector and the diverging preferences vector are perpendicular ($\vartheta = 90°$), hence the optimal SG level does not change relative to the baseline model. Due to the structure of the scenario, the perpendicular component of the diverging preferences vector (note that there is no parallel component in this specific example) and the baseline residual vector (which here is the combination of the violet diverging preferences vector and the red residual vector in the presence of diverging preferences) point into the same direction. The angle γ between the two vectors is 0°, since they are parallel, and the former compensates in part for the latter. In a two-region example, γ can only attain 0°and 180°, which is not the case when a higher number of regions is involved.

SG has a differential impact across different latitudes: Due to geophysical reasons, optimal SG overcompensates for warming induced by CO_2 in regions of low latitude and undercompensates in regions of high latitude (e.g. Ban-Weiss and Caldeira 2010, compare Table 5.2).[6] Based on this differential SG impact and the structure of the scenarios, the last part of the theoretical result suggests a certain pattern for the relationship between residual damages and preference strength. Consider a two-region example, with one high-latitude and one low-latitude region (compare Figure 5.3). When the high-latitude region prefers a warmer climate and the low-latitude region a colder one relative to the baseline climate, the diverging preference vector and the baseline residual vector point in the same general direction. The perpendicular component and the baseline residual vector are then parallel (since the example is two-dimensional) and the former compensates in part of the other, i.e. residual damages decrease due to the diverging preferences. However, this is only the case when the perpendicular component is not too long, in which case it overcompensates for the baseline residual vector. Residual damages are then larger than in absence of diverging preferences.

The theoretical result mirrors the result from the two-region example. It states that, for residual damages to decrease due to diverging preferences, the angle γ between the perpendicular component of the diverging preferences vector and the baseline residual vector has to be between $-90°$ and $90°$, while at the same time the former vector has to be short enough relative to the latter. Taking into account that in the two-region example γ is $0°$, the theoretical result exactly predicts the outcomes in two-region example. The theoretical result tells us that the pattern will qualitatively hold beyond two-region examples if only γ is between $-90°$ and $90°$. It is self-evident that for two-region examples γ is always in that range (even exactly $0°$). Whether this is the case in a specific multi-region example ultimately depends on the details of the scenario, such as which regions are considered hiigh-latitude, which are considered low-latitude, how strong individual regions' preferences are and the differential impacts of SG along other dimensions than latitude. However, for the type of scenario under consideration, it seems likely that γ is in that range – at least for many scenarios. The implementation delivers the actual γ for two specific such scenarios and delivers the range of preference strengths for which residual damages are smaller than in the baseline scenario, i.e. for which the *minimum climate damage metric* evaluates to below 100%. Lastly, the implementation may reveal whether there are similar patterns concerning other aspects of SG performance.

Results of the Implementation

I report the angles φ, ϑ and γ and the optimal SG level in absence and presence of diverging preferences. I report the results for the four metrics for several preference

[6] Note that potential non-uniform SG schemes optimized for latitude-dependent albedo reduction may ameliorate latitudinal differences in outcomes (Ban-Weiss and Caldeira 2010, MacMartin et al. 2013).

strengths between 0% and 100%. The angle φ and optimal SG in absence of diverging preferences x^* are independent of the scenario. Given x^*, ϑ and the strength of diverging preferences determine the SG level in presence of diverging preferences. Optimal SG is linear in the strength of preferences and it suffices to only explicitly report optimal SG for for one particular strength. The optimal SG level x^*_{DIV} reported refers to a preference strength of 100%.[7]

Optimal SG Levels and Angles between Vectors

Baseline Model			Scenario A			Scenario B		
x^*	φ	M	x^*_{DIV}	ϑ	γ	x^*_{DIV}	ϑ	γ
1.03	2.9°	99.7	1.16	80.1°	22.6°	0.96	94.5°	38.2°

Table 5.3: The left columns state the optimal SG level x^* in the baseline model, the angle φ between the CO_2 and the SG vector and the metric M (in percent) from the baseline model. The middle columns and the right columns state, for the scenarios A and B, respectively, the optimal SG level x^*_{DIV} when preference strength is 100%, the angle ϑ between the SG vector and the diverging preferences vector and the angle γ between the perpendicular component of the diverging preferences vector and the baseline residual vector.

The angle φ between the CO_2 vector and the SG vector is small (2.9°). In the absence of diverging preferences, optimal SG is therefore close to one and the relative effectiveness of SG in reducing CO_2 induced damages (metric M) is close to 100%. In both scenarios, the angle ϑ between the SG vector and the diverging preferences vector is close to 90° (scenario A: $\vartheta = 80.1°$, scenario B: $\vartheta = 94.5°$). Consequently, optimal SG differs only moderately between the absence and presence of diverging preferences. Even for very strong diverging preferences (100%), optimal SG is only 0.13 higher than in the baseline model (+12.6%) in scenario A and only 0.07 lower than in the baseline model (−6.8%) in scenario B. Optimal SG is larger in scenario A than in scenario B, corresponding to the on net stronger preferences for higher temperatures in scenario B.

The *total damage reduction metric M1* and the *minimum climate damage metric M3* exhibit a similar pattern for both scenarios. According to both metrics, the performance of optimal SG increases for weak diverging preferences relative to the baseline scenario and then decreases again for strong diverging preferences. Minimum residual damages are attained for a preference strength of 6.4% in scenario A and for a preference strength of 4.7% in scenario B. Residual damages start to exceed baseline residual damages for a preference strength of 12.8% in scenario A and for 9.4% in scenario B. The corresponding preferences strengths for the *total damage reduction metric* are very similar (differences $< 0.2\%$). Residual damages become very large relative to baseline residual damages for very strong diverging preferences, up to a factor of 184 and of 255 higher in scenario A and scenario B, respectively. Relative effectiveness in compensating for total damages

[7] Optimal SG for other strengths of preferences can be obtained by linear interpolation. For example, optimal SG for a strength of 50% is $x^* + 0.5 \cdot (x^*_{\text{DIV}} - x^*)$.

is substantially lower than in the baseline scenario for very strong diverging preferences, down to 72.9% and 56.9% in scenario A and scenario B, respectively. However, for both scenarios, SG can compensate for at least 86.1% of total damages when the preference strength is no higher than 50% and for at least 96.9% when the preference strength is no higher than 25%.

Results for the Four Metrics

Preference	Scenario A				Scenario B			
Strength	*M1*	*M2*	*M3*	*M4*	*M1*	*M2*	*M3*	*M4*
0%	99.7	99.7	100.0	100.0	99.7	99.7	100.0	100.0
5%	99.9	100.0	18.7	101.2	99.9	100.0	38.5	99.3
10%	99.8	100.4	42.3	102.5	99.7	100.4	117.3	98.6
15%	99.5	100.7	170.7	103.8	99.1	100.8	336.5	98.0
25%	98.2	101.4	742.1	106.5	96.9	101.5	1196	96.6
50%	91.6	103.2	4005	113.1	86.1	103.6	5801	93.3
75%	82.4	105.1	9890	120.0	71.5	106.0	13915	90.1
100%	72.8	106.9	18396	127.1	56.9	108.7	25539	86.9

Table 5.4: The metrics are the *total damage reduction metric M1*, the *CO_2 damage reduction metric M2*, the *minimum climate damage metric M3* and the *gross value metric M4*. *M1* and *M2* are in percent of their respective damage baseline. *M3* and *M4* are in percent relative to the respective outcome in the baseline scenario. Results are given in incremental steps in the strength of diverging preferences. For example, a strength of 50% means that one unit of diverging preferences corresponds to half the average regional temperature change from climate change. If regional mean temperature rises by 4°C in absence of SG, a strength of 50% corresponds to low-latitude regions having a diverging temperature preference of −2°C and high-latitude regions having a diverging temperature preference of +2°C in scenario A.

The changes in the *CO_2 damage reduction metric M2* and the *gross value metric M4* are monotone in the strength of diverging preferences. The relationship between the two metrics and preference strength is almost linear. Relative effectiveness in compensating for damages caused by CO_2 increases in both scenarios. For very strong diverging preferences, it is 106.9% in scenario A and 108.7% in scenario B. Relative effectiveness in compensating for damages caused by CO_2 is at least 100% in both scenarios when the strength of diverging preferences is at least 5%. The gross value of SG increases in scenario A, but decreases in scenario B. For very strong diverging preferences, the gross value of SG is 27.1% higher than the gross value of SG in absence of diverging preferences in the former and 13.1% lower than the gross value of SG in absence of diverging preferences in the latter.

Explanation and Interpretation of the Results

The *minimum climate damage metric M3* measures the maximum climate welfare SG can implement in a scenario relative to maximum climate welfare in the baseline scenario and is determined by the ratio of the respective residual damages. Since $\gamma = 22.6°$ in scenario A and $\gamma = 38.2°$ in scenario B, the *minimum climate damage metric* follows the pattern presumed in advance: The metric is smaller than 100% for weak diverging preferences and larger than 100% for moderate to strong diverging preferences in both scenarios. In both scenarios, γ is well within the range of $(-90°, 90°)$. One can therefore expect γ to be within this range for many scenarios following the basic premise, implying that one can expect the pattern to hold for many such scenarios. The values $\cos(\gamma)$ attains in the scenarios are close to one (0.92 in scenario A, 0.78 in scenario B). The preference strengths for which the *minimum climate damage metric* is smaller than 100% in scenario A are therefore close to the maximum of such strengths possible for any scenario. For strong diverging preferences, the perpendicular component of the diverging preferences vector is, irrespective of its direction, much larger than the baseline residual vector. The residual vector is then almost identical to the perpendicular component. Hence, the *minimum climate damage metric* becomes very large for strong diverging preferences, irrespective of the scenario.

The *total damage reduction metric M1* measures the relative effectiveness of optimal SG in compensating for total damages. It can be expressed as

$$1 - \frac{|T_{\mathrm{RES}}(x^*) - T_{\mathrm{DIV}}^{\perp}|^2}{|T_{\mathrm{CO2}} - T_{\mathrm{DIV}}|^2}.$$

In both scenarios, the relative effectiveness in compensating for total damages is higher than in the baseline scenario for weak diverging preferences and lower for moderate to strong ones. Since the CO_2 vector is comparatively large, the relative change in total damages (the denominator in the expression) from absence to presence of diverging preferences is comparatively small, while the relative change in residual damages (the nominator in the expression) is comparatively high. Therefore, residual damages govern the qualitative behavior of the *total damage reduction metric* at least for weak and moderate preferences, resulting in the observed pattern in both scenarios. This intuition is valid except for the rather unrealistic case that the perpendicular component is trivial.

The *CO_2 damage reduction metric M2* measures the relative effectiveness of optimal SG in compensating for damages caused by CO_2. It can be expressed as

$$1 + \frac{|T_{\mathrm{DIV}}|^2 - |T_{\mathrm{RES}}(x^*_{\mathrm{DIV}}, T_{\mathrm{DIV}})|^2}{|T_{\mathrm{CO2}} - T_{\mathrm{DIV}}|^2 - |T_{\mathrm{DIV}}|^2}.$$

From this characterization, it is evident that the *CO_2 damage reduction metric* evaluates to 100% when the diverging preferences vector and the residual vector are of the same

length, i.e. when SG restores the damage level of the baseline climate, thereby exactly compensating for pure CO_2 damages. In both scenarios, the *CO_2 damage reduction metric* exceeds 100% for even weak preferences. When $\gamma \in (-90°, 90°)$, the residual vector becomes shorter for weak preferences. The closer γ is to 0°, the faster the diverging preference vector becomes longer than the residual vector with increasing preference strength.[8] Since the baseline residual vector is very short and γ is in both scenarios of small to moderate size, the *CO_2 damage reduction metric* becomes larger than 100% for even weak preferences in both scenarios.

The *gross value metric M4* measures the gross value of SG in a scenario relative to its gross value in the baseline scenario. It can be expressed as

$$\frac{|T_{\mathrm{CO2}} - T_{\mathrm{DIV}}|^2 - |T_{\mathrm{RES}}(x^*) - T_{\mathrm{DIV}}^{\perp}|^2}{|T_{\mathrm{CO2}}|^2 - |T_{\mathrm{RES}}(x^*)|^2}.$$

The *gross value metric* exceeds 100% for all preference strengths in scenario A and is below 100% for all preference strengths in scenario B, coinciding with optimal SG increasing in scenario A and decreasing in scenario B relative to optimal SG in the baseline model. As a heuristic, the parallel component of the diverging preferences vector and the CO_2 vector can be considered as (anti)parallel, since the angle φ between the SG and the CO_2 vector is very small. Intuitively, the parallel component then increases total damages when it causes optimal SG to rise and reduces total damages when it causes optimal SG to fall, while not affecting the residual vector – in other words, the parallel component increases the gross value of SG when it causes optimal SG to rise and decreases the gross value of SG when it causes optimal SG to fall. However, this reasoning leaves out the influence of the perpendicular component on the change in gross value of SG. This means that, while a more positive change in optimal SG increases the gross value of SG *ceteris paribus*, a positive (negative) change in optimal SG relative to optimal SG in the baseline scenario does not necessarily correspond to a gross value of SG higher (lower) than in the baseline scenario.

The numerical implementation and its discussion yield two main results. Firstly, the performance of SG in presence of diverging preferences may increase or decrease relative to the performance in absence of diverging preferences. Which one is the case depends on the set of diverging preferences, the aspect of SG performance one is interested in and, for the aspect of relative effectiveness of optimal SG, on the damage baseline against which effectiveness is measured. In particular, this underpins the importance of specifying the aspect of SG performance and, if applicable, the damage baseline, when assessing SG in presence of diverging preferences: In the two scenarios, the *absolute*

[8] It can be easily be shown that the diverging preference vector always becomes longer than the residual vector for strong enough preferences when $\gamma \in (-90°, 90°)$: It holds that

$$|T_{\mathrm{DIV}}|^2 - |T_{\mathrm{RES}}(x^*_{\mathrm{DIV}}, T_{\mathrm{DIV}})|^2 = |T_{\mathrm{DIV}}^{\parallel}|^2 - |T_{\mathrm{RES}}(x^*)|^2 + T_{\mathrm{RES}}(x^*) \cdot T_{\mathrm{DIV}}^{\perp}.$$

The last term on the right-hand side is positive if $\gamma \in (-90°, 90°)$, from which the statement follows.

damage reduction metric indicates a reduction in the relative effectiveness of SG for at least moderately strong diverging preferences, while the *CO_2 damage reduction metric* indicates an increase in the relative effectiveness of SG for all preference strengths. Depending on scenario and preference strength, the *minimum climate damage metric* and the *gross value metric* both indicate an increase in SG performance, both indicate a decrease in SG performance or one indicates an increase while the other indicates a decrease. Therefore, how the maximum climate welfare that SG can implement and the gross value of SG change in the presence of diverging preferences are in fact independent.

Secondly, there are patterns in the change of SG performance which hold for both scenarios. The first pattern is that the relative effectiveness of SG in reducing absolute damages and the maximum climate welfare that SG can implement increase for weak diverging preferences and decrease for stronger diverging preferences. In particular, this means that the maximum climate welfare that SG can implement increases relative to the maximum climate welfare in the absence of diverging preferences, albeit only when regions' diverging temperatures preferences are small in magnitude compared to average CO_2 induced warming. The second pattern is that the relative effectiveness of SG in reducing damages caused by CO_2 is above 100% for all but very weak preferences. The implication here is that optimal SG can implement a climate welfare higher than the climate welfare in the baseline climate, i.e. in the climate before CO_2 driven temperature changes set in. These two patterns concerning maximum climate welfare demonstrate that diverging climate preferences do not necessarily lower SG performance, at least when optimal SG is feasible. I have argued that these pattern are very likely (in case of the *minimum climate damage metric* even sure) to emerge when the angle γ is small. Furthermore, scenarios in which high-latitude regions prefer a colder climate and low-latitude regions prefer a warmer one, will often lead to a small γ, due to the differential impact of SG across latitudes. Is is therefore likely that also other scenarios of that type will often lead to the patterns observed in the implementation.

5.5 Conclusion

Solar geoengineering (SG) has the potential to compensate for increased temperatures from climate change on a global level. However, SG has heterogeneous impacts at the regional level. Until now, studies examining these regional differences have employed the assumption that regions' temperature preferences correspond to a common baseline climate, e.g. to preindustrial or 1990 climate conditions. This assumption has been criticized (Heyen et al. 2015) and conflicts with empirical evidence supporting globally generalizable relationships between economic productivity and absolute temperature levels (Burke et al. 2015).

In this paper, I extended the Residual Climate Response (RCR) model (Moreno-Cruz et al. 2012) for assessing regional SG differences by formally introducing the possibility of regional temperature preferences diverging from the baseline climate, building on an illustrative example by Heyen et al. (2015). In the key theoretical result, I showed that the impact of these diverging preferences can be split into two components. The first component changes the optimal SG level, but does neither change optimal climate welfare nor affect regional disagreement over SG. The second component leaves the optimal SG level unaffected, but changes regional disagreement over SG and optimal climate welfare. This decomposition helps in understanding how specific diverging preferences affect globally optimal SG and the disagreement over SG by different regions. I introduced metrics for measuring three independent aspects of SG performance. The first aspect is the relative effectiveness of SG in reducing damages, the second aspect is the change of optimal climate welfare relative to the absence of diverging preferences and the third aspect is the change in the gross value of SG relative to the absence of diverging preferences.

I numerically implemented the extended RCR model, focusing on scenarios in which high-latitude regions prefer a warmer climate and low-latitude regions prefer a cooler climate relative to the baseline climate – a scenario structure which is both plausible and supported by empirical evidence (Burke et al. 2015). The results of the implementation suggest two welfare implications. The first is that optimal climate welfare may increase relative to the absence of diverging preferences. However, such a positive change in optimal climate welfare only occurs when diverging temperature preferences are small in magnitude compared to average climate change induced warming. The second implication is that optimal climate welfare is often higher than climate welfare in the baseline climate, i.e. in the climate before greenhouse gas driven temperature changes set in. These implications demonstrate that diverging climate preferences do not necessarily lower SG performance, at least when optimal SG can be implemented. I argue that these welfare implications are likely to emerge more generally in scenarios in which high-latitude regions prefer higher temperatures and low-latitude regions prefer lower temperatures than in the baseline climate.

One should keep in mind that both the baseline and the extended RCR model are deliberatively simple in nature. They derive their usefulness from conceptual understanding and from identifying first-order effects. Additionally, the results I obtained for the type of scenario I considered are temperature specific and do not necessarily hold for other climate variables, for which diverging preferences may be relevant as well, like precipitation. Several lines of research should be pursued in the future for further increasing the understanding of the relationship between diverging climate preferences and the assessment of SG. Firstly, I concentrated on the outcomes for globally optimal SG. Examining the potential impact on Pareto optimal SG levels and on SG levels in the free-driver outcome (Weitzman 2015), as well as examining the respective welfare implications, will

lead to a more complete picture regarding the relationship between diverging climate preferences and the assessment of SG. Secondly, investigating the potential impact of diverging preferences on coalition formation, in particular coalitions based on similar latitudes, may lead to further insights into the strategic dimensions of SG. Lastly, a further conceptual extension seems desirable: Regional damages may be conceptualized as being a combination of damages deriving from the variability-adjusted temperature difference to regional baseline climate temperatures and damages deriving from the absolute temperature difference to some absolute temperature preference. A generalization incorporating both, however, necessarily opens up the question of the relative weight of both types of damages.

Appendix

Implementation of the Extended RCR Model Using Normalized Temperatures

Here, I provide the results for the extended RCR model when using temperatures normalized to regional interannual variability. The main difference in results between using normalized and absolute temperatures in the presence of diverging preferences is that optimal SG levels are substantially higher when using normalized temperatures (compare Table 5.3 and Table 5.5). The differences in optimal SG between using normalized and absolute temperatures is explained by low-latitude regions having on average a much smaller interannual variability in mean temperature than high-latitude regions. Using normalized temperatures, diverging temperature preferences of the same absolute magnitude then translate into a much larger magnitudes in units of interannual variability for low-latitude regions than for high-latitude regions, while this effect is not present when using absolute temperatures.

Optimal SG Levels and Angles between Vectors

Baseline Model			Scenario A			Scenario B		
x^*	φ	M	x^*_{DIV}	ϑ	γ	x^*_{DIV}	ϑ	γ
0.99	2.5°	99.8	1.68	59.1°	23.2°	1.60	72.3°	32.7°

Table 5.5: Results for the implementation of the extended RCR Model using normalized temperatures. Corresponds to Table 5.3. The left columns state the optimal SG level x^* in the baseline model, the angle φ between the CO_2 and the SG vector and the metric M (in percent) from the baseline model. The middle columns and the right columns state for the scenarios A and B, respectively, the optimal SG level x^*_{DIV} when preference strength is 100%, the angle ϑ between the SG vector and the diverging preferences vector and the angle γ between the perpendicular component of the diverging preferences vector and the baseline residual vector.

The results for the four metrics when using normalized temperatures (compare Table 5.6) are qualitatively the same as when using absolute temperatures. The same general patterns can be observed when using normalized temperatures as when using absolute temperatures, regarding the inversely U-shaped relationship between SG performance as measured by metric *M1*, as well as metric *M3* and preference strength, regarding *M2* exceeding 100% for all but very weak diverging preferences and regarding the relationship between the change in optimal SG and metric *M4*.

Results for the Four Metrics

Preference	Scenario A				Scenario B			
Strength	*M1*	*M2*	*M3*	*M4*	*M1*	*M2*	*M3*	*M4*
0%	99.8	99.8	100.0	100.0	99.8	99.8	100.0	100.0
5%	99.9	100.1	17.3	107.1	99.9	100.1	30.0	106.2
10%	99.9	100.7	58.4	114.5	99.7	100.6	134.7	112.7
15%	99.6	101.4	223.4	122.1	99.3	101.3	414.3	119.3
25%	98.7	103.1	924.9	138.0	97.9	102.7	1497	133.1
50%	95.3	108.7	4845	182.0	92.7	107.6	7265	170.8
75%	91.4	115.5	11863	232.2	86.9	113.4	17403	213.3
100%	87.7	122.9	21976	288.5	81.6	119.9	31911	260.5

Table 5.6: Results for the implementation of the extended RCR Model using normalized temperatures. Corresponds to Table 5.4. The metrics are the *total damage reduction metric M1*, the *CO_2 damage reduction metric M2*, the *minimum climate damage metric M3* and the *gross value metric M4*. *M1* and *M2* are in percent of their respective damage baseline. *M3* and *M4* are in percent relative to the respective outcome in the baseline scenario. Results are given in incremental steps in the strength of diverging preferences. For example, a strength of 50% means that one unit of diverging preferences corresponds to half the average regional temperature change from climate change. If regional mean temperature rises by 4°C in absence of SG, a strength of 50% corresponds to low-latitude regions having a diverging temperature preference of −2°C and high-latitude regions having a diverging temperature preference of +2°C in scenario A.

Proofs

Proof of the Theoretical Result.

1. Since the residual damages are convex in the SG level, the first equation from the first part follows from computing the first order conditions of the residual damages with respect to the SG level. The second equation follows because $T_{\mathrm{DIV}}^{\parallel}$ is the projection of T_{DIV} onto T_{SG}.

2. Plugging in the optimal SG level into

$$T_{\mathrm{RES}}(x, T_{\mathrm{DIV}}) = (T_{\mathrm{CO2}} - T_{\mathrm{DIV}}) + x \cdot T_{\mathrm{SG}}$$

yields

$$\begin{aligned} T_{\mathrm{RES}}(x^*_{\mathrm{DIV}}, T_{\mathrm{DIV}}) &= (T_{\mathrm{CO2}} - (T^{\perp}_{\mathrm{DIV}} + T^{\parallel}_{\mathrm{DIV}})) + (x^* + \frac{|T_{\mathrm{DIV}}| \cdot \cos(\vartheta)}{|T_{\mathrm{SG}}|}) \cdot T_{\mathrm{SG}} \\ &= (T_{\mathrm{CO2}} + x^* \cdot T_{\mathrm{SG}}) - T^{\perp}_{\mathrm{DIV}} - T^{\parallel}_{\mathrm{DIV}} + \frac{|T_{\mathrm{DIV}}| \cdot \cos(\vartheta)}{|T_{\mathrm{SG}}|} \cdot T_{\mathrm{SG}} \end{aligned}$$

Because of

$$\cos(\vartheta) = \frac{T_{\mathrm{SG}} \bullet T_{\mathrm{DIV}}}{|T_{\mathrm{SG}}| \cdot |T_{\mathrm{DIV}}|},$$

it holds that

$$T^{\parallel}_{\mathrm{DIV}} = \frac{|T_{\mathrm{DIV}}| \cdot \cos(\vartheta)}{|T_{\mathrm{SG}}|} \cdot T_{\mathrm{SG}}$$

and the second part follows.

3. The third part follows from

$$\begin{aligned} &|T_{\mathrm{RES}}(x^*) - T^{\perp}_{\mathrm{DIV}}|^2 < |T_{\mathrm{RES}}(x^*)|^2 \Leftrightarrow \\ &|T_{\mathrm{RES}}(x^*)|^2 - 2 \cdot T_{\mathrm{RES}}(x^*) \bullet T^{\perp}_{\mathrm{DIV}} + |T_{\mathrm{RES}}(x^*)|^2 < |T_{\mathrm{RES}}(x^*)|^2 \Leftrightarrow \\ &\tfrac{1}{2} \cdot |T^{\perp}_{\mathrm{DIV}}|^2 < T_{\mathrm{RES}}(x^*) \bullet T^{\perp}_{\mathrm{DIV}} = |T_{\mathrm{RES}}(x^*)| \cdot |T^{\perp}_{\mathrm{DIV}}| \cdot \cos(\gamma). \end{aligned}$$

□

References

Allen, M. (2003). "Liability for climate change". In: *Nature* 421.6926, p. 891.

Allen, M., Pall, P., Stone, D., Stott, P., Frame, D., Min, S.-K., Nozawa, T., and Yukimoto, S. (2007). "Scientific challenges in the attribution of harm to human influence on climate". In: *University of Pennsylvania Law Review*, pp. 1353–1400.

Amir, R. (1996). "Cournot oligopoly and the theory of supermodular games". In: *Games and Economic Behavior* 15.2, pp. 132–148.

Aswathy, V., Boucher, O., Quaas, M., Niemeier, U., Muri, H., Mülmenstädt, J., and Quaas, J. (2015). "Climate extremes in multi-model simulations of stratospheric aerosol and marine cloud brightening climate engineering". In: *Atmospheric Chemistry and Physics* 15.16, pp. 9593–9610.

Ban-Weiss, G. A. and Caldeira, K. (2010). "Geoengineering as an optimization problem". In: *Environmental Research Letters* 5.3, p. 034009.

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

Baumann, F. and Friehe, T. (2016). "Learning-by-doing in torts: Liability and information about accident technology". In: *Economics Letters* 138, pp. 1–4.

Ben-Shahar, O. (1998). "Should Products Liability Be Based on Hindsight?" In: *Journal of Law, Economics, and Organization* 14.2, pp. 325–357.

Brown, J. P. (1973). "Toward an economic theory of liability". In: *The Journal of Legal Studies* 2.2, pp. 323–349.

Burk, D. L. and Boczar, B. A. (1993). "Biotechnology and tort liability: A strategic industry at risk". In: *U. pitt. L. rev.* 55, p. 791.

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.

Byrne, R. E. (1973). "Strict Liability and the Scientifically Unknowable Risk". In: *Marq. L. Rev.* 57, p. 660.

Calabresi, G. (1968). "Transaction Costs, Resource Allocation and Liability Rules–A Comment". In: *The Journal of Law and Economics* 11.1, pp. 67–73.

Calabresi, G. (1970). *The Costs of Accidents*. Yale University Press, New Haven, Conn.

Calabresi, G. and Melamed, A. D. (1972). "Property rules, liability rules, and inalienability: one view of the cathedral". In: *Harvard law review*, pp. 1089–1128.

Carmines, E. G. and Zeller, R. A. (1979). *Reliability and validity assessment*. Vol. 17. Sage publications.

Christidis, N., Stott, P. A., Scaife, A. A., Arribas, A., Jones, G. S., Copsey, D., Knight, J. R., and Tennant, W. J. (2013). "A new HadGEM3-A-based system for attribution of weather-and climate-related extreme events". In: *Journal of Climate* 26.9, pp. 2756–2783.

Coase, R. H. (1960). "The Problem of Social Cost". In: *The Journal of Law and Economics* 3, pp. 1–44.

Collins, W., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C., Joshi, M., Liddicoat, S., et al. (2011). "Development and evaluation of an Earth-System model-HadGEM2". In: *Geoscientific Model Development* 4.4, p. 1051.

Connolly, P. R. (1965). "The Liability of a Manufacturer for Unknowable Hazards Inherent in His Product". In: *Ins. Counsel J.* 32, p. 303.

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

Dari-Mattiacci, G. (2009). "Negative liability". In: *The Journal of Legal Studies* 38.1, pp. 21–59.

d'Aspremont, C. and Jacquemin, A. (1988). "Cooperative and noncooperative R & D in duopoly with spillovers". In: *The American Economic Review* 78.5, pp. 1133–1137.

De Volder, M. F., Tawfick, S. H., Baughman, R. H., and Hart, A. J. (2013). "Carbon nanotubes: present and future commercial applications". In: *Science* 339.6119, pp. 535–539.

Devaney, J. G. (2016). *Fact-finding before the International Court of Justice*. Cambridge University Press.

Diffenbaugh, N. S., Swain, D. L., and Touma, D. (2015). "Anthropogenic warming has increased drought risk in California". In: *Proceedings of the National Academy of Sciences* 112.13, pp. 3931–3936.

Dole, R., Hoerling, M., Perlwitz, J., Eischeid, J., Pegion, P., Zhang, T., Quan, X.-W., Xu, T., and Murray, D. (2011). "Was there a basis for anticipating the 2010 Russian heat wave?" In: *Geophysical Research Letters* 38.6.

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.

Endres, A. and Bertram, R. (2006). "The development of care technology under liability law". In: *International Review of Law and Economics* 26.4, pp. 503–518.

Endres, A. and Friehe, T. (2011). "Incentives to diffuse advanced abatement technology under environmental liability law". In: *Journal of Environmental Economics and Management* 62.1, pp. 30–40.

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

Finkelstein, A. (2004). "Static and dynamic effects of health policy: Evidence from the vaccine industry". In: *The Quarterly Journal of Economics* 119.2, pp. 527–564.

Fondazione Rosselli (2004). "Analysis of the economic impact of the development risk clause as provided by Directive 85/374/EEC on liability for defective products". In: *Report for the European Commission*.

Frich, P., Alexander, L. V., Della-Marta, P., Gleason, B., Haylock, M., Tank, A. K., and Peterson, T. (2002). "Observed coherent changes in climatic extremes during the second half of the twentieth century". In: *Climate research* 19.3, pp. 193–212.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., et al. (2013). "Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5". In: *Journal of Advances in Modeling Earth Systems* 5.3, pp. 572–597.

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.

Goeschl, T., Heyen, D., and Moreno-Cruz, J. (2013). "The intergenerational transfer of solar radiation management capabilities and atmospheric carbon stocks". In: *Environmental and resource economics* 56.1, pp. 85–104.

Goldman, A. I. (1986). *Epistemology and cognition*. Harvard University Press.

Graff Zivin, J. and Neidell, M. (2014). "Temperature and the allocation of time: Implications for climate change". In: *Journal of Labor Economics* 32.1, pp. 1–26.

Guttman, N. B. (1998). "Comparing the Palmer drought index and the standardized precipitation index". In: *JAWRA Journal of the American Water Resources Association* 34.1, pp. 113–121.

Haack, S. (2008). "What's Wrong with Litigation-Driven Science-An Essay in Legal Epistemology". In: *Seton Hall L. Rev.* 38, p. 1053.

Haack, S. (2010). "Federal Philosophy of science: A deconstruction-and a reconstruction". In: *NYUJL & Liberty* 5, p. 394.

Hannart, A., Pearl, J., Otto, F., Naveau, P., and Ghil, M. (2016). "Causal counterfactual theory for the attribution of weather and climate-related events". In: *Bulletin of the American Meteorological Society* 97.1, pp. 99–110.

Hauser, M., Gudmundsson, L., Orth, R., Jézéquel, A., Haustein, K., Vautard, R., Oldenborgh, G. J. van, Wilcox, L., and Seneviratne, S. I. (2017). "Methods and model dependency of extreme event attribution: the 2015 European drought". In: *Earth's Future* 5.10, pp. 1034–1043.

Heinzerling, L. (2006). "Doubting Daubert". In: *JL & Pol'y* 14, p. 65.

Herring, S. C., Christidis, N., Hoell, A., Kossin, J. P., Schreck III, C. J., and Stott, P. A. (2018a). "Explaining extreme events of 2016 from a climate perspective". In: *Bulletin of the American Meteorological Society* 97.12, S1–S157.

Herring, S. C., Christidis, N., Hoell, A., Kossin, J. P., Schreck III, C. J., and Stott, P. A. (2018b). "Introduction to explaining extreme events of 2016 from a climate perspective". In: *Bulletin of the American Meteorological Society* 97.12, S1–S6.

Herring, S. C., Hoell, A., Hoerling, M. P., Kossin, J. P., Schreck III, C. J., and Stott, P. A. (2016a). "Explaining extreme events of 2015 from a climate perspective". In: *Bulletin of the American Meteorological Society* 97.12, S1–S145.

Herring, S. C., Hoell, A., Hoerling, M. P., Kossin, J. P., Schreck III, C. J., and Stott, P. A. (2016b). "Introduction to explaining extreme events of 2015 from a climate perspective". In: *Bulletin of the American Meteorological Society* 97.12, S1–S3.

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

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.

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.

Howells, G. G. and Mildred, M. (1997). "Is European products liability more protective than the restatement (third) of torts: products liability". In: *Tenn. L. Rev.* 65, p. 985.

Immordino, G., Pagano, M., and Polo, M. (2011). "Incentives to innovate and social harm: Laissez-faire, authorization or penalties?" In: *Journal of Public Economics* 95.7, pp. 864–876.

IPCC (2014). *Climate Change 2014: Impacts, Adaptation, Vulnerability. Summary for policymakers. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

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

Jasanoff, S. (2005). "Law's knowledge: science for justice in legal settings". In: *American journal of public health* 95.S1, S49–S58.

Ji, S.-r., Liu, C., Zhang, B., Yang, F., Xu, J., Long, J., Jin, C., Fu, D.-l., Ni, Q.-x., and Yu, X.-j. (2010). "Carbon nanotubes in cancer diagnosis and therapy". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1806.1, pp. 29–35.

Kamien, M. I., Muller, E., and Zang, I. (1992). "Research joint ventures and R&D cartels". In: *The American Economic Review*, pp. 1293–1306.

Keith, D. (2013). *A case for climate engineering*. MIT Press.

Keith, D. W. (2010). "Photophoretic levitation of engineered aerosols for geoengineering". In: *Proceedings of the National Academy of Sciences* 107.38, pp. 16428–16431.

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

Kessler, F. (1967). "Products liability". In: *The Yale Law Journal* 76.5, pp. 887–938.

Klepper, G. and Rickels, W. (2012). "The real economics of climate engineering". In: *Economics Research International* 2012.

Kostarelos, K. (2008). "The long and short of carbon nanotube toxicity". In: *Nature biotechnology* 26.7, pp. 774–776.

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. (2014). "A multi-model assessment of regional climate disparities caused by solar geoengineering". In: *Environmental Research Letters* 9.7: 074013.

Kravitz, B., Rasch, P. J., Forster, P. M., Andrews, T., Cole, J. N., Irvine, P. J., Ji, D., Kristjánsson, J. E., Moore, J. C., Muri, H., et al. (2013). "An energetic perspective on hydrological cycle changes in the Geoengineering Model Intercomparison Project". In: *Journal of Geophysical Research: Atmospheres* 118.23.

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.

La Serre, D., Barbier, E., and Sibony, A.-L. (2008). "Expert evidence before the EC courts". In: *Common Market L. Rev.* 45, p. 941.

Landes, W. M. and Posner, R. A. (1985). "A positive economic analysis of products liability". In: *The Journal of Legal Studies* 14.3, pp. 535–567.

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

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.

Lusk, G. (2017). "The social utility of event attribution: liability, adaptation, and justice-based loss and damage". In: *Climatic Change* 143.1-2, pp. 201–212.

MacMartin, D. G., Kravitz, B., Long, J., and Rasch, P. J. (2016). "Geoengineering with stratospheric aerosols: What do we not know after a decade of research?" In: *Earth's Future* 4.11, pp. 543–548.

Mann, M. E., Lloyd, E. A., and Oreskes, N. (2017). "Assessing climate change impacts on extreme weather events: the case for an alternative (Bayesian) approach". In: *Climatic change* 144.2, pp. 131–142.

Marjanac, S. and Patton, L. (2018). "Extreme weather event attribution science and climate change litigation: an essential step in the causal chain?" In: *Journal of Energy & Natural Resources Law*, pp. 1–34.

Marjanac, S., Patton, L., and Thornton, J. (2017). "Acts of God, human influence and litigation". In: *Nature Geoscience* 10.9, pp. 616–619.

McAvaney, B., Covey, C., Joussaume, S., Kattsov, V., Kitoh, A., Ogana, W., Pitman, A., Weaver, A., Wood, R., Zhao, Z.-C., et al. (2001). "Model evaluation". In: *Climate Change 2001: The scientific basis. Contribution of WG1 to the Third Assessment Report of the IPCC (TAR)*. Cambridge University Press, pp. 471–523.

McCormick, S., Simmens, S. J., Glicksman, R. L., Paddock, L., Kim, D., Whited, B., and Davies, W. (2017). "Science in litigation, the third branch of US climate policy". In: *Science* 357.6355, pp. 979–980.

McGarity, T. O. (2004). "Our science is sound science and their science is junk science: Science-based strategies for avoiding accountability and responsibility for risk-producing products and activities". In: *U. Kan. L. Rev.* 52, p. 897.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M., Lamarque, J.-F., Matsumoto, K., Montzka, S., Raper, S., Riahi, K., et al. (2011). "The RCP greenhouse gas concentrations and their extensions from 1765 to 2300". In: *Climatic change* 109.1-2, p. 213.

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.

Murray Jr, W. R. (1982). "Requiring Omniscience: The Duty to Warn of Scientifically Undiscoverable Product Defects". In: *Geo. LJ* 71, p. 1635.

National Academy of Sciences (2016). *Attribution of extreme weather events in the context of climate change*. National Academies Press.

Niemeier, U., Schmidt, H., and Timmreck, C. (2011). "The dependency of geoengineered sulfate aerosol on the emission strategy". In: *Atmospheric Science Letters* 12.2, pp. 189–194.

Niemeier, U., Schmidt, H., Alterskjær, K., and Kristjánsson, J. (2013). "Solar irradiance reduction via climate engineering: Impact of different techniques on the energy balance and the hydrological cycle". In: *Journal of Geophysical Research: Atmospheres* 118.21.

Niemeier, U. and Tilmes, S. (2017). "Sulfur injections for a cooler planet". In: *Science* 357.6348, pp. 246–248.

Nordhaus, W. D. (2011). "The economics of tail events with an application to climate change". In: *Review of Environmental Economics and Policy* 5.2, pp. 240–257.

Nunner-Krautgasser, B. and Anzensberger, P. (2016). "Inadmissible Evidence: Illegally Obtained Evidence and the Limits of the Judicial Establishment of the Truth". In: *Dimensions of evidence in European civil procedure*. Kluwer Law International, pp. 195–212.

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

O'Reilly, J. T. (1987). "Biotechnology Meets Products Liability: Problems Beyond the State of the Art". In: *Hous. L. Rev.* 24, p. 451.

Otto, F. E. (2016). "Extreme events: The art of attribution". In: *Nature Climate Change* 6.4, p. 342.

Otto, F. E., Massey, N., Oldenborgh, G., Jones, R., and Allen, M. (2012). "Reconciling two approaches to attribution of the 2010 Russian heat wave". In: *Geophysical Research Letters* 39.4.

Otto, F. E., Skeie, R. B., Fuglestvedt, J. S., Berntsen, T., and Allen, M. R. (2017). "Assigning historic responsibility for extreme weather events". In: *Nature Climate Change* 7.11, p. 757.

Otto, F. E. L. (2012). "Modelling the earth's climate-an epistemic perspective". PhD thesis. Freie Universität Berlin.

Owen, D. G. (2010). "Bending Nature, Bending Law". In: *Fla. L. Rev.* 62, p. 569.

Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G., Lohmann, D., and Allen, M. R. (2011). "Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000". In: *Nature* 470.7334, p. 382.

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

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

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

Petersen, A. C. (2012). *Simulating nature: a philosophical study of computer-simulation uncertainties and their role in climate science and policy advice*. CRC Press.

Pfrommer, T. (2018). "A model of solar radiation management liability". In: *AWI Discussion Paper Series* 644. Available at: https://www.uni-heidelberg.de/md/awi/institut/awlecture/dp644.pdf.

Pindyck, R. S. (2011). "Fat tails, thin tails, and climate change policy". In: *Review of Environmental Economics and Policy* 5.2, pp. 258–274.

Pitari, G., Aquila, V., Kravitz, B., Robock, A., Watanabe, S., Cionni, I., Luca, N. D., Genova, G. D., Mancini, E., and Tilmes, S. (2014). "Stratospheric ozone response to sulfate geoengineering: Results from the Geoengineering Model Intercomparison Project (GeoMIP)". In: *Journal of Geophysical Research: Atmospheres* 119.5, pp. 2629–2653.

Poland, C. A., Duffin, R., Kinloch, I., Maynard, A., Wallace, W. A., Seaton, A., Stone, V., Brown, S., MacNee, W., and Donaldson, K. (2008). "Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study". In: *Nature nanotechnology* 3.7, pp. 423–428.

Polinsky, A. M. (1980). "Strict Liability vs. Negligence in a Market Setting". In: *The American Economic Review*, pp. 363–367.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Routledge.

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

Quaas, M. F., Quaas, J., Rickels, W., and Boucher, O. (2017). "Are there reasons against open-ended research into solar radiation management? A model of intergenerational decision-making under uncertainty". In: *Journal of Environmental Economics and Management* 84, pp. 1–17.

Rahmstorf, S. and Coumou, D. (2011). "Increase of extreme events in a warming world". In: *Proceedings of the National Academy of Sciences* 108.44, pp. 17905–17909.

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.

Reimann, M. (2003). "Liability for defective products at the beginning of the twenty-first century: Emergence of a worldwide standard?" In: *The American Journal of Comparative Law* 51.4, pp. 751–838.

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.

Rhee, C. van (2016). "Evidence law in an international context: The principles of transnational civil procedure". In: *Dimensions of evidence in European civil procedure*. Kluwer Law International, pp. 11–28.

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.

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.

Riddell, A. and Plant, B. (2009). *Evidence before the international court of justice*. British Institute of International and Comparative Law.

Robb, G. C. (1982). "Practical Approach to Use of State of the Art Evidence in Strict Products Liability Cases". In: *Nw. UL Rev.* 77, p. 1.

Robock, A. (2008). "20 reasons why geoengineering may be a bad idea". In: *Bulletin of the Atomic Scientists* 64.2, pp. 14–18.

Robock, A., Marquardt, A., Kravitz, B., and Stenchikov, G. (2009). "Benefits, risks, and costs of stratospheric geoengineering". In: *Geophysical Research Letters* 36.19.

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.

Rogelj, J., Den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., Schaeffer, R., Sha, F., Riahi, K., and Meinshausen, M. (2016). "Paris Agreement climate proposals need a boost to keep warming well below 2 C". In: *Nature* 534.7609, p. 631.

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.

Schäfer, S., Lawrence, M., Stelzer, H., Born, W., Low, S., Aaheim, A., Adriázola, P., Betz, G., Boucher, O., Carius, A., et al. (2015). "The European transdisciplinary assessment of climate engineering (EuTRACE): Removing greenhouse gases from the atmosphere and reflecting sunlight away from Earth". In: *Funded by the European Union's Seventh Framework Programme under Grant Agreement* 306993.

Schlenker, W. and Roberts, M. J. (2009). "Nonlinear temperature effects indicate severe damages to US crop yields under climate change". In: *Proceedings of the National Academy of sciences* 106.37, pp. 15594–15598.

Seager, R., Hoerling, M., Schubert, S., Wang, H., Lyon, B., Kumar, A., Nakamura, J., and Henderson, N. (2015). "Causes of the 2011–14 California drought". In: *Journal of Climate* 28.18, pp. 6997–7024.

Shavell, S. (1980). "Strict liability versus negligence". In: *The Journal of Legal Studies*, pp. 1–25.

Shavell, S. (1984a). "A model of the optimal use of liability and safety regulation". In: *The Rand Journal of Economics* 15.2, pp. 271–280.

Shavell, S. (1984b). "Liability for harm versus regulation of safety". In: *The Journal of Legal Studies* 13.2, pp. 357–374.

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

Shavell, S. (1992). "Liability and the incentive to obtain information about risk". In: *The Journal of Legal Studies*, pp. 259–270.

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

Shiogama, H., Watanabe, M., Imada, Y., Mori, M., Ishii, M., and Kimoto, M. (2013). "An event attribution of the 2010 drought in the South Amazon region using the MIROC5 model". In: *Atmospheric Science Letters* 14.3, pp. 170–175.

Sillmann, J., Kharin, V., Zhang, X., Zwiers, F., and Bronaugh, D. (2013). "Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate". In: *Journal of Geophysical Research: Atmospheres* 118.4, pp. 1716–1733.

Sladič, J. and Uzelac, A. (2016). "Assessment of Evidence". In: *Dimensions of Evidence in European Civil Procedure*. Wolters Kluwer, pp. 107–132.

Stein, A. and Parchomovsky, G. (2008). "The Anti-Innovation Bias of Tort Law". In: *Mich. L. Rev.* 107, pp. 285–305.

Stier, P., Feichter, J., Kinne, S., Kloster, S., Vignati, E., Wilson, J., Ganzeveld, L., Tegen, I., Werner, M., Balkanski, Y., et al. (2005). "The aerosol-climate model ECHAM5-HAM". In: *Atmospheric Chemistry and Physics* 5.4, pp. 1125–1156.

Stott, P. A., Christidis, N., Otto, F. E., Sun, Y., Vanderlinden, J.-P., Van Oldenborgh, G. J., Vautard, R., Von Storch, H., Walton, P., Yiou, P., et al. (2016). "Attribution of extreme weather and climate-related events". In: *Wiley Interdisciplinary Reviews: Climate Change* 7.1, pp. 23–41.

Stott, P. A., Christidis, N., Herring, S. C., Hoell, A., Kossin, J. P., and Schreck III, C. J. (2018). "Future Challenges in Event Attribution Methodologies". In: *Bulletin of the American Meteorological Society* 99.1, S155–S157.

Stott, P. A., Karoly, D. J., and Zwiers, F. W. (2017). "Is the choice of statistical paradigm critical in extreme event attribution studies?" In: *Climatic change* 144.2, pp. 143–150.

Stott, P. A., Stone, D. A., and Allen, M. R. (2004). "Human contribution to the European heatwave of 2003". In: *Nature* 432.7017, p. 610.

Swinehart, M. W. (2007). "Remedying Daubert's Inadequacy in Evaluating the Admissibility of Scientific Models Used in Environmental-Tort Litigation". In: *Tex. L. Rev.* 86, p. 1281.

Thornton, J. and Covington, H. (2016). "Climate change before the court". In: *Nature Geoscience* 9.1, p. 3.

Tilmes, S., Fasullo, J., Lamarque, J.-F., Marsh, D. R., Mills, M., Alterskjær, K., Muri, H., Kristjánsson, J. E., Boucher, O., Schulz, M., et al. (2013). "The hydrological impact of geoengineering in the Geoengineering Model Intercomparison Project (GeoMIP)". In: *Journal of Geophysical Research: Atmospheres* 118.19.

Tilmes, S., Müller, R., and Salawitch, R. (2008). "The sensitivity of polar ozone depletion to proposed geoengineering schemes". In: *Science* 320.5880, pp. 1201–1204.

Tomka, H. and Proulx, V.-J. (2015). "The evidentiary practice of the world court". In: *SSRN Working Paper*. Available at: https://papers.ssrn.com/abstract=2693558.

Trenberth, K. E. and Dai, A. (2007). "Effects of Mount Pinatubo volcanic eruption on the hydrological cycle as an analog of geoengineering". In: *Geophysical Research Letters* 34.15.

Trenberth, K. E., Fasullo, J. T., and Shepherd, T. G. (2015). "Attribution of climate extreme events". In: *Nature Climate Change* 5.8, p. 725.

Vicente-Serrano, S. and Beguería-Portugués, S. (2003). "Estimating extreme dry-spell risk in the middle Ebro valley (northeastern Spain): a comparative analysis of partial duration series with a general Pareto distribution and annual maxima series with a Gumbel distribution". In: *International Journal of Climatology* 23.9, pp. 1103–1118.

Viscusi, W. K. and Moore, M. J. (1993). "Product liability, research and development, and innovation". In: *Journal of Political Economy*, pp. 161–184.

Wagner, W. (2005). "The perils of relying on interested parties to evaluate scientific quality". In: *American Journal of Public Health* 95.S1, S99–S106.

Weitzman, M. L. (2009). "On modeling and interpreting the economics of catastrophic climate change". In: *The Review of Economics and Statistics* 91.1, pp. 1–19.

Weitzman, M. L. (2014). "Fat tails and the social cost of carbon". In: *American Economic Review* 104.5, pp. 544–46.

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.

Wilhite, D. A. and Glantz, M. H. (1985). "Understanding: the drought phenomenon: the role of definitions". In: *Water international* 10.3, pp. 111–120.

Williams, A. P., Seager, R., Abatzoglou, J. T., Cook, B. I., Smerdon, J. E., and Cook, E. R. (2015). "Contribution of anthropogenic warming to California drought during 2012–2014". In: *Geophysical Research Letters* 42.16, pp. 6819–6828.

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.