Master's Thesis - Integrated Climate System Sciences

Decision Under Climate Uncertainty: Learning in Target-Based Approaches

Felix Schreyer

Matriculation number: 6640854

Born: 21st of March 1991 in Leipzig

School of Integrated Climate System Sciences

University of Hamburg

Hamburg, 15th of December 2016

Title:

Decision Under Climate Uncertainty: Learning in Target-Based Approaches.

This thesis has been accepted as a Master Thesis by the Department of Geowissenschaften der Universität Hamburg.

1. Reviewer: Prof. Dr. Hermann Held, Research Unit Sustainability and Global Change (FNU), Universität Hamburg.

2. Reviewer: Prof. Dr. Andreas Lange, Department of Economics, Universität Hamburg.

Abstract

We investigate different decision criteria for treating climate targets in a setting of uncertainty and learning. Developed as an implementation of strong sustainability, climate targets have been traditionally understood as maximum acceptable limits of global warming. Reflecting this idea in decision criteria for the climate problem implies using a lexicographic structure: Meeting the climate target should be the primary criterion, while reducing mitigation cost is secondary. The structure can be preserved under climate-related uncertainty by formulating probabilistic climate targets. Yet, extending this position to a situation of future learning may give rise to normatively unappealing effects as a probabilistic cost-effectiveness analysis deviates from expected utility theory. We point to these effects by discussing the relevance of the von Neumann-Morgenstern axioms in the context of the climate problem. Instead of cost-effectiveness analysis, cost-risk analysis can be used which is an expected utility criterion based on a risk measure of overshooting the climate target. We discuss the features of this criterion against the background of strong sustainability. Finally, we explore a third target-based decision criterion that minimizes the target overshoot for a given budget of mitigation cost.

Nomenclature

BAU	Business-as-usual (baseline scenario without emission reduction)
CBA	Cost-benefit analysis
CEA	Probabilistic cost-effectiveness analysis
CRA	Cost-risk analysis
IAM	Integrated assessment model
MIND	Model of Investment and technological Development
Mitigation	Mitigation of climate change (emission reduction)
MRA	Minimum Risk Analysis
Posterior	After learning
Prior	Before learning
Temperature	Global mean surface temperature relative to preindustrial times

Contents

Introduction

Imagine the international community follows the Paris climate agreement of 2015 and sets off to hold global temperature increase to “well below 2°C” (UNFCCC, 2015). Every nation strictly keeps to its plan for meeting this common climate target. Now, suppose by 2030 climate scientists discover that climate change is much stronger than expected. The envisaged amount of greenhouse gas emissions cause considerably more warming than 2°C. How should climate policy react? Should decision makers take more action to reduce greenhouse gas emissions? If so, how much more? Should they still aim at a 2°C target? No matter how we respond to any of these questions, as in the future we may have more knowledge about the climate system, such situations should be consistently anticipated in today’s decision making. This study will investigate different approaches for taking this prospect of future learning into account.

Different decision criteria have been developed in climate economic modeling to derive policy recommendations with respect to climate targets. Edenhofer et al. (2005) introduced cost-effectiveness analysis for investigating optimal mitigation policies to reach climate targets. It is the standard approach used in the IPCC (2014b, pp. 413-510) to address numerous questions on economic transformation pathways for climate stabilization. Held et al. (2009) extended the approach to take relevant uncertainty in the climate and the economic system into account. Finally, Schmidt et al. (2011) developed cost-risk analysis, a criterion which can consistently deal with future learning about the climate system. Neubersch et al. (2014) applied this decision criterion to perfect learning about climate sensitivity in the model MIND-L and derived necessary conditions for the convexity of the risk function employed.

This study asks how to consistently formulate target-based criteria to make decisions on the climate problem under learning. In fact, considerable parts can be seen as a review of Schmidt et al. (2009) and Schmidt et al. (2011) who laid the foundations to this discussion. Our study revisits the problem, discussing it against the normative background of strong sustainability: Traditionally, strong sustainability implies using a lexicographic structure of decision criteria: Meeting the climate target should be the primary criterion, while reducing mitigation cost is secondary. Yet, extending this position to a situation of future learning may imply breaking with expected utility theory, the standard framework of decision making under uncertainty. This is why we will discuss the relevance of the von Neumann-Morgenstern axioms of expected utility theory in the context of the climate problem. We aim to answer the following question:

How can strong sustainability be formalized in a consistent decision criterion for the climate problem under uncertainty and future learning about climate sensitivity?

Let us specify the role that we wish to take as decision analysts when discussing different criteria against the normative background of strong sustainability. Keeney (1982) conceives decision analysis as a dialog between a decision maker and a decision analyst: Dealing with simple decision problems, the decision maker can normally rely on her intuition. However, suppose the problem is sufficiently complex, the decision maker could fail to see the full implications of her choice. Here, the decision analyst comes in by formalizing the choice and detecting possible inconsistencies to let her reconsider the decision. The goal of the procedure is to elucidate the decision problem and find the best option for the decision maker.

Value-free decision analysis is neither possible nor desirable, as Keeney (1982) emphasizes. First, Sen (1993) famously argued that "internal consistency" without reference to motivations or principles outside of decision theory does not exist. Consistency principles, no matter how natural they may appear, require normative discussion and justification against the background of the specific decision problem. Second, "decision analysts try to formalize the thinking and the feelings that the decision maker wishes to use on the problem" (Keeney, 1982, p. 819). This clearly requires acts of translation and interpretation. It is not a deficiency, rather a necessity when engaging in the aspired dialog. Decision analysis aims at clarifying the thought process in decision making as to make it accessible to review and adjustment.

We will structure our considerations as follows: Chapter 1 presents the climate problem and the role of climate targets according to strong sustainability. It explains the complexities raised by uncertainty and learning and introduces the two existing decision criteria. Chapter 2 presents formal concepts of decision theory and presents the expected utility framework, the standard approach used in decision making under uncertainty. Subsequently, chapters 3 and 4 will analyze the two existing target-based decision criteria under learning: probabilistic cost-effectiveness and cost-risk analysis. Finally, chapter 5 introduces minimum-risk analysis as a new target-based decision criterion and gives a brief outlook on its possible chances and problems. Chapters 3 to 5 respectively address the question, under which condition the investigated criterion can be seen as an adequate formalizations of strong sustainability under learning.

1 The Decision Problem: Meeting Climate Targets under Uncertainty

The broad framing of this first chapter may be justified by providing a background that helps to understand the key normative perspective this study deals with. The climate problem raises the complex question of how the manifold risks of climate change can be related to the economic cost of emission reduction (section 1.1). There are two major normative framings to generally approach this question: the cost-benefit approach and the climate target approach. The first assesses monetized climate damages to be traded-off against mitigation cost (section 1.2), while the second works with maximum acceptable limits of global warming (section 1.3). How this target-based framing can be formalized in a decision criterion under uncertainty and learning will be the question tackled by this study (section 1.4). So far, two criteria have been suggested: Probabilistic cost-effectiveness analysis and cost-risk analysis (section 1.5). They will be discussed in more detail in chapters 3 and 4.

1.1 The Climate Problem

The climate problem is a dilemma of modern economic development. On the one hand, to seriously cut emissions, a costly and technologically demanding transformation of the energy system, the backbone of modern industrialized economies, would be necessary. The fossil-fuel sector would need to be replaced by hitherto more expensive and less stable low-carbon energy sources. On the other hand, depending on the amount of emissions, climate change may pose very severe threats to human societies and natural ecosystems. The list of potential adverse impacts compiled by the IPCC (2014a) is long and alarming. The key question is how much greenhouse gases should still be emitted in the face of climate change. Without a doubt, the weights to be balanced for this question are huge.

Integrated Assessment Models (IAMs) can help to guide decision making on the climate problem by analyzing emission scenarios from a social planer perspective (IPCC, 2014b, pp. 178-181). The models simulate interactions between the economic system, the energy system and the climate system. Not only do they allow for estimating the negative effect of emission reductions on macroeconomic growth, i.e. the *mitigation cost* of climate change. They also run climate and impact simulations to determine and assess the *climate impacts* caused by emissions. There are various comprehensive IAMs such as DICE[1], FUND[2] or REMIND[3], some of which aim to cover a broad range of aspects such as energy markets, crop technologies, climate-induced sea-level rise or extreme weather.

In this study, we will focus on the decision analysis which can be done with IAMs. Different decision criteria have been suggested to deal with the climate problem. In the most general sense, they all weigh some assessment of climate impacts against mitigation cost. Obviously, given the vast empirical as well as normative complexities of the problem, it is not surprising that already the meta-decision on an adequate decision criterion is difficult and highly contentious. Any approach sets the multi-faceted climate problem into a particular normative framing. The two major framings are the *cost-benefit framing* (section 1.2) and the *climate target framing* (section 1.3). As decision analysts, we will not focus on justifying either of them. Rather, we will investigate more formally how the target-based framing can be adequately formalized in a decision criterion under uncertainty and learning.

The climate problem is in many ways an exceptional decision problem. Let us point to some of its complexities more specifically:

First, there are numerous uncertainties when using IAMs for estimating the economic or environmental implications of emission reductions. They are partly as large that some authors question the usefulness of IAMs in general (Pindyck, 2013). Second, climate change affects human societies in manifold and potentially existential ways such that even if impacts could be well predicted, it would be ethically far from clear how to assess and compare them (e.g. Ackerman et al., 2009).

Third, the climate problem is an intergenerational problem. Most of the people affected are not yet born, which raises fundamental questions about the nature of their rights and the duties of present generations (IPCC, 2014b, pp. 216-220, 223-224). Furthermore, the mitigation cost generally occur decades earlier than the associated benefits of mitigation, i.e. the avoided climate impacts. The main

[1] e.g. Nordhaus (2013)

[2] e.g. Tol (2013), www.fund-model.org

[3] e.g. Luderer et al. (2013)

reason is that the warming response to greenhouse gas emissions is delayed by the ocean heat uptake which is slow in comparison to land. The oceans receive more than 90% of the additional energy input to the Earth system and need decades for restoring radiative balance with greenhouse forcing as to reach equilibrium temperature (IPCC, 2013a, pp. 264-265).

Fourth, this time delay allows to continuously re-evaluate the decisions on emission policy. The mentioned uncertainties may be reduced by additional climate observations or advances in the scientific understanding of physical processes. This knowledge will be used to update IAMs in the future. Adapting our choices consistently to new information will be referred to as decision making under learning and is the focus of this study.

In the next two sections, we will present the two major normative framings to approach decision making on the climate problem. The first is the cost-benefit framing, the standard approach in environmental economics. The second framing works with climate targets to be understood as maximum acceptable limits to climate change.

1.2 The Cost Benefit Approach

The climate problem confronts us with a trade-off between mitigation cost and climate impacts. Cost-benefit analysis (CBA) tackles the problem in the most straightforward manner: It weighs the cost of mitigation against aggregated economic damages from all kinds of climate impacts in monetary terms.

CBA represents the standard economic approach to deal with environmental problems by internalizing unconsidered externalities into a welfare optimization. Marginal damage functions of environmental harm are assessed and balanced with the marginal cost of emission abatement (see e.g. Perman et al., 1996, pp. 204-208). As the “the mother of all externalities”[4], climate change can be tackled accordingly.

[4] In the words of Richard Tol (Tol, 2009, p. 29).

CBA conducts the following optimization:

$$\text{Min}_E \quad \int (C(E) + D(E))e^{-\rho t} dt. \tag{1}$$

Here, $E = E(t)$ denotes the pathway of global greenhouse gas emissions over time, $C(E)$ the total mitigation cost and $D(E)$ the total damages induced by climate change[5]. In general, we refer to these quantities as functions over time, although for notational convenience we omit writing the time dependence explicitly. The mitigation cost are the welfare losses relative to a business-as-usual (BAU) scenario, i.e. a future growth scenario in which climate impacts are ignored and emissions continue to grow efficiently. The climate damages represent the total market and non-market impacts of climate change in terms of welfare. Climate consequences are modeled explicitly and internalized into the social planer analysis. CBA then seeks to find the optimal level of climate change mitigation.

As is common for intertemporal optimization, the objective function is discounted by the factor $e^{-\rho t}$ and aggregated over time, where ρ is rate of pure time preference. It is a measure of how much future welfare is devalued relative to present welfare. The discount factor significantly influences the optimum as mitigation cost occur generally earlier than climate damages. A value of $\rho = 0$ implies that the welfare at any future point in time is equally valued as the present welfare. The proper way of discounting has been subject to extensive ethical debates (see Davidson, 2015). The sensitivity of the results to this contentious parameter is one of multiple problems CBA faces.

Cost-benefit analyses of the climate problem have been conducted with several IAMs. Many of the results indicate that the economic benefits of emissions outweigh the climate damages up to a considerable warming of several degrees in global mean temperature above preindustrial times[6]: In the version of Nordhaus (2008), the DICE model suggests an optimal temperature of 3.5°C by the end of

[5] Throughout this study, climate damages and mitigation cost will be considered as aggregated quantities. Questions of interpersonal comparisons, distribution and adequate welfare measures are not treated here. We refer the interested reader to Adler and Treich (2015) or the IPCC (2014b, pp. 221-223).

[6] The preindustrial time as defined by the (IPCC, 2013b, p. 1264) refers somewhat roughly to the period before 1750 i.e. before significant anthropogenic greenhouse gas emissions occurred.

the 22th century (pp. 82-83)[7]. The MERGE model used by Manne and Richels (2005) recommends a pathway of around 3°C warming by 2150 similar to the optimum found by the PAGE model of Hope (2008) for a best-estimate scenario.

However, many critics consider a temperature rise of this scale temperature rise to be unreasonably high, as can be seen in the discourse on the 2°C target (see Jaeger and Jaeger, 2011). The focal point of the CBA criticism is the construction of a damage function that is to be traded-off against mitigation cost. Three lines of argument have been brought forward.

First, there is fundamental uncertainty about the specific impacts associated with a climate change of this magnitude (Charlesworth and Okereke, 2010). A damage function, it is argued, needs to assume a degree of climate predictability which cannot be reached for such a complex system featuring non-linear feedback mechanisms and tipping points. Uncertainties about this system are partly irreducible and can often not even be represented by probability distributions.

Second, CBA comes with precise valuations of all possible kinds of climate damages, a practice which must necessarily rely on many ethically contestable methods and assumptions (Ackerman et al., 2009). It needs to monetize, for instance, the value of life, health, cultural heritages and biodiversity. Obviously, many of these values at stake seem "incomparable" (IPCC, 2014b, p. 220). Yet, CBA must set explicit trade-off parameters to integrate each of them. It forces detailed quantitative balancing where, in general, qualitative reasoning might be more adequate.

Third, also the question of compensatory justice points to the limits of economic assessment of climate change (IPCC, 2014b, p. 224). Reasoning on the basis of trade-offs implies that compensation of future generations is generally acceptable for any sort of climate damages. However, it has been brought forward that climate change leads to infringements of future people's (human) rights, which could be avoided, so the argument goes, at comparatively reasonable cost (Caney, 2008). The cost-benefit framing has difficulties to integrate such rights-based ethical reasoning.

[7] Nordhaus (2013) obtains an optimum of 2.3°C maximum temperature in a more recent version of DICE and for zero discounting (pp. 207-208).

1.3 Climate Targets

The second framing to deal with the climate problem is based on climate targets. It features prominently in the international climate policy discourse. The recent 21st Conference of the Parties formulated the objective of "holding the increase in global average temperature to well below 2°C above preindustrial levels" (UNFCCC, 2015). Such climate targets are generally not justified by detailed impact analysis but by more qualitative arguments. The target-based framing evades the main criticism of CBA as there is no damage function and diverse ethical aspects such as the role of rights or precaution in the light of fundamental uncertainty may feature the justification of the climate target. It externalizes the evaluation of climate change to political discourse and leaves economic analysis with cost-effectiveness questions of reaching exogenous targets.

The question arises of how strict such a target needs to be understood. Although other interpretations are possible[8], we consider the interpretation promoted by the influential German Advisory Council of Global Change (WBGU). It is the very institution which introduced the 2°C target to the policy debate and shaped major arguments (see Jaeger and Jaeger, 2011). The WBGU (2014) speaks of "planetary guard rails" to be understood as

> "damage thresholds whose transgression either today or in future would have such intolerable consequences that even large-scale benefits in other areas could not compensate these." (p. 11)

The WBGU endorses a strict interpretation. They claim that holding global temperature below the guard rail should be the primary concern, while mitigation cost are secondary. They reject any trade-off between the "intolerable consequences" and mitigation cost. The guard rail is a *maximum acceptable limit*.

The idea of maximum acceptable limits is known in the sustainability discourse as *strong sustainability*. The positions of weak and strong sustainability differ depending on whether the natural capital at stake in the environmental problem (e.g. clean air, ecosystems or resources) is generally substitutable or

[8] See the "Three Views on 2 degrees" by Jaeger and Jaeger (2011).

non-substitutable by human-made capital (e.g. machines, software or knowledge) (Neumayer, 2013; Perman et al., 1996, pp. 22-29, 59-60). Considering the many forms of natural capital that may be adversely affected by climate change, Neumayer (2013, pp. 40-46) argues that the main controversy between the cost-benefit and the target-based framing is about the substitutability of environmental and economic values. Applied to the climate system as a whole, holding “safe minimum standards”, as suggested by strong sustainability (Neumayer, 2013, pp. 110-115, 118), corresponds to the very idea behind the guard rail concept of the WBGU. This is why, in the remainder, we will refer to their understanding of the target-based normative framing as *strong sustainability*. It implies to structure the problem into a primary climate criterion and a secondary cost criterion.

Proponents of strong sustainability needs to be able to justify a specific climate target. The WBGU (1995) argued that global mean temperature has never been higher than 1.5°C above preindustrial over the late Quaternary, i.e. the last 800,000 years. It is a range which is still relatively familiar to climate research by paleoclimatic evidence. Adding somewhat arbitrarily a tolerance of 0.5°C, they put forward a limit of 2°C as to preserve an environment similar to present-day conditions. Meanwhile, more impact estimates have been done and lead the IPCC (2014a, pp. 61-62) to conclude that the risks from climate change increase starkly between 2-3°C global warming. At higher temperatures, large-scale tipping points could be triggered to induce major irreversible changes in the Earth system. Yet, there are large uncertainties about the level of these thresholds (IPCC, 2013a, pp. 1114-1119). This has raised a second line of argument: In the light of these disastrous and hardly quantifiable events, the WBGU (2014) suggests the 2°C level as a precautionary limit to stay away from large-scale changes.

However, making a climate target the primary objective of global economic policy is obviously a strong claim, too. The corresponding criticism brought forward by, unsurprisingly, proponents of CBA is that “the simple target approach is unworkable because it ignores the cost of attaining the goals” (Nordhaus, 2013, p. 7). If at all, cost considerations have featured the arguments for a 2°C target more implicitly. The WBGU (1995) advocated the 2°C limit as a “tolerable window” of climate change that does not impose “excessive cost”. The IPCC (2014b, pp. 448-451) estimated mitigation cost in the range of a

0.04%-0.14% average reduction in global consumption growth over the course of the 21th century for close to 2°C scenarios [9]. From different perspectives it has been argued on qualitative grounds that the mitigation cost necessary for the 2°C target do not compare to the climate risks looming beyond that level (e.g. Caney, 2008; Steigleder, 2016). If mitigation cost were very high, the argument of strong sustainability would lose its force.

Finding a decision criterion for strong sustainability without uncertainty is straightforward. The question is reduced to how to cost-effectively comply with the climate target. The corresponding decision criterion is (deterministic) cost-effectiveness analysis given by

$$\begin{aligned} &\mathrm{Min}_E \int C(E)e^{-\rho t}dt, \\ &\text{s.t. } T_{max}(E) \leq T_g. \end{aligned} \tag{2}$$

As above, E is the emission pathway, $C(E)$ the mitigation cost and ρ the pure rate of time preference. Furthermore, $T_{max}(E)$ denotes the maximum temperature reached over time and T_g represents the temperature guard rail. Cost-effectiveness analysis selects the emission pathway which keeps temperature below the guard rail T_g at minimal mitigation cost. As the discounting only affects mitigation cost, it has less influence on the optimum than in CBA where also climate damages are discounted.

1.4 Uncertainty and Learning

The above idea of strong sustainability relies on the assumption that for any given emission pathway it can be exactly determined whether or not the guard rail will be transgressed. Unfortunately, this is far from the reality. Due to its physically complex nature, the climate system is inherently difficult to predict. There are limits to resolving its uncertainties in the light of hardly predictable feedback mechanisms.

Target-based decision criteria primarily face uncertainty about the response of temperature to emissions. There are several uncertainties in the carbon cycle

[9] They refer to a range of greenhouse gas concentration of 430-480 ppm CO2eq.

as well as uncertainty about climate sensitivity. Climate sensitivity is defined as the difference in global mean temperature reached in equilibrium for a doubling of preindustrial greenhouse gas concentrations. To make our point in this study, it will be sufficient to consider the uncertainty about climate sensitivity: Numerous probability distributions have been estimated based on different methods and datasets (IPCC, 2013a, pp. 921-926). They indicate that climate sensitivity is likely between 1.5°C and 4.5°C. However, not only are there questions about the reliability of these estimates, the resulting distributions have uncomfortable implications: It is impossible to give an upper bound to climate sensitivity (long tail) and for high values the probability density declines less fast than in exponential distributions, making extreme outcomes relatively more likely (fat tail).

The struggle with climate sensitivity is as hard since it is an aggregation of numerous feedback mechanisms in the climate system (Roe and Baker, 2007; Allen et al., 2009). As a complex system the climate becomes harder to predict the more it moves away from its initial state. The climate sensitivity to increased greenhouse gas concentrations without feedbacks is relatively straightforward to calculate (1.2-1.3°C, see Planck feedback parameter in Bony et al., 2006). However, this temperature increase changes many processes of the climate system that further reinforce the warming. They are known as positive *feedback mechanisms*. Finally, in a world of e.g. 4°C warming, the conditions would be so different from today that it is almost impossible to say anything about these processes and predict whether warming would stop. This is what Roe and Baker (2007) argue in a more formal sense, examining the implications of fat tails in climate sensitivity distributions. Climate sensitivity has shown to be inherently difficult to determine so far.

The uncertainty about climate sensitivity poses a serious problem for strong sustainability. The primary concern to stay below the guard rail was based on the assumption that this is possible under any circumstance. However, the probability density distributions have infinite support (long tails), which implies that no climate target can be met with certainty. For any temperature level, there is at least a small chance that it will be transgressed due to past emissions alone. Strong sustainability indeed appears too strong.

However, there has been an attempt to extend the idea of strong sustainability to a probabilistic setting. The temperature guard rail is replaced by a maximum probability to transgress it. This is suggested by Baumgärtner and Quaas (2009) as a general approach for strong sustainability under uncertainty. In climate economics, it has been introduced by den Elzen and Van Vuuren (2007) and Held et al. (2009) who developed probabilistic cost-effectiveness analysis. The climate target can be reformulated as to maintain a minimum probability to stay within the temperature guard rail. Henceforth, this will be referred to as a *probabilistic climate target*.

A probabilistic climate target raises the question of a maximum acceptable level of exceedance probability. It is a second normative parameter in addition to the temperature guard rail and there seems no obvious way to set it. A suggestion has been made on the basis of the agreement at the 17th Conference of the Parties (UNFCCC, 2012). The Parties accepted the objective to "have a *likely* chance of holding the increase in global average temperature below 2°C [author's emphasis]." Along the lines of IPCC language, Neubersch et al. (2014) interpret "likely" as a two thirds probability. Nevertheless, even if this policy reference is considered too weak and contentious, increasing the number of arguable modeling choices by one may still be manageable for policy advice as results can be presented for different parameter values.

Finally, increasing future knowledge about climate sensitivity adds another level of complexity to the decision problem. Webster et al. (2008) estimate that the uncertainty about climate sensitivity can be reduced by 20-40% from climate observations in the next two to five decades. Moreover, advances in the conceptual understanding of crucial physical processes as for example in clouds may grant more knowledge (IPCC, 2013a, pp. 593-594). Taking future learning into account, we obtain a multi-stage decision problem. Now, there are at least two decisions to make, one before and one after a learning event. The event provides the decision maker with a updated probability distribution of climate sensitivity from Bayesian learning. Hence, both decisions are based on a different state of knowledge. To consistently anticipate and adapt our choices in the light of new information, an adequate target-based decision criterion must be able to integrate future learning.

Altogether, the previous considerations were meant to give some background to the question asked by this study:

How can strong sustainability be formalized in a consistent decision criterion for the climate problem under uncertainty and future learning about climate sensitivity?

The following section introduces the two existing decision criteria which will be investigated more thoroughly in chapters 3 and 4.

1.5 Probabilistic Cost-Effectiveness and Cost-Risk Analysis

Two decision criteria have been proposed for implementing the target-based framework under uncertainty and learning: probabilistic cost-effectiveness analysis (CEA)[10] and cost-risk analysis (CRA). CEA without learning has been introduced by den Elzen and Van Vuuren (2007) and Held et al. (2009). It can be formulated as

$$\begin{aligned} &\mathrm{Min}_E \quad C(E), \\ &\text{s.t.} \quad R_{ex}(E|p(\theta)) \leq R_g. \end{aligned} \tag{3}$$

As usual, $E = E(t)$ denotes the emission pathway over time. For notational convenience, the time integral as in (2) is dropped and $C(E)$ refers to the discounted time-aggregated mitigation cost. Climate sensitivity θ, the temperature response to emissions, is only known up to the probability distribution $p(\theta)$. Moreover, $R_{ex}(E|p(\theta))$ is the probability of exceeding the temperature guard rail T_g under emissions pathway E and probability distribution $p(\theta)$. It is calculated by

$$R_{ex}(E|p(\theta)) = \mathbb{E}_{\theta|p(\theta)}[\Theta(T_{\max}(E,\theta) - T_g)]. \tag{4}$$

Here, $T_{max}(E,\theta)$ is the maximum temperature resulting from the emission path-

[10] The literature has often used the acronym 'CEA' for deterministic cost-effectiveness analysis as introduced in section 1.3. Due to the focus of this study, we follow Schmidt et al. (2011) and refer to probabilistic cost-effectiveness analysis as CEA.

way E under climate sensitivity θ. Thereby, 'temperature' always refers to global mean temperature above preindustrial. The notation $\Theta(.)$ represents the Heaviside function which is one for positive arguments and zero otherwise. The probability of exceeding T_g is given by the expectation $\mathbb{E}_{\theta|p(\theta)}$ of $\Theta[T_{max}(E,\theta) - T_g]$ over climate sensitivity θ distributed by $p(\theta)$. Hence, the decision criterion implied by (3) chooses minimal mitigation cost as long as exceedance probability is not higher than the risk guard rail R_g.

To integrate learning about climate sensitivity, we consider the following two-stage decision problem. The emission pathway is split up into emissions before and after learning (E_0 and E_m): First, the decision maker decides for emissions E_0 before learning. After learning and obtaining one of n messages $m \in \{1, ...n\}$, she updates her prior distribution $p(\gamma)$ to the posterior distribution $p_m(\gamma)$. Based on these information, she decides on the emissions E_m after learning. The n-dimensional vector $\boldsymbol{E} = (E_1, ..., E_n)$ represents the set of emission pathways after learning conditional on the message obtained.

CEA is extended to model learning by Schmidt et al. (2011) in the following way:

$$\begin{aligned} &\text{Min}_{(E_0,\boldsymbol{E})} \quad \mathbb{E}_m[C(E_0, E_m)] \\ &\text{s.t.} \quad \forall m: \quad R_{ex}(E_0, E_m|p_m(\theta)) \leq R_g. \end{aligned} \tag{5}$$

Now, the decision maker chooses an emission plan $(E_0, \boldsymbol{E})$. It specifies the emission pathway before learning E_0 as well as the emission pathway after learning (E_m) conditional on the posterior distribution $p_m(\theta)$. The mitigation cost and exceedance probability depend on both decisions E_0 and E_m. Each learning scenario m occurs with a respective likelihood π_m and the expectation over all learning scenarios is $\mathbb{E}_m[.] := \sum_m \pi_m(.)$. Thus, CEA solves for the emission plan with the least expected mitigation cost that keeps the exceedance probability $R_{ex}(E_0, E_m|p_m(\theta))$ below R_g in all learning scenarios. In other words, the probability threshold criterion is applied to all possible posterior distributions of climate sensitivity.

However, it turns out that this type of CEA may be infeasible under learning and can have a negative expected value of information (see chapter 3). This is why Schmidt et al. (2011) develop CRA, an alternative target-based criterion within the expected utility framework. Following the formulation of Neubersch et al. (2014), CRA can be formulated under learning as

$$\mathrm{Min}_{(E_0,\boldsymbol{E})} \quad \mathbb{E}_m[C(E_0,E_m)+\beta R_{DY}(E_0,E_m|p_m(\theta))]. \tag{6}$$

Here, R_{DY} represents the time-discounted expected "degree years", another measure of climate risk introduced by Schneider and Mastrandrea (2005). Degree years quantify the overshoot of a temperature guard rail. They represent the area between the temperature trajectory over time $T(t)$ and the temperature guard rail T_g. The measure is given by

$$\begin{aligned}&R_{DY}(E_0,E_m|p_m(\theta)) = \\ &\int \mathbb{E}_{\theta|p_m(\theta)}[\Theta(T(t)(E_0,E_m,\theta)-T_g)(T(t)(E_0,E_m,\theta)-T_g)e^{-\rho t}]dt.\end{aligned} \tag{7}$$

As the mitigation cost $C(E_0,E_m)$, this risk measure is obtained by a discounted time-aggregation. Unlike exceedance probability, it also takes the magnitude and the duration of the potential guard rail overshoot into account.

However, the main reason not to use exceedance probability R_{ex} as the risk function in CRA is that it may give results which are obviously not in the sense of strong sustainability. Anticipated by Schmidt et al. (2009), Neubersch et al. (2014) find such CRA to suggest that emissions should be strongly increased in the "bad" learning scenarios of high climate sensitivity. This is because if exceedance is almost certain, the risk can hardly increase anymore with additional emissions. Hence, mitigation cost savings become more attractive than risk reductions. Formally, they argue that the risk function must be convex, i.e. at least linear, in temperature to avoid such behavior in high learning scenarios.

CRA minimizes a weighted sum of mitigation cost and climate risk. As an unconstrained optimization, it mimics cost-benefit analysis. However, it is not based on classical damage functions, but rather on the willingness to prevent temperature exceeding a normatively predefined guard rail. Obviously, the optimum depends on the trade-off parameter β between cost and risk. The parameter is calibrated such that without learning the optimum is just at the level of the probabilistic climate target. Chapter 4 will discuss this calibration procedure in more detail.

Finally, we introduce some generalization of terminology. Motivated by the distinction made by strong sustainability, this study focuses on investigating different ways to relate two kinds of values to each other: mitigation cost and climate risk. First, mitigation cost refer to welfare losses from emission reductions relative to a BAU scenario. Second, climate risk can be understood as the risk of "intolerable consequences" looming beyond the guard rail that the WBGU speaks of. It should qualitatively not be confused with the expected damages from a detailed impact assessment in the cost-benefit framing. We may say that CEA and CRA use different measures of climate risk: exceedance probability R_{ex} and expected degree years R_{DY}. Generally, more convex functions of temperature overshoot $(T(t) - T_g)$ than the linear form in R_{DY} can be thought of. Large part of the analysis however will not need these distinctions such that the general term *climate risk* shall refer to either of them.

* * *

This chapter sketched the background to the question examined by this study. The two normative framings used in decision analysis of the climate problem are the cost-benefit approach and the climate target approach. The latter is most prominently advocated by the WBGU who understand the target as a maximum acceptable limit in the sense of *strong sustainability.* It implies to structure the problem into a primary climate criterion and a secondary cost criterion. However, target-based criteria need to take uncertainty about climate sensitivity into account. The question arises how strong sustainability can be formalized in a decision criterion to consistently integrate uncertainty and anticipated future learning about climate sensitivity. So far, probabilistic cost-effectiveness and cost-risk analysis have been suggested which will be investigated more thoroughly in chapter 3 and 4.

2 The Standard Framework: Expected Utility Theory

This chapter presents the standard framework for decision-making under uncertainty: expected utility theory. Understanding its basis will be helpful for analyzing the properties of different decision criteria, first of all CEA and CRA in chapters 3 and 4. Since the foundational work by Von Neumann and Morgenstern (1944), expected utility theory has become the most influential framework against which other theories of decision under uncertainty have to be benchmarked (IPCC, 2014b, p. 168). It provides the decision maker with a strong set of consistency principles to guide her choice. However, as will become apparent, strong sustainability may suggest to drop expected utility theory. Developing the von Neumann-Morgenstern Theorem will help to see possible chances and problems of decision criteria outside expected utility theory.

First, basic terminology of decision analysis is introduced and the climate problem with learning is presented in a simple guiding model that will be used for illustration throughout this study (section 2.1). Second, the von Neumann-Morgenstern axioms and the corresponding representation theorem are presented (section 2.2). Finally, it will be shown that CRA is an expected utility criterion, while CEA violates several of the von Neumann-Morgenstern axioms (section 2.3).

2.1 Concepts of Decision Theory

In decision analysis, choosing between options with different possible outcomes is considered a choice on *lotteries*. A finite lottery L consists of n possible outcomes $(X_1, ..., X_n)$, each occurring with a certain probability $(p_1, ..., p_n)$, where $\sum p_i = 1$. The decision problem tackled by CEA and CRA can be formalized accordingly. Every emission pathway is associated with a certain outcome of mitigation cost and corresponds to a lottery on different outcomes of maximum temperature determined by the probability distribution of climate sensitivity.

In the theory of choice under uncertainty, decision criteria such as CEA or CRA can be formalized in terms of preference relations over lotteries. By a preference relation, or simply ‘preferences’, we mean a set of statements, indicating which lottery the decision maker would choose from pairwise offers. For two lotteries L_1 and L_2, $L_1 \succ L_2$ expresses that the decision maker prefers L_1 over L_2, while $L_1 \sim L_2$ expresses that she is indifferent between the two. The weak preference relation $L_1 \succeq L_2$ expresses that the decision maker's preferences are either $L_1 \succ L_2$ or $L_1 \sim L_2$.

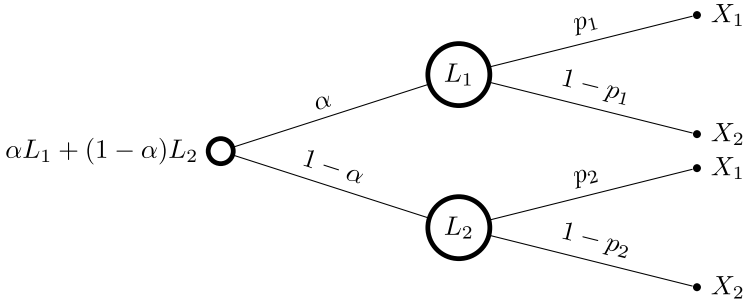


Figure 1: *The compound lottery $\alpha L_1 + (1-\alpha)L_2$ is a lottery that gives the lottery L_1 with a probability α and L_2 with a probability $(1-\alpha)$. The lotteries L_1, L_2 have the probabilities p_1, p_2 of outcome X_1 to occur from a set $\{X_1, X_2\}$.*

To include learning about climate sensitivity, we need to introduce the concept of compound lotteries. A compound lottery is a lottery whose outcomes are again lotteries. The lottery tree in Figure 1 shows the compound lottery $\alpha L_1 + (1-\alpha)L_2$. The notation implies that the decision maker has a probability α of obtaining lottery L_1 and a probability $(1-\alpha)$ of obtaining lottery L_2. In the example, L_1 and L_2 are defined as lotteries on the same outcomes X_1 and X_2 with different probabilities p_1 and p_2. Strictly speaking, outcomes can be considered as lotteries with certainty and the above notation can be used for outcomes, too. For instance, the lottery L_1 could be written as $p_1X_1+(1-p_1)X_2$.

Generally, decision makers accept the axiom of reduction of compound lotteries. It allows to reduce compound lotteries to simple lotteries using Bayes' Law. The probability of an outcome is calculated by multiplying the probabilities to reach it along each path in the lottery tree and summing over all possible paths. Thus, we could rewrite the lottery in Figure 1 as $[\alpha p_1 + (1-\alpha)p_2]X_1 + [\alpha(1-p_1) + (1-\alpha)(1-p_2)]X_2$. In the following, we will always assume the axiom of reduction to hold.

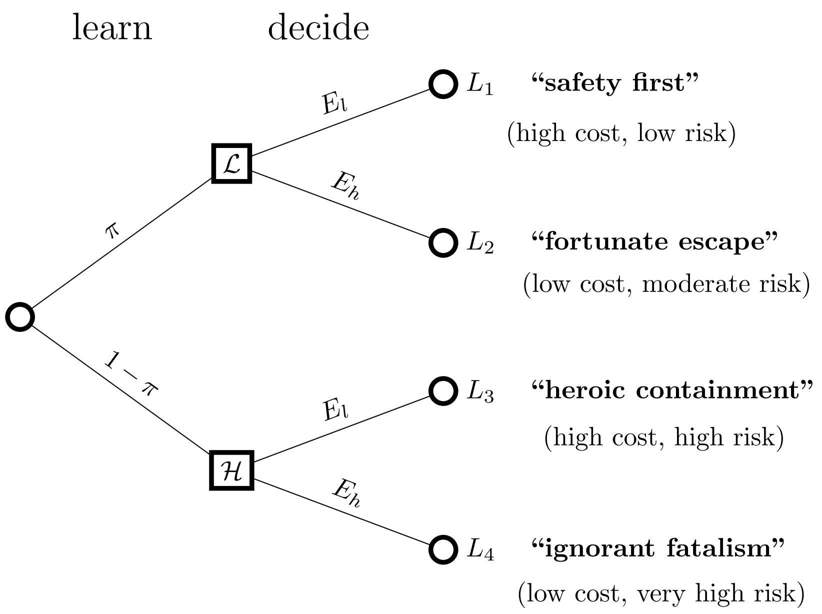

learn
decide
π
$1-\pi$
$\mathcal{L}$
$\mathcal{H}$
E_l
E_h
E_l
E_h
L_1 **“safety first”**
(high cost, low risk)
L_2 **“fortunate escape”**
(low cost, moderate risk)
L_3 **“heroic containment”**
(high cost, high risk)
L_4 **“ignorant fatalism”**
(low cost, very high risk)

Figure 2: *The “guiding model” for decision-making on the climate problem under learning: Learning reveals that climate sensitivity is either “low” ($\mathcal{L}$) or “high” ($\mathcal{H}$). In each learning scenario low emissions (E_l) or high emissions (E_h) can be chosen. This can give four different climate lotteries (L_1 to L_4) associated with different combinations of mitigation cost and (remaining) climate risk. In the decision tree, squares denotes decision nodes, while circles denote lottery nodes.*

Now, let us present the climate problem under learning using the illustrative example in Figure 2. For the moment, we only consider the anticipated decisions after learning. The question is how to choose emissions once there is more knowledge about the response of the climate system. We follow a common notation, depicting lottery nodes by circles and decision nodes by squares. Imagine we anticipate that at a future point in time we will have learned from new climate observations whether climate sensitivity is either “low” ($\mathcal{L}$) or “high” ($\mathcal{H}$). In each learning scenario we will have to choose between subsequent low emissions (E_l) or high emissions (E_h). We assume partial learning such that some uncertainty about the eventual climate outcome still remains. Today, we do not know yet which of the two learning scenarios will occur. Nonetheless, we know that we have a probability of π to learn $\mathcal{L}$ and $(1-\pi)$ to learn $\mathcal{H}$. Thus, it is possible to prepare our choice and plan for emissions conditional on what is learned.

There are four possible climate lotteries we could end up with: First, suppose we learned that climate sensitivity is low ($\mathcal{L}$). Either, we still reduce emissions (L_1) to make sure that the residual risk of transgressing the 2°C is as small as possible ("safety first"). Or we save the money of an ambitious energy transformation and invest it into other fields (L_2). Here, we would accept a moderate climate risk maybe about as high as it had been for low emissions without learning ("fortunate escape"). However, it could also turn out that we have underestimated the magnitude of climate change and we receive the information that climate sensitivity is high ($\mathcal{H}$). Now we know that no matter what we do, we will likely face "intolerable consequences" from climate change. We could either try to keep the inevitable damage as small as possible (L_3) by devoting much effort into emission reduction ("heroic containment"). Or we give up climate mitigation (L_4) and instead use the funds for adaptation or other purposes ("ignorant fatalism").

Our framework requires to define preferences over climate lotteries such as the L_1,...,L_4 in Figure 2. In general, we define a simple climate lottery as a probability distribution of possible temperature trajectories. Choosing the climate lottery $L_{E(t),p(\theta)}$ can be interpreted as pursuing the emission pathway $E(t)$ under a probability distribution of climate sensitivity $p(\theta)$. In each leaning scenario, only climate lotteries with the corresponding climate sensitivity distribution of this scenario can be selected.

Now, decisions under learning can be modeled as a choice on compound climate lotteries. For example, suppose in Figure 2 we decide before learning on the emission plan of choosing E_h if $\mathcal{L}$ and E_l if $\mathcal{H}$. Then, we effectively choose the compound climate lottery $\pi L_2 + (1-\pi)L_3$[11]. Any emission plan we may choose corresponds to a compound climate lottery.

As stylized as this model may be, we will refer to it throughout this study for illustrating some more abstract considerations. It may provide helpful intuitions to put relevant properties of decision criteria into context. Four such properties will be introduced in the following: the *von Neumann-Morgenstern axioms*.

[11] Strictly speaking, identifying the decision problem depicted in Figure 2 as a choice on compound lotteries as shown in Figure 1 presupposes some consistency principles to be discussed in section 3.2.

2.2 The von Neumann-Morgenstern Theorem

The consistency requirements that expected utility theory based on Von Neumann and Morgenstern (1944) makes towards the preferences over lotteries are the *von Neumann-Morgenstern axioms*. The axioms are completeness, transitivity, continuity and independence. We present them following the formulation of Gollier (2001, pp. 4-6).

Completeness: *Preferences $\succeq$ on the lottery space $\boldsymbol{L}$ are such that for any two lotteries $L_1, L_2 \in \boldsymbol{L}$ it is either $L_1 \succ L_2$, $L_1 \prec L_2$ or $L_1 \sim L_2$.*

Completeness demands from the decision maker to be able to state for any pair of lotteries whether she prefers one over the other or whether she is indifferent between the two. There is no third category.

Transitivity: *Preferences $\succeq$ on the lottery space $\boldsymbol{L}$ are such that for any $L_1, L_2, L_3 \in \boldsymbol{L}$ with $L_2 \succeq L_1$ and $L_3 \succeq L_2$, it is $L_3 \succeq L_1$.*

Transitivity requires preferences to be consistent over a triple. A decision maker with complete and intransitive preferences can be "money pumped" (e.g. Mandler, 2005). Transitivity is indispensable for as to avoid preference cycles.

Completeness and transitivity are the standard minimum requirements. Yet, having well-defined preferences on the whole option space may be difficult if the lotteries are hard to compare and feature incommensurable values. The decision maker may feel incapable of finding any argument for why to prefer one or the other or be indifferent between the two. It has been argued that completeness is too demanding when we face moral dilemmas or uncertainty close to ignorance (Gilboa et al., 2009). This may apply to the climate problem, too. For instance, as it remains difficult to draw clear distinctions between scenarios of strong climate change with potentially disastrous domino effects (see section 1.4), it could be argued that there is no basis at all for making choices between the scenarios L_3 and L_4 of high warming in Figure 2.

Yet, accepting the existence of incomparable options points to the limits of what decision analysis can do. There is no way to guide a decision maker on the basis of ambiguous or undefined preferences. The requirement we make here towards preferences on the climate problem is that there always exists a unique optimal choice in each learning scenario, i.e. a lottery L^* such that for all other available lotteries at this node $L \neq L^*$ it is $L^* \succ L$. Strictly speaking, this does not make completeness necessary, however there is no reason why under this requirement incompleteness should make any difference. Whether or not preferences over non-optimal lotteries are defined, does not affect the optimal choice.

Continuity: *Preferences $\succeq$ on the lottery space $\boldsymbol{L}$ are such that for any $L_1, L_2, L_3 \in \boldsymbol{L}$ with $L_3 \succeq L_2 \succeq L_1$ there exists a probability $p \in [0,1]$ such that: $pL_1 + (1-p)L_3 \sim L_2$.*

Here, the decision maker accepts that there is always a (prior) probability p for which she would be indifferent between receiving a compound lottery on a "good" (L_3) and a "bad" lottery (L_1), $pL_1 + (1-p)L_3$, and receiving an intermediate lottery L_2. Continuity implies that the probability of obtaining either L_1 or L_3 always makes a *gradual* difference to the decision maker. It can be interpreted by analogy with the continuity in mathematics only for a mapping from probability to a preference ranking: A small change in the probability of obtaining some good or bad outcome should only translate to a small change in the preference ranking of the lotteries, i.e. it should make the decision maker only slightly better or slightly worse-off. In other words, the ranking of two lotteries with a small difference in probabilities should not be too different.

This already indicates why CEA violates continuity. Suppose the primary criterion asserts that any probability for transgressing the 2°C level of 50% or lower is preferred, while the secondary criterion implies to minimize mitigation cost. Then, exchanging an option with a 50% probability for an option with only a slightly higher probability makes an enormous difference as the primary criterion is triggered. It exchanges the optimal choice (50%) for a choice ($> 50\%$) which is worse than any choice with a chance below 50%. Hence, the probability threshold criterion as in CEA is at odds with continuity. This is will be shown formally in section 2.3.

Discussing the normative justification of continuity is somewhat difficult. As Gilboa (2009, pp. 80-81) points out, not only is it impossible to conduct real-world experiments on whether or not people comply with it. Its violation requires infinitely many observations. Moreover, the infinite number of options also challenges normative argument. He suggests to engage in thought experiments such as whether one would still go for some small benefit if the action requires risking one's life with a very low probability. An example would be crossing a frequented street for a time gain of only a few seconds. Yet, although Gilboa concedes that the framing of such thought experiments is often crucial for the choice, he concludes that for most applications we readily accept continuity.

It is possible to challenge continuity for a choice between a 100%-certain and an uncertain option. Absolute certainty could make a qualitative difference. However, since the probability distribution of climate sensitivity has a long tail, no temperature guard rail can be met with certainty. We need to live with a small chance of disastrous climate change. Eventually, setting a specific (non-zero) level of probability as the maximum acceptable risk appears somewhat arbitrary and artificial. Probabilities within the open interval $(0, 1)$ seem continuous as real numbers. A probabilistic climate target as introduced in section 1.4 may be a pragmatic policy move to maintain the idea of a climate target in a probabilistic setting. However, the radical exclusion of any option slightly above the probability threshold is at odds with the consistency intuition behind the continuity axiom.

Independence: *Preferences* $\succeq$ *on the lottery space* $\boldsymbol{L}$ *are such that for any* $L_1, L_2, L_3 \in \boldsymbol{L}$ *and* $p \in [0, 1]$*:* $L_1 \succeq L_2 \Longleftrightarrow pL_1 + (1-p)L_3 \succeq pL_2 + (1-p)L_3$.

Independence asserts that whether the decision maker prefers $L_1 \succ L_2$ should not depend on other possible lotteries that she could receive. The decision maker accepts that comparing $pL_1 + (1-p)L_3$ to $pL_2 + (1-p)L_3$ is nothing but comparing L_1 to L_2, irrespective of p and L_3. It can be interpreted as a condition to consistently extend the choice from simple to compound lotteries and vice versa. As presented in section 2.1, making choices while anticipating future learning corresponds to a decision on compound lotteries. Hence, violating independence has normatively uncomfortable implications under learning: As uncertainty resolves, i.e. the decision maker learns, she may change her mind about a given pair of lotteries.

Let us illustrate this for the climate problem. Suppose prior to learning the decision maker prefers to go for low emissions in both learning scenarios in Figure

2.1, i.e. $\pi L_1 + (1-\pi)L_3 \succ \pi L_2 + (1-\pi)L_3$. Now, violating independence would imply that she can prefer $L_1 \prec L_2$. Hence, once she actually learns that climate sensitivity is low she does not go for the option she initially preferred. The decision maker does not correctly anticipate the choice which raises a problem of time-inconsistency. Section 3.2.2 will discuss this issue more thoroughly and show that by violating independence at least one other standard consistency principle must be violated, too.

Having introduced the axioms, we can turn to the von Neumann-Morgenstern Theorem. First, we define the notion of a *utility representation* of preferences:

Utility Representation: *The function $V : \boldsymbol{L} \rightarrow \mathbb{R}$ is a utility representation of the preferences $\succeq$ over the lottery space $\boldsymbol{L}$ if and only if for all L_1, $L_2 \in \boldsymbol{L}$:*

$$\begin{aligned} V(L_1) > V(L_2) \Leftrightarrow L_1 \succ L_2 \\ V(L_2) = V(L_2) \Leftrightarrow L_1 \sim L_2. \end{aligned} \tag{8}$$

The utility representation is a function that stands for the degree of satisfaction the decision maker has according to her preferences by holding a certain lottery. Such utility function over lotteries is ordinal, i.e. the magnitude of the differences in utility does not have a meaning. It only provides a ranking from the most to the least preferred lottery. For preferences which have a utility representation, the optimal choice can be obtained by finding the maximum of the utility representation. Finally, we can formulate the von Neumann-Morgenstern Theorem following Gollier (2001, p. 7):

von Neumann-Morgenstern Theorem: *If and only if preferences $\succeq$ on the lottery space $\boldsymbol{L}$ satisfy completeness, transitivity, continuity and independence, there exists a utility representation $V : \boldsymbol{L} \rightarrow \mathbb{R}$ of $\succeq$ that is linear in probabilities. That is, for a lottery L with n possible outcomes X_i to occur with probability p_i, where $i = 1, ...n$, there exists a scalar U_i such that*

$$V(L) = \sum p_i U_i. \tag{9}$$

The essence of the theorem is that a decision maker accepting the above axioms is an expected utility maximizer. The U_i can be interpreted as the utility representation over outcomes X_i known from choice under certainty in consumer theory. An expected utility maximizer sums the probability-weighted utilities of all outcomes for each of the lotteries and chooses the lottery with the highest sum.

Let us sketch the proof of (9) for a finite set of outcomes following Gollier (2001, pp. 7-8). This provides some insight into the function each axiom has in expected utility theory: By completeness and transitivity, we can pairwise compare all lotteries in the lottery space $\mathbf{L}$ and arrange them on a scale from 'worst' to 'best'. Let us denote the worst lottery by $\underline{L}$ and the best lottery by $\overline{L}$. Then, by continuity we can find a $\lambda_i \in [0,1]$ for any lottery L_i such that

$$\lambda_i \overline{L} + (1-\lambda_i)\underline{L} \sim L_i. \tag{10}$$

As the probability of an equally preferred compound lottery on $\overline{L}$ and $\underline{L}$, λ_i is a utility representation of L_i. We can define $V(L_i) := \lambda_i$. Thus, completeness, transitivity and continuity are sufficient for the existence of a utility representation of the preferences over lotteries.

Finally, independence requires the utility representation to be linear in probabilities. Using the definition of $V(L)$, the axiom of reduction and independence, it is possible to show (see appendix A.1) that for any $\beta \in [0,1]$ and $L_1, L_2 \in \mathbf{L}$

$$V(\beta L_1 + (1-\beta)L_2) = \beta V(L_1) + (1-\beta)V(L_2). \tag{11}$$

Hence, the utility of the compound lottery needs to be calculated by the probability-weighted sum of the corresponding simple lotteries. In other words, the utility function is linear in the probability vector.

This section presented expected utility theory and its axiomatic basis in the von Neumann-Morgenstern framework. We have seen that only a decision maker whose preferences satisfy completeness, transitivity, continuity and independence is an expected utility maximizer. Before turning to the discussion of

CEA and CRA, we will formally show how both criteria differ in their compliance with the von Neumann-Morgenstern axioms.

2.3 Axiomatic Deviation of CEA

The first step for characterizing CEA and CRA is to check their compliance with the von Neumann-Morgenstern axioms. The case for CRA is straightforward. Without learning, CRA as in (6) can be written as

$$\begin{aligned}&\mathrm{Max}_E\ \ \mathbb{E}_\theta[U(E,\theta)]\\&U := -\int\{C(t)(E)+\beta\Theta[T(t)(E,\theta)-T_g](T(t)(E,\theta)-T_g)\}e^{-\rho t}dt.\end{aligned} \tag{12}$$

In this form, it becomes obvious that CRA is an expected utility maximization. The von Neumann-Morgenstern theorem implies that CRA satisfies completeness, transitivity, continuity and independence.

Schmidt et al. (2009) check the von Neumann-Morgenstern axioms for CEA. Let us present their reasoning here. First, the CEA criterion needs to be formulated in terms of a preference relation on climate lotteries as defined in section 2.1: As usual, we consider n learning scenarios with probability distributions of climate sensitivity $p_1(\theta),...,p_n(\theta)$. Let $\mathbf{L}$ denote the set of simple climate lotteries that can be received with learning. It comprises all possible probability distributions of temperature trajectories that result from any combination of an emission pathway $E(t)$ with one of the distributions $p_1(\theta),...,p_n(\theta)$. According to the functional relations used in section 1.5, any simple climate lottery $L \in \mathbf{L}$ can be associated with a unique pair of mitigation cost $C(L)$ and climate risk $R(L)$.

Now, for choosing between two lotteries $L_1, L_2 \in \mathbf{L}$, CEA applies two decision criteria in lexicographic order. The primary criterion is the probabilistic guard rail criterion $\succ_1$ given by

$$\begin{aligned}&L_1 \succ_1 L_2 \Leftrightarrow \{R(L_1) \le R_g\}\ \wedge\ \{R(L_2) > R_g\}\\&L_1 \sim_1 L_2 \Leftrightarrow \{R(L_1) \le R_g\}\ \wedge\ \{R(L_2) \le R_g\}.\end{aligned} \tag{13}$$

Any option below the guard rail is strictly preferred to any option above the guardrail. If both options are below the guard rail, the primary climate criterion values them equally. However, the criterion is not complete, as no preferences are defined for two lotteries which both transgress the risk guard rail. Now, by the secondary criterion $\succeq_2$ a simple climate lottery is preferred if and only if mitigation cost are smaller, i.e.

$$L_1 \succeq_2 L_2 \Leftrightarrow C(L_1) \leq C(L_2). \tag{14}$$

Now, CEA is a *lexicographic composition* of the two criteria. This implies that the decision maker follows the primary criterion as long as it gives a strict preference. The secondary criterion only applies if the decision maker is indifferent by the first criterion. Accordingly, the CEA criterion $\succeq$ is given by

$$\begin{aligned} L_1 \succ L_2 \; &\Leftrightarrow \{L_1 \succ_1 L_2\} \vee \{(L_1 \sim_1 L_2) \;\wedge\; (L_1 \succ_2 L_2)\} \\ L_1 \sim L_2 \; &\Leftrightarrow \{L_1 \sim_1 L_2\} \;\wedge\; \{L_1 \sim_2 L_2\}. \end{aligned} \tag{15}$$

After having formulated CEA as a preference relation over the set of simple climate lotteries $\mathbf{L}$, we can check the axioms by definition. Completeness and transitivity are straightforward: The primary criterion is incomplete because preferences are not defined over two lotteries that both transgress the risk guard rail. This implies that CEA itself is incomplete. Transitivity is satisfied by CEA as both the primary and the secondary criterion are transitive.

Now, let us turn to the continuity axiom. For this, we need to specify how the CEA criterion applies to compound lotteries. The climate risk can be obtained by the axiom of reduction:

$$R(pL_1 + (1-p)L_3) = pR(L_1) + (1-p)R(L_3) \tag{16}$$

Continuity would be satisfied if for any $L_1, L_2, L_3 \in \mathbf{L}$ we had

$$L_3 \succeq L_2 \succeq L_1 \Rightarrow \exists p \in [0,1] : pL_1 + (1-p)L_3 \sim L_2. \tag{17}$$

However, it is possible to construct counterexamples for the CEA criterion.

Let us consider $L_1, L_2, L_3 \in \mathbf{L}$ with $R(L_2) < R(L_3) = R_g < R(L_1)$ and $C(L_2) > C(L_3)$. If such triple did not exist, there would obviously be no climate problem.

Applying the CEA criterion gives $L_2 \succ L_1$ and $L_3 \succ L_1$ since L_1 transgresses the risk guard rail R_g, while L_2 and L_3 comply with it. Moreover, the secondary criterion implies $L_3 \succ L_2$. Hence, we obtain $L_3 \succ L_2 \succ L_1$. However, any $p \in (0, 1]$ would make $pL_1 + (1-p)L_3$ transgress the risk guard rail R_g as $R(L_3) = R_g$ and $R(L_1) > R_g$ which implies $L_2 \succ pL_1 + (1-p)L_3$. For $p = 0$, we would only recover $L_2 \prec L_3$. Thus, it is not possible to find a probability $p \in [0, 1]$ to make the decision maker indifferent between $pL_1 + (1-p)L_3$ and L_2. This violates the continuity axiom.

Finally, we check the compliance with the independence axiom. It would require for any L_1, L_2, L_3 and $p \in [0, 1]$:

$$L_1 \succeq L_2 \Leftrightarrow pL_1 + (1-p)L_3 \succeq pL_2 + (1-p)L_3. \tag{18}$$

A similar counterexample can be made. Let us consider $L_1, L_2, L_3 \in \mathbf{L}$ with $R(L_2) < R(L_1) < R_g < R(L_3)$ and $C(L_1) < C(L_2)$. For emission pathways which both comply with the guard rail smaller mitigation cost are chosen such that $L_1 \succ L_2$. However, due to $R(L_1) > R(L_2)$ it is possible to find a $p \in (0, 1]$ such that

$$R(pL_1 + (1-p)L_3) > R_g \geq R(pL_2 + (1-p)L_3). \tag{19}$$

Applying the guard rail criterion gives $pL_1 + (1-p)L_3 \prec pL_2 + (1-p)L_3$. Thus, the CEA preferences $\succeq$ violate independence, too.

* * *

In this chapter, we introduced expected utility theory in the von Neumann-Morgenstern framework. We showed that CRA is an expected utility approach, while CEA violates completeness, continuity and independence. In general, each of the four von Neumann-Morgenstern axioms is a desirable consistency principle. However, consistency alone is not sufficient as can be seen with CBA, which also is an expected utility criterion. Approaching the climate problem from the perspective of strong sustainability suggests CEA as a lexicographic decision criterion outside of expected utility theory. After all, the question arises how dispensable the von Neumann-Morgenstern axioms are in this context. Analyzing the properties of CEA and CRA in the following two chapters will reveal the implications of the axioms and allow to discuss whether the criteria are adequate formalizations of strong sustainability.

3 The Limits of CEA: Problems of a Probabilistic Target under Learning

In this chapter, we will discuss whether CEA can adequately formalize strong sustainability under uncertainty and learning. Under learning, CEA still applies a primary climate criterion and a secondary mitigation cost criterion. The former implies that the risk guard rail needs to be held in all learning scenarios, while the latter aims at minimizing cost. It extends the general idea of strong sustainability under uncertainty to decisions under learning. Yet, as this chapter will show, there are some normatively unappealing consistency problems of CEA and, more generally, of probabilistic constraints if learning is taken into account.

First, depending on how “much” is learned, CEA may become infeasible as it cannot deal with overshoots of the (probabilistic) climate target (section 3.1). Second, CEA can have a negative expected value of information (EVOI), a phenomenon that we will explore in three steps (section 3.2): The negative EVOI emerges due to the posterior probabilistic constraint in CEA that leads to the violation of the independence axiom (section 3.2.1). In general, dropping this axiom implies to relax at least one of three standard consistency requirements of dynamic choice (section 3.2.2). However, an actual negative EVOI only occurs in CEA if the mitigation cost function is sufficiently convex (section 3.2.3). Both issues, the infeasibility and the negative EVOI, point to conceptual problems of probabilistic constraints under learning.

3.1 Infeasibility and the Probabilistic Contradiction

CEA works only as long as the probabilistic constraint can be met. However, given past emissions, this may be impossible for any emission pathway under certain combinations of temperature and risk guard rails. On the axiomatic level, the infeasibility problem is reflected by the incompleteness of the CEA preference relation as defined in (15). CEA cannot deal with overshoots of the (probabilistic) climate target. Likewise, CEA can be thought of as a constrained expected utility maximization whose constraint may leave an empty option space.

Let us illustrate the feasibility limits of CEA with some numbers from the Model of Investment and Technological Development (MIND) (Edenhofer et al., 2005).

Schmidt et al. (2009) find that the 2°C target can be met in MIND without learning for climate sensitivities up to 6.4°C. This corresponds to the 92%-ile of their prior distribution. Restricting the maximum rate of emission reduction to 13.3% annually as done in Lorenz et al. (2012) and Roth et al. (2015), the model cannot reach the 2°C target for climate sensitivities above 3.5°C. Finally, under learning feasibility depends on the posterior distributions and their numerical sampling since a posterior constraint requires to meet the probabilistic target in all learning scenarios.

While for an analysis without learning the feasibility of CEA can be guaranteed by choosing an achievable target, any target can be out of reach if learning is included. Consider the extreme case of perfect learning, i.e. if the exact value of climate sensitivity was revealed. Here, CEA is obviously infeasible for any guard rail as long as the prior distribution has infinite support. Meeting the target in all learning scenarios would require meeting it for any climate sensitivity, which makes it effectively equivalent to deterministic CEA. Considering partial learning, feasibility depends not only on how "much" is expected to be learned. Above all, the numerical sampling of the climate sensitivity values has a crucial influence because CEA requires compliance in all learning scenarios. As pointed out by Schmidt et al. (2009), feasibility of CEA under learning is a "numerical artefact".

The potential infeasibility is a symptom of a more conceptual problem inherent to CEA under learning. In fact, requiring to meet a probabilistic climate target in all possible learning scenarios contradicts the very idea of such target. To see this, let us put aside the infeasibility problem and suppose that it was actually possible to meet the 2°C target with certainty. Now, for some reason, e.g. because 100%-compliance would demand very high cost, the decision maker still favors a probabilistic climate target. Suppose she would like to aim for a chance $R_g = 80\%$ of meeting the 2°C target. By making this statement, she obviously accepts the possibility of missing the 2°C target. However, applying CEA under perfect learning would make her meet the 2°C target with certainty as the risk constraint needs to hold in all learning scenarios. The probabilistic target has effectively become a deterministic target. The posterior constraint of CEA under learning contradicts the very probabilistic basis on which it was developed in the first place.

To resolve the problem, it must be clarified what exactly is meant by the concept of a probabilistic climate target. The above argument shows that if the probabilistic target is referred to any state of knowledge, the specific probability threshold R_g becomes irrelevant. Thus, it makes more sense to consider the probabilistic climate target as the statement that given the current state of knowledge the decision maker aims to meet the temperature target T_g with a certain probability R_g. This conception implies that a probabilistic climate target needs to be defined with respect to some probability distribution of climate sensitivity. This would allow for overshoots of the target in certain states of knowledge. The overshoots would need to be determined on the basis of the probabilistic climate target and the state of knowledge it refers to. As we will see in section 4.1, this is the very approach a calibrated CRA makes. Before, we will point to a second problem raised by a posterior risk constraint.

3.2 The Expected Value of Information

The motivation to learn and reduce uncertainty in a decision problem is to make better decisions. Quite paradoxically, CEA appears to violate this fundamental principle since it can give a negative expected value of information (EVOI). The EVOI generally refers to the welfare gain due to learning. It can be interpreted as the maximum amount to spend for receiving new information. Schmidt et al. (2009) analyze a scenario of partial future learning about climate sensitivity with CEA in MIND and find a negative EVOI for most risk guard rails on the 2°C target. They show that the problem is conceptually linked to the violation of the independence axiom. Generally, the issue is well known and was already investigated by Blau (1974) and Lavalle (1986) among others[12]. The negative EVOI seems to imply that a decision maker is better-off without new information and must consequently reject costless learning.

Let us first define the expected value of information in general terms. It is the difference between the expected maximum welfare under learning and the maximum welfare without learning:

$$EVOI = \mathbb{E}_m[W^{(l)}(E_0^*, E_m^*, p_m(\theta))] - W^{(nl)}(E^*, p(\theta)). \tag{20}$$

[12] They refer to to it as a “dilemma” of information in “chance-constrained programming” which is their term to denote expected utility maximization subject to a probabilistic constraint.

Here, $W^{(l)}(E_0^*, E_m^*, p(\theta))$ denotes the expectation of the maximum welfare obtained with learning over n learning scenarios $m \in \{1, ..., n\}$. It depends on the optimal emission plan $(E_0^*, \boldsymbol{E_1^*})$, i.e. the emission pathway before learning E_0^* and the vector of optimal emission pathways after learning $\boldsymbol{E_1^*} = (E_1^*, ..., E_n^*)$ conditional on the probability distributions $p_1(\theta), ..., p_n(\theta)$. Correspondingly, $W^{(nl)}(E^*, p(\theta))$ represents the maximum welfare obtained without learning for the optimal emission pathway E^* under the probability distribution $p(\theta)$. The welfare refers to the objective function to be maximized. Hence, for CEA the welfare is given by $W_{\text{CEA}} = -C(E)$, while for CRA it is $W_{\text{CRA}} = -[C(E) + \beta R_{DY}(E, p(\theta))]$.

We will analyze the problem of a negative EVOI in CEA in three steps. First, based on a simple example, we will illustrate why it is possible that CEA exhibits a negative EVOI and draw the link to the violation of the independence axiom. Second, we will discuss in general whether dropping this axiom can be defended in the case of the climate problem by relaxing other, hitherto implicit, consistency principles. Third, we will point to three factors that constitute the EVOI of CEA in the full climate problem and show why it can generally be positive or negative.

3.2.1 Information Aversion

The problem of negative EVOI in CEA can be understood by analyzing the option space of emission plans admissible under the CEA risk constraint. The sign of the EVOI depends on how the risk constraint restrains this admissible option space with and without learning respectively. If the option space is not restricted due to learning, the optimal emission pathway for no learning E^* can still be chosen in every learning scenario. Now, as there is no uncertainty in the mitigation cost function, i.e. the objective function of CEA, this will still give the same amount of maximum (expected) welfare. The EVOI is at least zero.

There are two effects on the admissible option space comparing learning to no learning in CEA. First, learning always adds options to the option space since the decision maker can adapt her choice in the light of new information. This we will refer to as the *enlarging effect* of learning. If the enlarging effect adds an option with a higher expected welfare than in the no-learning optimum, the EVOI is strictly positive. However, as we will show, the specific way of extending the risk constraint in CEA under learning induces a second *restrictive effect* on the admissible option space. In fact, CEA exhibits a negative EVOI if and only if this restrictive effect dominates the enlarging effect, i.e. if the optimal choice without learning is not available anymore with learning and no at least equally good option is added.

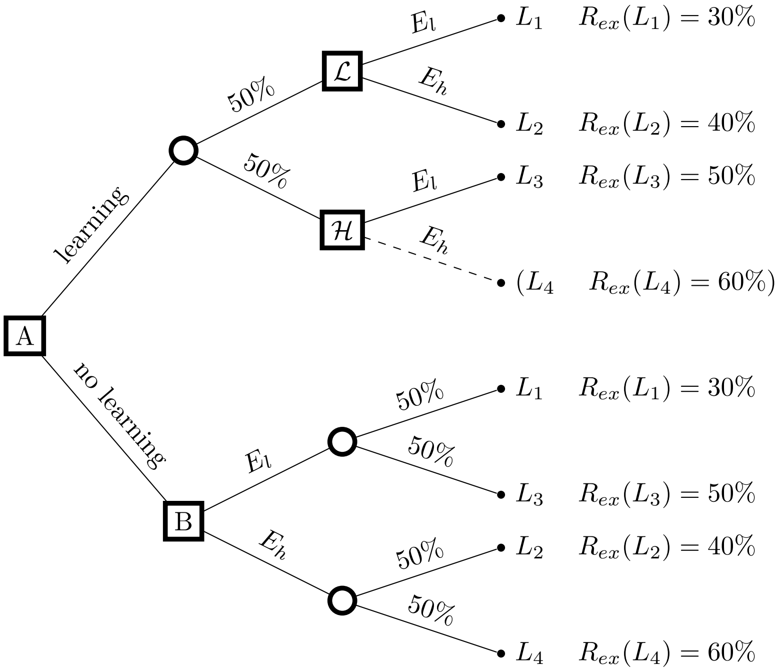

A
learning
no learning
50%
50%
$\mathcal{L}$
$\mathcal{H}$
E_l
E_h
E_l
E_h
L_1 $R_{ex}(L_1) = 30\%$
L_2 $R_{ex}(L_2) = 40\%$
L_3 $R_{ex}(L_3) = 50\%$
$(L_4$ $R_{ex}(L_4) = 60\%)$
B
E_l
E_h
50%
50%
50%
50%
L_1 $R_{ex}(L_1) = 30\%$
L_3 $R_{ex}(L_3) = 50\%$
L_2 $R_{ex}(L_2) = 40\%$
L_4 $R_{ex}(L_4) = 60\%$

Figure 3: *Illustration of how a negative EVOI emerges in CEA: A decision maker with CEA preferences as in (15) with respect to the risk guard rail $R_g = 50\%$ can choose at A whether or not to learn about climate sensitivity before deciding on low emissions (E_l) or high emissions (E_h). The information aversion, i.e. preferring not to learn, arises since the decision maker satisfies the risk constraint by choosing E_h for no learning, while under learning the posterior constraint requires her to choose E_l at $\mathcal{H}$. The restrictive effect of the posterior risk constraint eliminates the cost-minimal option $1/2L_2 + 1/2L_4$ (brackets), which is available to the decision maker only without learning.*

First, let us understand how a negative EVOI can occur in CEA using the guiding model from section 2.1. In the decision problem of Figure 3 inspired by Wakker (1988), we may choose between learning and no learning. Choosing to learn, we face the very situation known from Figure 2. We will first receive the information whether climate sensitivity is “low” ($\mathcal{L}$) or “high” ($\mathcal{H}$) and can subsequently decide for low emissions (E_l) or high emissions (E_h). As before, this will give us one of the climate lotteries $L_1, ..., L_4$. Now, each of them is associated with a specific exceedance probability: $R_{ex}(L_1) = 30\%, R_{ex}(L_2) = 40\%, R_{ex}(L_3) = 50\%, R_{ex}(L_4) = 60\%$. The mitigation cost decrease with emissions such that $C(L_1) = C(L_3) > C(L_2) = C(L_4)$. Finally, the likelihood of obtaining either $\mathcal{L}$ or $\mathcal{H}$ is 50%. Without learning, we face the same situation only that we decide on emissions first, while the state of the world, either $\mathcal{L}$ or $\mathcal{H}$, is revealed afterwards.

Now, imagine we apply CEA at node A in Figure 3 with respect to the risk guard rail $R_g = 50\%$. Remember that in the framework introduced in section 2.1 emission plans, i.e. sets of emission pathways conditional on the learning scenario, correspond to compound climate lotteries. From the viewpoint of A, the choice of both branches, learning and no learning, can be characterized by a different set of compound lotteries on $L_1, ..., L_4$. The decision at A can be determined by comparing these two option spaces for learning and no learning.

Let us find the enlarging and the restrictive effect on the admissible option space in this simple example. In the no learning case, we have the choice between $1/2L_1 + 1/2L_3$ and $1/2L_2 + 1/2L_4$. First of all, with learning our option space is enlarged as we may pick the compound lottery $1/2L_2 + 1/2L_3$ which is not available without learning. The enlarging effect is because we can make the choice about E_l or E_h conditional on whether we find ourselves at $\mathcal{L}$ or at $\mathcal{H}$. Formally, no learning is the special case of learning where only emission plans with the same emission pathway for every learning scenario can be chosen.

However, there is also a restrictive effect from learning since the risk constraint in CEA is imposed *posterior*. This means that in each of the learning scenarios $\mathcal{L}$ and $\mathcal{H}$ respectively the exceedance risk must not be higher than 50%. Consequently, with learning the option of high emissions E_h at $\mathcal{H}$ has been removed (dotted line) and the scenario L_4 cannot occur anymore. The compound lottery $1/2L_2 + 1/2L_4$ available for no learning cannot be chosen with learning. This is because the posterior risk constraint (limiting the risk in each learning scenario) is strictly stronger than a prior risk constraint (limiting the expected risk over all learning scenarios).

Finally, our example is chosen such that the restrictive effect of the constraint dominates the enlarging effect of learning. Applying CEA, the decision maker chooses the least mitigation cost from a constrained option space. The point is that the compound lottery $1/2L_2 + 1/2L_4$ which incurs the least mitigation cost $C(E_l)$ is only available without learning. With learning, the optimal choice would be $1/2L_2 + 1/2L_3$, which gives the expected mitigation cost $1/2(C(E_l) + C(E_h)) > C(E_l)$. Hence, the EVOI is negative and the decision maker is better-off by choosing the lower branch (no learning) at A. Consequently, she should reject costless learning about climate sensitivity. This is what Wakker (1988) referred to as *information aversion.*

On the level of preferences, we may say that the information aversion in Figure 3 is due to the violation of the independence axiom by CEA preferences as defined and discussed in section 2.3. The problem is that at $\mathcal{H}$ the decision is determined by $L_3 \succ L_4$, while at B it is based on $1/2L_3 + 1/2L_2 \prec 1/2L_4 + 1/2L_2$. The first preference follows from the primary guard rail criterion as $R(L_3) = R_g < R(L_4)$. The second preference follows from comparing the compound lotteries by reduction which gives $R(1/2L_4 + 1/2L_2) = R_g$ and $C(1/2L_4 + 1/2L_2) < C(1/2L_3 + (1 - 1/2)L_2)$. Prior to learning, the high risk of L_4 can be balanced by the low risk of L_2. Yet, after learning this is not possible anymore since the uncertainty about $\mathcal{L}$ or $\mathcal{H}$ has been resolved.

3.2.2 Implications of Non-Independence

We have seen how the information aversion is linked to the violation of the independence axiom. Now, we will discuss the implications of dropping this axiom more generally. Violating independence implies violating at least one of three consistency principles of dynamic choice which have been implicit to our above argument. This will also allow us to see whether the EVOI problem of CEA could be fixed by relaxing these requirements using other non-independent decision criteria.

The point is that identifying the plans made at node A in Figure 3 as compound lotteries presupposes two consistency principles: time-consistency and consequentialism. So far, we have always silently accepted them by referring to anticipated choices after learning as compound lotteries for which we could then check independence using the axiom of reduction. In the following, we will discuss the principles by explaining the argument made by Wakker (1999) for defending the necessity of the independence axiom against the background of the climate problem.

The argument is illustrated in Figure 4. Each of the depicted conditions requires that the decision maker chooses the upper branch at the decision node (square) in one situation if and only if she chooses the upper branch at the decision node in the other situation. Hence, if independence is violated at least one of the other three conditions must be violated, too. The equivalence between the decision problem in (1) and (4) in Figure 4[13]. The necessity of the axiom can be defended by arguing that each of the other three consistency principles should be satisfied.

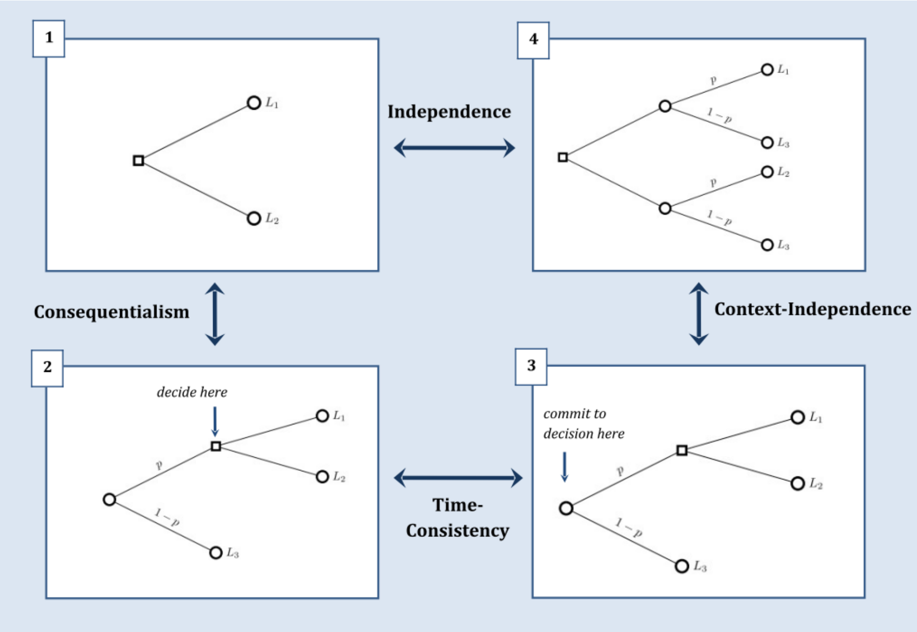


Figure 4: *Relation between independence, consequentialism, time-consistency and context independence following Wakker (1999). Each consistency condition implies that the decision maker chooses the upper branch at the squared decision node in one situation if and only if she chooses the upper branch in the other situation, too.*

First, let us consider consequentialism (Wakker (1999) calls it 'foregone-event independence'). It implies that the decision at some node should not depend

[13] Often the independence axiom is understood in a broader sense to denote the four principles together. Then, consequentialism or time-consistency are considered as reasons to accept or reject the axiom in some decision context (e.g. Wakker, 1988; Bradley and Stefansson, 2016). All four principles are necessary for expected utility theory.

on what could have happened in the past but eventually did not occur. Hence, the same choice must be made between “up” or “down” in (2) and (1). Second, time consistency requires that the decision maker can correctly anticipate her choice beforehand. The best decision should not depend on the point in time at which the decision maker thinks about it. This implies deciding in (2) as in (3). Third, holding (3) there should be no objection to frame this time-consistent plan about a future decision as a prior decision, i.e. a choice on two compound lotteries as in (4). Finally, making the same choice (“up” or “down”) on the simple lotteries in (1) as on the compound lotteries in (4) is the exact form of independence axiom as introduced in section 2.2.

For each condition, there have been suggestions for relaxing it as to allow for non-expected utility models (see references in Wakker, 1999). First, time-inconsistency would imply that our plans before learning may deviate from the actual decision made after learning. We would choose the upper branch at A in Figure 3 only because we do not see it coming that at $\mathcal{H}$ we will not be able to choose E_h. Such reasoning would make the entire consideration of learning pointless. Time-consistency is indispensable.

Second, discarding context-independence must argue that a negative EVOI is actually acceptable. Note that the learning and the no learning branch in Figure 3 correspond to the situations (3) and (4) in Figure 4 respectively. Rejecting costless learning is a violation of context-independence. To defend it, one would need to argue that there is a relevant difference between a time-consistent plan about a future decision and a prior decision. Apart from that the actual decision is made at a different stage, the decision maker faces the same situation in (3) as in (4). It is hard to see any difference here.

As Schmidt et al. (2009) point out, for a problem where the risk is induced by someone who does not necessarily bear the risk, a negative EVOI can be plausible. They give the example of risk thresholds in nuclear power plants imposed by a regulating government. Here, acquiring more information about what may cause an accident can make the company worse-off, while the society can only benefit. The company will obviously ignore such new information until being forced to acknowledge it by updated government regulation. However, the benevolent social planer of the climate problem induces and bears the risk at the same time. Dropping this principle and accepting a negative EVOI would imply to ignore new information. Here, holding deliberately counterfactual beliefs is deeply troubling.

Finally, we are left with the principle that Schmidt et al. (2009) and Gollier (2001, p. 12) call 'consequentialism'. In the example of Figure 3, dropping consequentialism would allow to make the choice on L_1 or L_2 at $\mathcal{L}$ dependent on situations that could have been obtained at earlier stages of the decision tree. For example, replacing the posterior risk constraint by a prior risk constraint would give us a non-consequentialist criterion. It would imply that the decision maker minimizes mitigation cost subject to a maximum limit of the risk induced before learning. We may refer to this as 'Prior-CEA' (as opposed to the standard 'Posterior-CEA') given by

$$\begin{aligned} &\text{Min}_{(E_0,\boldsymbol{E})} \quad \mathbb{E}_m[C(E_0, E_m)] \\ &\text{s.t.} \quad \mathbb{E}_m[R_{ex}(E_0, E_m|p_m(\theta))] \leq R_g \end{aligned} \tag{21}$$

Prior-CEA cannot have a negative EVOI because there is no restrictive effect on the option space anymore (Lavalle, 1986). By choosing the optimal emission pathway without learning E^* in every learning scenario the constraint is met:

$$\sum_{m=1}^{n} \pi_m \int p_m(\theta)\Theta[T(E_0, E_m, \theta) - T_g]d\theta = \int p(\theta)\Theta[T(E^*, \theta) - T_g]d\theta \leq R_g \tag{22}$$

Any option available without learning is still available under learning.

Figure 5 summarizes our considerations about admissible option spaces in CEA with and without learning. First of all, learning always enlarges the option space. For an unconstrained cost minimization, any option without learning (light blue) is still available with learning (white). Imposing the risk constraint removes options. Yet, unlike for Prior-CEA (light red) the option space of Posterior-CEA (orange) does not contain the option space of CEA without learning (dark blue). Hence, the optimal choice without learning may not be available anymore under a posterior risk constraint and the EVOI can become negative.

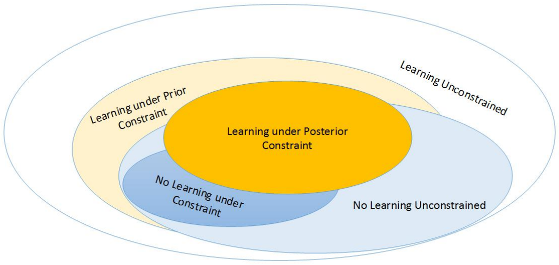

Learning Unconstrained
Learning under Prior Constraint
Learning under Posterior Constraint
No Learning under Constraint
No Learning Unconstrained

Figure 5: *Venn diagramm for illustrating different option spaces of emission plans available for cost minimization in CEA. Generally, learning always enlarges the option space as it allows to adjust emissions in the respective learning scenarios. However, unlike for Prior-CEA (light red) the option space of Posterior-CEA (orange) does not contain the option space of CEA without learning (dark blue).*

However, Prior-CEA is a problematic decision criterion, too. As Schmidt et al. (2009) point out, this violation of consequentialism is clearly unsatisfactory. Let us illustrate the problem in Figure 3. After learning $\mathcal{H}$ the decision maker would be allowed to choose L_4 by arguing that she would have chosen L_1 if $\mathcal{L}$ had occurred. This would have given her a probability of 50% prior to learning. Such argument can justify any high level of emissions after learning only by claiming that this risk was balanced by scenarios which eventually did not occur.

So far, we have not touched upon the question why using a non-consequentialist criterion for the climate problem should be desirable in the first place. In a different context, it has been repeatedly argued to drop consequentialism for acknowledging the Allais Preferences (Loomes and Sugden, 1982; Machina, 1989; Bradley and Stefansson, 2016). They go back to a famous counterexample to the independence axiom given by Allais (1953). Here, non-consequentialism may be justified by anticipated regret.

Let us briefly explain the Allais Paradox with an example. Kahneman and Tversky (1979) were the first to conduct behavioral experiments on the phenomenon. They repeatedly observed the following preferences: A 20% chance to win 4,000$ is preferred over a 25% chance to win 3,000$. Yet, at the same time a 80% chance to win 4,000$ is exchanged for a sure gain of 3,000$. This is a violation of independence since the former choice is obtained by receiving the latter choice with a chance of one quarter. When explaining their results, Kahneman and Tversky (1979) referred to this as the "certainty effect". There is a qualitative difference between a certain and a uncertain option because the decision maker might anticipate her regret of winning nothing when she could have won something for sure. Taking into account what could have happened if the safe option had been chosen may make the loss even worse. This can be considered in decision theory by dropping consequentialism and making decisions path-dependent (e.g. Loomes and Sugden, 1982; Machina, 1989).

However, as the climate problem does not feature choices between certain and uncertain options, it is hard to see how there can be anticipated regret. For partial learning all options still have uncertain consequences. It would need to be argued, for instance, that we should be even more prudent about climate risks in a good learning scenario because we know we could have ended up in a bad learning scenario. Yet, this would make future decision dependent on a quite arbitrary outdated state of knowledge. Consequentialism, it seems, is not dispensable either.

After all, violating the independence axiom has undesirable implications. Relaxing one of the other three consistency principles, time-consistency, context-independence or consequentialism, is no attractive alternative to solve the EVOI problem.

3.2.3 The Sign of the EVOI

Obviously, the simple example in Figure 3 does not reflect the decision problem tackled by IAMs. It was supposed to illustrate how the restrictive effect arises. In the full climate problem, there are infinitely many options and different combinations of parameters and distributions are possible. Here, the restrictive effect may not always dominate the enlarging effect. Schmidt et al. (2009) obtain a negative EVOI for most but not all parameter combinations in their analysis with MIND. In the following, we will show how this mixed result can be explained by the superimposition of two effects: the convexity of the mitigation cost function and the asymmetry of learning.

Again, let us illustrate the two effects using a simplified example. It was developed by Schmidt et al. (2011). Here, we consider decisions on emissions $E \in \mathbb{R}_{\geq 0}$ instead of emission pathways. Cumulative emissions since preindustrial times are a good approximation of the maximum temperature reached over time (Allen et al., 2009). We assume $T_{max} = a\theta E$ with the climate sensitivity θ and constant conversion factor a. Moreover, the mitigation cost function $C(E)$ is decreasing and refers to the cost-minimal emission distribution over time. We consider perfect learning of three equally likely states of the world in which climate sensitivity is either θ_1,θ_2 or θ_3, where $\theta_1 > \theta_2 > \theta_3$. Finally, we suppose the decision maker applies CEA and aims at a minimum probability of 50% to meet the 2°C target.

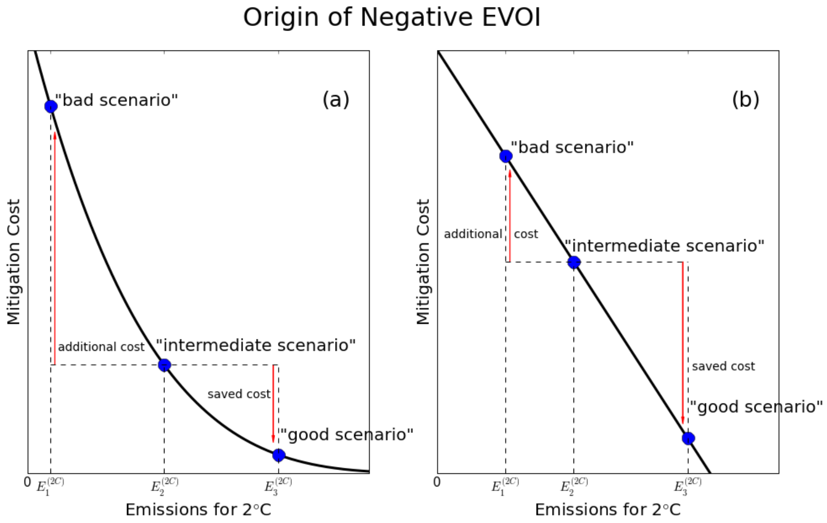


Figure 6: *Illustration of the cost convexity and the asymmetry effect that affect the EVOI in CEA: There are three scenarios of perfect learning in which the 2°C level can be met by emissions E_1, E_2 and E_3 respectively, where $E_1 < E_2 < E_3$. The EVOI refers to the difference in (expected) welfare, i.e. negative mitigation cost, with and without learning. The EVOI decreases the more convex the mitigation cost function $C(E)$ is (left). The EVOI increases the more emissions are allowed in the "good" scenario relative to the emissions reduced in the "bad" scenario (right).*

Let us compare the expected mitigation cost with perfect learning to the mitigation cost without learning in this simplified decision problem. We assume that in both cases compliance is possible and CEA is feasible. Under learning, the decision maker chooses optimal emissions E_1, E_2 and E_3 to meet the 2°C target in the learning scenarios θ_1, θ_2 and θ_3 respectively. Since climate sensitivity is decreasing, cost-minimal emissions have to increase: $E_1 < E_2 < E_3$. Now, without learning the decision maker could choose E_2 because this gives her an exceedance probability of 33% and is the cost-minimal choice for obtaining a probability not higher than 50%. Hence, the EVOI is

$$EVOI = C(E_2) - \frac{1}{3}[C(E_1) + C(E_2) + C(E_3)]. \tag{23}$$

Whether the EVOI in (23) is positive or negative depends on two factors: The distance between E_1, E_2 and E_3 and the convexity of the mitigation cost function $C(E)$. Figure 6 illustrates how both effects may counteract each other. The left panel shows that the EVOI is reduced the more convex the mitigation cost function is. The right panel shows that this effect is counteracted if the amount of emission gained in the “good scenario” $(E_3 - E_2)$ exceeds the required reduction in the “bad scenario” $(E_2 - E_1)$. This second effect depends on the symmetry of the learning scenarios, i.e. on the differences between θ_1, θ_2 and θ_3. If $(\theta_1 - \theta_2) = (\theta_2 - \theta_3)$, the effect increases the EVOI as depicted in Figure 6b since $E(\theta) = T_g/(a\theta)$ is a convex function. More emissions are gained in the “good scenario” (E_3) than are lost in the “bad scenario” (E_1). Hence, the EVOI as in (23) is only negative for a sufficiently convex mitigation cost function which outweighs the asymmetry effect.

Considering the full climate problem, it is theoretically not clear how the superimposition of these two effects plays out for the sign of the EVOI. In the more general case of partial learning, the second effect mainly depends on the skewness of the posterior distributions. The mitigation cost function implicit in an IAM is certainly convex, yet the asymmetric learning effect may outweigh it. Moreover, both effects obtain different weight if the learning scenarios do not occur with equal likelihoods.

Additionally, the expected value of anticipation (EVOA) is always negative in CEA. The EVOA is the component of the EVOI due to the adjustment of decisions before learning relative to the situation of no learning (Lorenz et al., 2012). With learning, the decision maker may also want to alter first-period emissions as she anticipates the learning event. However, the EVOA in CEA is always negative because the posterior risk constraint requires the climate target to be met in all learning scenarios. It implies that first-period emissions are completely determined by the worst-case learning scenario. Not only is this an extremely risk-averse behavior, the negative EVOA suggests that the decision maker would be better-off by ignoring the fact that she will learn until she eventually learns. The negative EVOA by itself is a normatively unappealing feature of CEA.

To conclude, we have seen that CEA may give a negative EVOI and even become infeasible as the set of emission plans which comply with the target changes depending on what is learned. In general, violating independence as implied by probabilistic constraints has normatively unappealing implications under learning. This motivates using a target-based expected utility criterion as CRA which will be presented in the following chapter.

* * *

This section pointed to the problems of CEA under learning. First, CEA can become infeasible if “too much” is learned. Here, the probabilistic climate target cannot be met in all learning scenarios due to past emissions. Second, CEA can exhibit a negative expected value of information. This implies that the decision maker would be better-off if she did not learn in the first place. The EVOI problem arises due to the violation of the independence axiom. As long as a probabilistic constraint is imposed, the problem can only be evaded by dropping other desirable consistency principles. Whether the actual EVOI is positive or negative in CEA depends on the superimposition of two effects: the convexity of the mitigation cost function and the asymmetry of learning scenarios. The infeasibility as well as the EVOI problem conceptually originate in the fact that the set of emission plans the constraint allows for changes depending on what is learned. Hence, alternative criteria are needed to formulate strong sustainability under uncertainty and learning in a self-consistent way.

4 Understanding CRA: A Target-based Expected Utility Criterion

CRA has been developed by Schmidt et al. (2009) as a target-based expected utility criterion to overcome the indicated problems of CEA. This chapter aims at understanding CRA and its properties in more detail. First, section 4.1 introduces the objective function of CRA and explains its calibration to a probabilistic climate target. Second, we show that as an expected utility criterion CRA overcomes the EVOI problem of CEA (section 4.2). Third, we discuss the trade-off criticism brought forward against CRA (section 4.3). Fourth, we address the question under which conditions CRA can be seen as an adequate formalization of strong sustainability (section 4.4).

4.1 Objective Function and Calibration of CRA

When dealing with the climate problem under uncertainty in the target-based normative framing, two values are at stake: climate-induced risk $R(E, p(\theta))$ and mitigation cost $C(E)$. A target-based expected utility criterion needs an objective function to be maximized that relates these to values to each other. This function $W(C, R)$ should obviously decrease in both mitigation cost and climate risk.

Now, as Schmidt et al. (2009) note, the scope of possible functional forms for an expected utility criterion is very restricted. This is because climate risk itself is an expected value. Most generally, we can write it as $R(E, p(\theta)) = \mathbb{E}_{\theta|p(\theta)}[X(E, \theta)]$ with some non-decreasing exceedance function $X(E, \theta)$. To comply with expected utility theory, the objective function must be linear in the probabilities and hence $W(C, R)$ must be linear in climate risk. The only linear form which aims to reduce both mitigation cost and climate risk, is the additive approach $W = -(C(E) + \beta R(E))$ employed by CRA.[14].

The delicate issue in CRA is the trade-off parameter β. It can be interpreted as the willingness to pay for reducing the climate risk by one unit. If climate risk is exceedance probability R_{ex} as given in (4), then $\beta/100$ is the welfare loss incurred by an 1% increase in the probability to transgress the temperature guard

[14] The other linear form would be the multiplicative approach $W = -C(E)R(E, p(\theta))$. However, it would imply that zero cost are optimal which corresponds to a BAU-scenario.

rail. Likewise, it may be interpreted as the shadow price of the risk constraint of a CEA without learning. However, CRA as defined in (6) uses expected degree years, a risk measure which is based on the magnitude and the duration of potential target overshoots. Here, β can be interpreted as the willingness to pay for reducing the expected overshoot of the temperature guard rail by one degree year.

Now, the crucial question is how to set this normative trade-off parameter. Neubersch et al. (2014) refer the parameter to a probabilistic climate target with respect to the current state of knowledge, i.e. a prior distribution. Their CRA is calibrated such that without learning the optimum provides a 66% chance to meet the 2°C target. As mentioned in section 1.4, this target represents their interpretation of the agreement made at the 17th Conference of the Parties in 2009. Applying CRA, the decision maker attaches the same value to risk reduction in terms of mitigation cost as if she wanted to reach this probabilistic climate target cost-effectively under prior knowledge.

It can be asked whether a repeated calibration of CRA is time-inconsistent in case that the target is politically reconfirmed, while real-world decisions deviate from the optimal pathway. However, as soon as the actual pathway differs from the optimum, circumstances have changed and time-consistency is unaffected. Hence, nothing impedes to recalibrate CRA under the impression of the latest climate policy agreement.

There are various ways to modify CRA by using different risk measures. Neubersch et al. (2014) and Roth et al. (2015) test more stringent risk functions, too, which penalize overshoots with higher order terms of temperature exceedance. Their numerical results show that this enhances environmental stringency but does not change the picture qualitatively: The higher the penalty on target overshoots, the smaller the optimal overshoot will be. The (linear) measure of degree years is, as Neubersch et al. (2014) show, the least environmentally stringent form to avoid the problem of BAU-solutions indicated in section 1.5. Concerns about whether the criterion attaches too large or too small value to avoiding target overshoots can be met by adjusting the risk function. This generally grants CRA much flexibility.

4.2 The EVOI in Expected Utility Theory

CRA overcomes the problems of CEA discussed in chapter 3. First, as an unconstrained optimization it is always feasible. Second, it is consistent with the idea of a probabilistic climate target as CRA actually allows for overshoots of the temperature guard rail under certain circumstances. It references the target to a specific state of knowledge and not to any state of knowledge as CEA does. Third, as an expected utility criterion the EVOI of CRA is always non-negative. Let us present this final point in more detail.

Previously, we have seen that CEA can have a negative EVOI due to its violation of the independence axiom. As an expected utility criterion, CRA cannot have a negative EVOI. Here, we sketch the general argument following Gollier (2001, pp. 357) for a finite number of states of the world. As we will see, the independence axiom ensures that any expected utility criterion exhibits a non-negative EVOI.

Let us consider an expected utility maximization on a finite space of options given by

$$\mathrm{Max}_c \;\; W(c, \boldsymbol{p}) = \mathbb{E}_s[U(c, s_i)] = \sum_{i=1}^{n} p_i U(c, s_i). \tag{24}$$

Here, $U(c, s_i)$ is the outcome utility gained for option c if state of the world s_i occurs. The knowledge of the decision maker is represented by the probability vector $\boldsymbol{p} = (p_1, ..., p_n)$ for n possible states of the world. The probability vector satisfies $\sum_{i=1}^{n} p_i = 1$. The total welfare $W(c, \boldsymbol{p})$ is the expected utility over all states of the world.

Now, learning is modeled by compound lotteries as introduced in chapter 2.1. The decision maker anticipates that in the future she will either obtain the knowledge $\boldsymbol{p}$ or the knowledge $\boldsymbol{p'}$. The first learning scenario occurs with a likelihood of q, while the second occurs with a likelihood of $(1 - q)$. Hence, according to Bayes' Law her prior knowledge is $q\boldsymbol{p} + (1 - q)\boldsymbol{p'}$.

We now consider maximum welfare as a function of the probability vector, i.e. $W^*(\boldsymbol{p}) := \mathrm{Max}_c \; W(c, \boldsymbol{p})$. The key idea is to see that $W^*(\boldsymbol{p})$ is a convex function

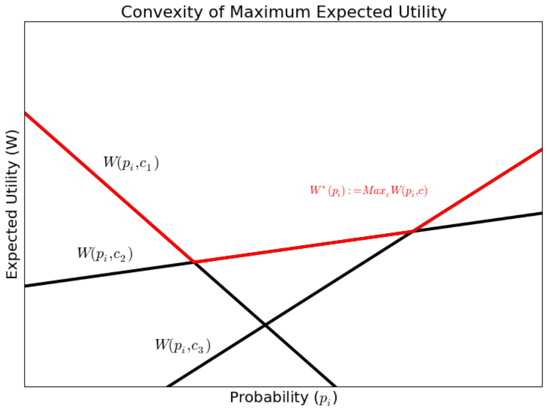


Figure 7: *Illustration of the convexity of maximum expected utility for one dimension p_i of the probability vector $\boldsymbol{p}$. Since the expected utility functions $W(p_i, c_1), W(p_i, c_2)$ and $W(p_i, c_3)$ for the options c_1, c_2 and c_3 (black lines) are linear in p_i, the function $W^*(p_i) := \text{Max}_c\ W(p_i, c)$ is convex (red line).*

as long as $W(c, \boldsymbol{p})$ is linear in $\boldsymbol{p}$. This is illustrated in Figure 7 for one dimension p_i of the probability vector $\boldsymbol{p}$. It shows the welfare obtained for options c_1, c_2 or c_3 depending on the probability p_i. Since the expected utility functions $W(p_i, c_1), W(p_i, c_2)$ and $W(p_i, c_3)$ are linear in p_i (black lines), the function $W^*(p)$ is convex (red line). This is because for increasing p_i the maximum welfare follows the expected utility function with the larger slope at intersection points.

Finally, as $W^*(\boldsymbol{p})$ is convex, it follows for any two probability vectors $\boldsymbol{p}$ and $\boldsymbol{p'}$ obtained by the likelihoods q and $(1 - q)$:

$$qW^*(\boldsymbol{p}) + (1 - q)W^*(\boldsymbol{p'}) \geq W^*(q\boldsymbol{p} + (1 - q)\boldsymbol{p'}). \tag{25}$$

The left hand-side is the expectation over the respective maximum welfare gained in the learning scenarios $\boldsymbol{p}$ and $\boldsymbol{p'}$. The right hand-side is the maximum welfare obtained with the prior knowledge. The first is the expected utility for learning, while the second is the expected utility for no learning. The difference between the two is the EVOI. Hence, the linearity in probability required by the independence axiom ensures that the EVOI is never negative in expected utility maximization.

4.3 The Trade-Off Criticism

CRA has been criticized for its trade-off structure between mitigation cost and climate risk (Hermann Held, personal communication). The criterion, it is argued, is not in the sense of strong sustainability since it does not attach first priority to keeping a "safe minimum standard" of climate change (section 1.3). CRA treats both values, mitigation cost and climate risk, interchangeably. As indicated, the position of the WBGU (2014) asserts that transgressing the guard rail leads to "intolerable consequences that even large-scale benefits in other areas could not compensate" (p. 11). This would, the argument goes, require a lexicographic decision criterion such as CEA under uncertainty and learning, too.

The problem raised is qualitatively different from the consistency problems of CEA. CRA is not inconsistent, but inadequate to represent this presupposed normative framing. Discussing such objection has a different character. Although consistency conditions such as the von Neumann-Morgenstern axioms require normative justification, too, they have a formal definition. Yet, strong sustainability is not a clear-cut term as section 1.3 pointed out. At this point, we need to carefully differentiate between different possible understandings of strong sustainability.

Suppose we accept the CRA trade-off criticism and require that the principle of non-substitutability applies to probabilistic measures of climate-related and economic values as well. Let us try to find an adequate decision criterion for this position. We can then discuss to what extend the implications of such criterion are in the general sense of strong sustainability as introduced in section 1.3.

For obtaining a decision criterion in this presupposed sense of strong sustainability, the continuity axiom would need to be dropped. As section 2.2 showed, assuming completeness and transitivity, continuity is sufficient for the existence of a utility representation. As long as both climate risk and mitigation cost matter, the utility function will feature some trade-off between both values. Hence, CRA is not an adequate decision criterion to formalize strong sustainability under learning. CEA is lexicographic, yet it features the consistency problems discussed in chapter 3. They do not occur if we require the completeness and the independence axiom. Hence, it makes sense to look for alternative criteria which satisfy all von Neumann-Morgenstern axioms except for continuity.

As Schmidt et al. (2009) point out, such axiomatic basis is provided by lexicographic expected utility theory as in e.g. Blume et al. (1991). Here, the continuity axiom is replaced by a weaker axiom while the other axioms of expected utility theory remain unaffected. Hence, it is a more general framework. Applied to the climate problem under learning, lexicographic expected utility theory contains decision criteria of the following form:

$$\begin{aligned}&\text{lex. Max}_{(E_0,\boldsymbol{E})}\\ &\{\mathbb{E}_m\ \mathbb{E}_{\theta|p_m(\theta)}\ [U_1(E_0,E_m,\theta)],\ \dots\ ,\mathbb{E}_m\ \mathbb{E}_{\theta|p_m(\theta)}\ [U_k(E_0,E_m,\theta)]\}.\end{aligned} \tag{26}$$

Here, the expression lex. $\text{Max}_x\{V_1(x), V_2(x), ..., V_k(x)\}$ denotes a lexicographic maximization: First V_1 is maximized, if multiple x give the maximum of V_1, then V_2 is maximized and so on. As usual, E_0 denotes the emission pathway before learning and $\boldsymbol{E} = (E_1, ..., E_n)$ the vector of emission pathways of learning scenarios $\{1, ..., n\}$. Moreover, $\mathbb{E}_m$ and $\mathbb{E}_{\theta|p_m(\theta)}$ are the expectation over the learning scenarios and the expectation over climate sensitivity θ with respect to the probability distribution $p_m(\theta)$. The $U_1(E_0, E_m, \theta), ...U_k(E_0, E_m, \theta)$ are k utility functions on emissions before and after learning (E_0, E_m) under climate sensitivity θ in descending lexicographic order of importance.

Now, the question arises how to specify the $U_1(E_0, E_m, \theta), ...U_k(E_0, E_m, \theta)$ in order to obtain a criterion in the above sense of strong sustainability. This strong sustainability would require U_1 to only represent climate-related value with respect to a potential overshoot of the guard rail T_g. For this we define a general exceedance function based on the notation of section 1.5 as

$$\begin{aligned}&X(E_0, E_m, \theta) :=\\ &\int [\Theta[T(t)(E_0, E_m, \theta) - T_g]x(T(t)(E_0, E_m, \theta))]e^{-\rho t}dt,\end{aligned} \tag{27}$$

where $x(T(t))$ is some non-decreasing function of the temperature trajectory $T(t)$. We formalize strong sustainability by setting $U_1 := X(E_0, E_m, \theta)$. Hence,

$V_1(E_0, \boldsymbol{E}) := \mathbb{E}_m \mathbb{E}_{\theta|p_m(\theta)} \left[U_1(E_0, E_m, \theta) \right]$ is the most general form of a risk function as we understand it. For instance, if x is a constant, then V_1 is exceedance probability. If $x(T)$ is linear, then V_1 is expected degree years and so on. The problem is that lexicographic expected utility theory would only be helpful if the primary utility function U_1 could be chosen such that $V_1(E_0, \boldsymbol{E})$ had multiple optima.

However, the combination of two features specific to the climate problem make this impossible. First, temperature $T(E_0, E_m, \theta)$ is strictly increasing in both emissions as well as climate sensitivity and has no upper bound (I). Second, it is most likely that future learning will not be able to give an upper bound to climate sensitivity either. Hence, we can assume that the posterior distributions have infinite support (II), i.e. $\forall m \in \{1, ..., n\}, \theta > 0 : p_m(\theta) > 0$.

Now, as past emissions have already occurred, (I) implies that any amount of temperature rise may be observed with arbitrarily high climate sensitivity. So (II) implies that the primary expected utility function can never be zero, i.e. for all emission plans $(E_0, \boldsymbol{E})$ we have $V_1(E_0, \boldsymbol{E}) > 0$. Finally, (I) implies that maximum mitigation, i.e. zero emissions for all times and learning scenarios, is optimal due to the primary criterion in (26) and no secondary criterion would make any difference. Hence, the only way of meeting the CRA trade-off criticism would be by minimizing a measure of climate risk without considering mitigation cost at all.

Let us rephrase the argument in a less formalistic manner: The trade-off criticism imposes the requirement of non-substitutability by strong sustainability also on probabilistic quantities. Now, if reducing mitgiation cost is always subordinated to reducing the chance of a possible temperature overshoot even by the smallest amount and, moreover, limited knowledge does not allow to exclude arbitrarily high values of climate sensitivity, maximum mitigation is optimal. Decreasing emissions will always decrease climate risk, which will never be zero. Hence, accepting the trade-off criticism against CRA implies that mitigation cost are irrelevant. A proponent of strong sustainability, who sincerely argues to extend the idea of non-substitutability also to probabilistic quantities, must support an immediate shutdown of the economy for climate reasons.

This position may be consistent, but it is at odds with a certain aspect of climate targets presented in section 1.3. Always prioritizing mitigation because of a non-zero probability to transgress the 2°C target obviously induces "excessive cost", a consequence that already the WBGU (1995) rejected. As section 1.3 pointed out, the discourse around the target-based normative framing has always had both sides in mind: A (probabilistic) climate target is supposed to avert high climate risk at reasonable cost. CRA can be defended against the trade-off criticism as the objection would imply to ignore mitigation cost completely.

Finally, let us ask about the motivation behind a lexicographic criterion in the first place. Requiring non-substitutability also for probabilistic measures of climate-related and economic value violates the continuity axiom. Section 2.2 pointed out that it is hard to defend specific probabilistic thresholds within the open interval $(0, 1)$. However, essentially a similar objection can be made against deterministic CEA: Any pathway reaching 2.2°C or 2.1°C is considered worse than any pathway that stays below the 2°C level. The numerically exact threshold level is quite arbitrary which makes the resulting discontinuity somewhat artificial.

This brings us back to the interpretation of the 2°C target. Jaeger and Jaeger (2011) provide a comprehensive review of the different arguments made in support of the 2°C target. Interestingly, they highlight a pragmatic argument which is that a number as 2°C provides a "focal point" and "collective narrative" for coordinating political action. They compare the function of the target to speed limits in traffic: There is no particular reason for a limit of 50 instead of 47 or 53 kilometer per hour except that it is easier to communicate. There would be no significantly lower or higher risk with those limits. It is a smooth transition zone to unacceptably dangerous traffic. As they emphasize, the focal point argument hardly provides sufficient reason for a 2°C target. The general regime of the target level needs to be determined by climate-related and economic argument. However, their analogy implies that the exact level is not supposed to be a cut-off point. Climate risk is not actually non-continuous at 2°C. Following this interpretation, the climate problem does not require a lexicographic criterion.

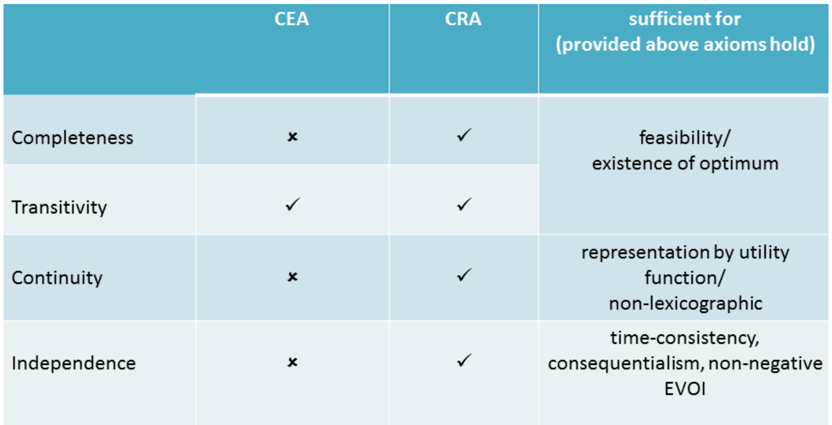
4.4 The Role of CRA

Finally, let us ask how decision making in CRA can be interpreted. The question remains, under which conditions it is an adequate formalization of strong sustainability. CRA overcomes the infeasibility and the EVOI problem of CEA by meeting the consistency standard of expected utility theory. We will explain in which sense a calibrated CRA refers to the concept of a probabilistic climate target suggested by strong sustainability.

The decision maker of CRA can be characterized as follows: She is an expected utility maximizer who makes decisions on the climate problem under learning only on the basis of mitigation cost and climate risk. Here, climate risk is understood as in section 4.3 by $R(E_0, E_m|p_m) := \mathbb{E}_{\theta|p_m(\theta)}\left[X(T(E_0, E_m, \theta))\right]$, where X is a general exceedance function as defined in (27). Now, it depends on whether or not strong sustainability can identify with such characterization. CRA does not have a lexicographic structure with a primary climate criterion. However, as section 4.3 showed, it is questionable whether this is the point strong sustainability wishes to make under learning after all.

In general, complying with the von Neumann-Morgestern axioms should not be regarded as the shrine of rationality. As we have seen, they are mainly motivated by excluding certain normatively unappealing effects under learning such as those discussed in section 3.2. However, if their occurrence can be ruled out for other, maybe empirical, reasons or is not considered to be a problem in the first place, non-expected utility criteria can still be used. For example, CEA without learning may still be an informative decision criterion in IAM analysis.

Let us shortly recapitulate the implications of accepting the von Neumann-Morgestern axioms as summarized in Figure 8. Accepting completeness and transitivity implies the existence of an optimal choice. Without continuity, this choice may be found by a lexicographic criterion. Accepting continuity in addition implies to maximize a utility function. As long as independence is not added to these requirements, utility functions which are non-linear in the probabilities can be employed (see e.g. Machina, 1989). Finally, also requiring independence implies using an expected utility criterion which ensures time-consistency, consequentialism and a non-negative EVOI.



	CEA	CRA	sufficient for (provided above axioms hold)
Completeness	×	✓	feasibility/ existence of optimum
Transitivity	✓	✓	
Continuity	×	✓	representation by utility function/ non-lexicographic
Independence	×	✓	time-consistency, consequentialism, non-negative EVOI

Figure 8: *Compliance of CEA and CRA with the von Neumann-Morgenstern axioms and summary of the axioms' implications.*

Following our discussion of the von-Neumann Morgenstern axioms against the background of the climate problem, we think that there are good reasons for applying CRA as a proponent of strong sustainability under learning. Given a probabilistic climate target, CRA can refer to it under learning using the calibration of Neubersch et al. (2014). Let us extent on this point and explain how the decisions in CEA without learning are related to a calibrated CRA.

Suppose a proponent of strong sustainability agrees to apply CEA under uncertainty and without learning. For example, she favors a mitigation policy which gives a probability of 66% for meeting the 2°C target. Section 3.1 pointed out that applying CEA with a posterior risk constraint under learning conceptually misunderstands the probabilistic climate target as it refers it to any possible state of knowledge. The CRA calibration of Neubersch et al. (2014) implies that the target is only met under the prior state of knowledge. Hence, the initial statement of the proponent of strong sustainability, i.e. that a 66% probability of meeting the 2°C target is optimal without learning, still holds. Now, under a different state of knowledge CRA attaches the same value to climate risk in terms of mitigation cost. It is the value which would make cost-effective target compliance optimal under the prior distribution.

The crucial step from CEA under uncertainty to CRA with learning is the risk function. It is required for structuring preferences above the temperature guard rail. Different forms can be employed depending on how dangerous an overshoot of the guard rail temperature is perceived. As a target-based criterion, CRA is meant to refer not to impact-based damage estimates but to positions and pledges in the climate policy discourse. Additional to a probabilistic climate target, CRA thus requires some input about what should be done once the temperature guard rail is transgressed.

The modeling choice of Neubersch et al. (2014) is to use expected degree years as the risk function. As pointed out in section 1.5, it is the least penalizing functional form for excluding solutions that suggest to ignore climate change in “bad learning scenarios”. Without learning, their CRA suggests more emission reduction in the long run than does CEA. The reason is that the duration of the overshoot is taken into account and the risk increases the longer the overshooting temperature trajectories of high climate sensitives stay above the guard rail. There is an incentive to reduce temperature in these scenarios to the guard rail level in the long run. When using more convex functional forms than the degree years, emissions are reduced further such that the solution of CRA for no learning is even more prudent than in CEA. Hence, without learning the calibrated CRA always suggests environmentally more stringent decisions than CEA.

Finally, CRA is similar to CBA in its mathematical structure. Yet, the crucial difference is how economic damage functions and the calibrated risk functions are constructed. A damage function aggregates economic assessments of specific climate impacts. Some evaluations are referred to market data, other impacts need to be based on more indirect assessments (e.g. Nordhaus, 2013, pp. 69-135). CRA takes a different perspective. Regardless of actual impact estimates, it analyzes decision on the basis on climate targets. It attempts to relate its main normative assumptions to certain positions and agreements in climate policy.

CRA and CBA provide answers to different questions. Both may be informative for policy makers at different stages of a decision process. One of the advantages of CRA is that it is based on a few, more explicit normative assumptions. The criterion requires a risk function and three normative parameters, constituting the probabilistic climate target: the temperature guard rail T_g, the risk guard rail R_g and the state of knowledge $p(\theta)$ the risk guard rail is supposed to refer to. This still relatively simple structure makes CRA attractive for informing decision makers on the economic implications of climate targets under uncertainty and learning.

* * *

The chapter presented CRA as an expected utility criterion that overcomes the consistency problems of CEA. CRA is the only way to conduct an unconstrained expected utility maximization featuring mitigation cost and climate risk. It works with a normative trade-off parameter β which represents the willingness to pay for reducing one unit of climate risk. As an expected utility criterion, it always exhibits a non-negative EVOI. CRA has been criticized for using a trade-off structure between mitigation cost and climate risk which was not in the sense of strong sustainability. However, accepting this objection would imply ignoring mitigation cost completely. CRA characterizes an expected utility maximizer who decides about the climate problem on the basis of mitigation cost and climate risk. It is an attractive decision framework to deal with climate targets in a setting of uncertainty and learning.

5 Exploring Alternatives: Minimum Risk Analysis

Two target-based decision criteria have been discussed in the previous chapters. The first was CEA, a risk-constrained cost minimization. The second was CRA, a minimization of a weighted sum of cost and risk. This last chapter will present the remaining alternative: a cost-constrained risk minimization. As before, we will also ask how decision making based on this criterion can be interpreted.

Section 5.1 will introduce minimum-risk analysis (MRA) and discuss some of its features. Subsequently, section 5.2 will reflect on the role of MRA in target-based decision analysis and point to its chances and limitations.

5.1 Introducing a Cost-Constrained Approach

The previous chapters investigated two existing target-based criteria to make decisions on the climate problem under learning. In the following, we will present a third criterion which we will refer to as minimum-risk analysis (MRA). Using the notation introduced in section 1.5, MRA without learning is given by

$$\begin{aligned} &\mathrm{Min}_E \quad R(E|p(\theta)) \\ &\text{s.t.} \quad C(E) \leq C_g. \end{aligned} \tag{28}$$

Here, the emission pathway with the least climate risk is chosen that does not incur higher mitigation cost than the cost guard rail C_g allows. Now, as discussed for CEA in section 3.2, there are again two ways to formulate this constraint under learning. We can either place a cost guard rail on the expected mitigation cost over all learning scenarios (prior constraint) or on every learning scenario respectively (posterior constraint). However, similar to Prior-CEA in section 3.2, the former criterion violates consequentialism. Whether a certain choice is admissible after learning depends on what would have been chosen in other learning scenarios. This is why we only consider MRA with a posterior cost constraint under learning. It is given by the optimization

$$\begin{aligned} &\mathrm{Min}_{(E_0,\boldsymbol{E})} \quad \mathbb{E}_m[R(E_0, E_m|p_m(\theta))] \\ &\text{s.t.} \quad \forall m: \quad C(E_0, E_m) \leq C_g. \end{aligned} \tag{29}$$

MRA is structurally similar to CEA only that cost are constrained and risk is minimized. It violates the von Neumann-Morgenstern axioms (see Appendix A.2) such that it is no expected utility criterion.

However, the two problems of CEA do not occur. First, MRA is always feasible. Any emission pathway which complies with the cost constraint in (28) also complies with any of the cost constraints in (29). Learning cannot make MRA infeasible as there is no restrictive effect on the option space and MRA without learning is feasible as long as $C_g \geq 0$. The BAU-pathway $E_{BAU}(t)$ with $C(E_{BAU}) = 0$ can always be chosen.

Second, since the constraint is not probabilistic and the objective function has expected utility form, the EVOI of MRA cannot be negative. There is no restrictive effect on the admissible option space for risk minimization. The optimal emission pathway without learning can still be chosen in any learning scenario. This emission plan gives the same risk level as in the no-learning optimum since the risk function is linear in the probabilities:

$$\sum_m \pi_m \int p_m(\theta) X(E_0, E_m, \theta) d\theta = \int p(\theta) X(E_0, E_m, \theta) d\theta. \tag{30}$$

Hence, the EVOI in MRA is at least zero. It corresponds to the amount of climate risk that can be reduced due to learning.

The neglect of any cost minimization raises a conceptual problem of MRA as soon as we drop the (quite realistic) assumption that the posterior probability distributions have infinite support. In fact, if a posterior distribution allows to hold the temperature guard rail with certainty at less than guard rail cost, there are multiple optima in this learning scenario. The zero risk level can be reached with more or less admissible mitigation cost. Specifically, this situation occurs for perfect learning. The problem can be overcome by conducting a lexicographic expected utility maximization as introduced in section 4.3 to obtain the risk minimum with the least mitigation cost. This is why we write MRA in its most general cost-efficient form as

$$\begin{aligned} &\text{lex. Min}_{(E_0,\boldsymbol{E})} \ \{\mathbb{E}_m[R(E_0,E_m|p_m(\theta))], \mathbb{E}_m[C(E_0,E_m)]\} \\ &\text{s.t.} \quad \forall m: \ \ C(E_0,E_m) \le C_g. \end{aligned} \tag{31}$$

As in (26) the expressions $\mathbb{E}_m[R(E_0,E_m|p_m(\theta))]$ and $\mathbb{E}_m[C(E_0,E_m)]$ are minimized in lexicographic order. So, if multiple emission plans $(E_0,\boldsymbol{E})$ give minimal climate risk, the cost-minimal solution among them is preferred.

MRA in its cost-efficient form always incurs less or equal (expected) mitigation cost when learning is included. In fact, as argued in section 4.3, if the risk cannot be zero, the secondary criterion makes no difference. Then, the mitigation cost without learning correspond to the expected mitigation cost with learning. The full cost budget is used up in every learning scenario. Yet, if the temperature guard rail can be held with certainty in at least one learning scenario at lower than guard rail cost, the two criteria differ and cost-efficient MRA incurs strictly less (expected) mitigation cost with learning than without learning.

In lexicographic optimization under learning, we face the problem that the EVOI can only refer either to the primary or the secondary function. Extending the concept of the EVOI to a lexicographic structure is questionable. For instance, it could be suggested to refer the EVOI to the risk function, if optimal risk levels for learning and no learning are different, and to the saved mitigation cost if the former are equal. However, this would change the interpretation of the EVOI from reduced climate risk to saved mitigation cost depending on the specific optimum reached. Yet, if the EVOI refers only to the primary risk function, it ignores saved mitigation cost from the second criterion. Hence, in the special case that a cost-efficient MRA is required, it is not clear how to quantify the benefit gained by new information.

5.2 The Role of MRA

Finally, we discuss how applying MRA under learning can be interpreted. Although MRA is lexicographic, it is hardly in the sense of strong sustainability since it inverts the order of importance. MRA employs a primary cost criterion and a secondary risk criterion. It characterizes a decision maker who spends the same amount of mitigation cost regardless of the learning scenario she receives.

Yet, MRA may still be an interesting decision criterion, although from a very different perspective. The criterion shows how much climate change can be mitigated at maximum with a fix budget of mitigation cost. MRA provides an answer to the question which temperature rise we would obtain if in any learning scenario we could only spend the same predefined amount of mitigation cost. This can be interesting for investigating scenarios of real-world politics where decision makers effectively face economic budget constraints.

However, there are two problems of MRA due to its structure as a constrained optimization. First, the EVOI only takes the reduced climate risk into account, although cost-effective MRA may also reduce the expected mitigation cost under learning. Second, MRA faces the very problem of a negative EVOI as CEA if there is uncertainty and learning about mitigation cost, i.e. uncertainty in the economic system. The familiar deficiencies of probabilistic constraints discussed in section 3.2 reoccur.

The question is whether the uncertainty about mitigation cost is a significant factor to consider in IAM analysis on the climate problem. Held et al. (2009) include uncertainty about the fossil resource base and the future learning rates as well as the floor cost of renewable energy production. Estimates of these parameters may differ considerably. For instance, the IPCC (2014b, p. 525) estimates the fossil resource base between 8,500 and 13,600 GtC. This is the total amount of fossil fuels that can be potentially extracted at economic levels. Furthermore, estimating the cost development of renewable energy depends on a number of complex factors such as future technological change and energy market structures which are inherently difficult to predict (IPCC, 2014b, pp. 538).

Using the model MIND, Held et al. (2009) show that the optimal pathway for reaching the 2°C is also sensitive to economic uncertainty, although less than to uncertainty about climate sensitivity. They find that, if climate sensitivity is high, the mitigation cost for meeting the 2°C target vary by up to 1% in BAU-welfare due to the economic uncertainty. MRA would only be a viable criterion if these uncertainties can be neglected.

* * *

In this final chapter, we introduce MRA as a third target-based decision criterion. MRA minimizes climate risk subject to a mitigation cost constraint imposed on every learning scenario. If the temperature guard rail can be held with certainty in at least one learning scenario and at admissible cost, a more general cost-efficient MRA is required. MRA is not an expected utility criterion but evades the two general problems of CEA. It is always feasible and gives a non-negative EVOI. However, the EVOI takes only the reduced climate risk into account. MRA is not an adequate criterion for strong sustainability. Still, it might be informative for investigating the climate scenarios that can be obtained with a given cost budget.

Summary and Conclusion

This study tackled the question of how to consistently formalize strong sustainability in an adequate decision criterion for the climate problem under learning. Chapter 1 pointed out that strong sustainability has originally interpreted climate targets as maximum acceptable levels of warming. This implies using lexicographic decision criteria where the primary criterion is to meet the climate target, while minimizing mitigation cost is secondary. Under uncertainty probabilistic climate targets can be formulated. They limit the probability of exceeding some predefined temperature level. Two decision criteria have been suggested to deal with probabilistic climate targets under learning: CEA and CRA. CEA finds cost-effective solutions for meeting the target in all learning scenarios, while CRA minimizes a weighted sum of mitigation cost and climate risk, where the latter is a measure of the expected temperature overshoot of the target.

Chapter 2 presented the main tool of the analysis: The von Neumann-Morgenstern axioms. They are the necessary and sufficient consistency conditions for applying expected utility theory, the standard framework of decision-making under uncertainty. While CRA is an expected utility criterion, CEA violates the completeness, the continuity and the independence axiom.

Chapters 3 and 4 developed the discussion of CEA and CRA against the background of the von Neumann-Morgenstern axioms. The violation of completeness and independence by CEA leads to troubling inconsistencies: First, CEA becomes infeasible as soon as the target cannot be met in all learning scenarios. Preferences are not defined over target-overshooting pathways. Moreover, by referring the target to any possible state of knowledge, CEA is at odds with the idea of a probabilistic target. Second, violating the independence axiom, CEA can have a negative expected value of information. This implies that the decision maker may be better-off without new information. More generally, non-independent decision criteria must either be time-inconsistent or take into account counterfactual events to avoid this problem. Dropping this axiom under learning, we think, is hard to justify.

As an expected utility criterion, CRA overcomes the consistency problems of CEA. Yet, while CEA still captures the idea of a primary climate criterion, CRA is a non-lexicographic criterion. Lexicographic criteria are only obtained by dropping the continuity axiom. However, insisting on a primary climate criterion also under learning implies accepting any amount of mitigation cost as long as the probability density distribution of climate sensitivity has infinite support. Moreover, regardless of this extreme implication, we find no convincing argument why the climate problem would require to drop continuity. Not only is there no perfectly safe option, it is also hard to see why a small overshoot of some probability level should be penalized exorbitantly. Our result is the following: CRA is an adequate formalization of strong sustainability if and only if this position accepts the von Neumann-Morgenstern axioms and agrees to represent climate-related concern by a separate expected measure of guard rail overshoot, i.e. a climate risk function. We think that there are good reasons for both.

Lastly, chapter 5 briefly presents MRA, a possible third target-based decision criterion under learning. Here, risk-minimal solutions are found such that mitigation cost do not transgress a predefined cost guard rail in any learning scenario. Although MRA is not an expected utility criterion either, it evades the problems of CEA as long as there is no uncertainty about mitigation cost. Moreover, if the temperature guard rail can be held with certainty in at least one learning scenario, MRA requires a secondary cost criterion to yield cost-effective solutions. MRA is hardly in the sense of strong sustainability as it prioritizes mitigation cost over climate risk. Rather, it can be seen as an analysis of the minimum climate impact obtainable with a given amount of mitigation effort. This can be interesting for investigating scenarios of real-world politics where decision makers effectively face economic budget constraints.

We suggest proponents of strong sustainability to employ CRA for analyzing the climate problem under learning. However, this framework is only a blueprint and a couple of normative assumptions need to be clarified in dialog with decision makers. After all, applying CRA requires a specific risk function and a probabilistic climate target.

Our discussion of the von Neumann-Morgenstern against the background of the climate problem revealed that applying non-expected utility criteria under learning comes with concessions to generally desirable consistency principles. Yet, the normative weight attached to these concessions is certainly debatable.

For instance, violating the continuity axiom by using lexicographic decision criteria may in many contexts not be particularly troubling. The strength of the axioms is to exclude certain normatively unappealing effects in decision making. Yet, if their occurrence can be ruled out for other, maybe empirical, reasons or is not considered to be a problem in the first place, non-expected utility criteria can still be used. If learning is not taken into account, for instance, CEA is still an informative decision criterion for analyzing the economic implications of climate targets.

Several questions are left for future research: First, CRA has only been applied to learning about climate-related uncertainty. However, future decision makers will likely have more knowledge about mitigation cost, too. One source of current uncertainty that will be resolved as the renewable energy sector grows in the future are the floor cost of renewable energies, i.e. the cost once technologies are mature. CRA could also investigate the climate problem by taking into account learning about these economic uncertainties.

Second, the relation of CBA and CRA requires further inquiry. Essentially, CRA replaces the impact-based damage function of CBA by a risk function. The difference is not so much in the mathematical formalism but in the normative reference of decision analysis on the climate problem. CBA is based on economic impact assessments, while CRA refers to positions and pledges in policy making. Discussing this difference would require the much broader perspective of political theory. For example, it may be asked under which conditions decision analysis can meaningfully infer normative parameters as required in CRA by reference to policy discourse or stakeholder processes.

Third, MRA still needs to be implemented in an actual IAM. Moreover, a risk minimization subject to an economic constraint can be done in different ways. Global time-aggregated mitigation cost may not be the most useful indicator for the "economic feasibility" of climate policies. Other formulations such as restrictions on minimum annual growth rates can be thought of. Still, all those approaches are only promising as long as uncertainty and learning about the economic implications of emission reductions may be neglected. The idea of MRA could only be sketched in this study and requires further investigation.

A Appendix

A.1 Independence and Linear Probabilities

Here, we give the last part of the proof of the von Neumann-Morgenstern Theorem following Gollier (2001, pp. 7-8). The decision maker has preferences $\succeq$ over the set of lotteries $\mathbf{L}$. The utility function $V(L), L \in \mathbf{L}$, is defined as the probability for which $L \sim V(L)\overline{L} + (1 - V(L))\underline{L}$, where $\overline{L}$ is the most preferred and $\underline{L}$ the least preferred lottery.

Using this definition, the independence axiom and the axiom of reduction, we show that for any $\beta \in [0, 1]$ and $L_1, L_2 \in \mathbf{L}$:

$$V(\beta L_1 + (1 - \beta)L_2) = \beta V(L_1) + (1 - \beta)V(L_2). \tag{32}$$

First, let us rename $V_1 := V(L_1)$, $V_2 := V(L_2)$ for notational convenience. The above notion of a "definition" may be a bit misleading since without independence we could not simply "plug in" $V_1\overline{L} + (1 - V_1)\underline{L}$ for L_1 in a preference statement. Both are different but equally valued lotteries. However, we can mix both lotteries with lottery L_2 and conclude by using independence twice (for both directions $\succeq$ and $\preceq$):

$$\beta L_1 + (1 - \beta)L_2 \sim \beta[V_1\overline{L} + (1 - V_1)\underline{L}] + (1 - \beta)L_2. \tag{33}$$

By the same argument we are allowed to "plug in" the definition of L_1. Hence, the first lottery in (33) must still be valued equally as

$$\beta[V_1\overline{L} + (1 - V_1)\underline{L}] + (1 - \beta)[V_2\overline{L} + (1 - V_2)\underline{L}]. \tag{34}$$

The axiom of reduction allows to move β inside the brackets. Then, by rearranging (34), we can conclude that the decision maker must be indifferent between the initial lottery mix $\beta L_1 + (1 - \beta)L_2$ and

$$[\beta V_1 + (1 - \beta)V_2]\overline{L} + [1 - (\beta V_1 + (1 - \beta)V_2)]\underline{L}. \tag{35}$$

The latter is by definition valued by the utility $V(\beta L_1+(1-\beta)L_2)$ which implies the equivalence stated in (32).

A.2 Compliance of MRA with the von Neumann-Morgenstern Axioms

As for CEA in section 2.3, we check the compliance of MRA with the von Neumann-Morgenstern axioms. The argument is similar to the one for CEA in section 2.3. MRA can be expressed as a lexicographic composition of two preference relations $\succ_1$ and $\succ_2$ over simple climate lotteries $L \in \mathbf{L}$. Any climate lottery induces a climate risk $R(L)$ and incurs mitigation cost $C(L)$. The primary and the secondary preference relation of MRA are given by

$$\begin{aligned} &L_1 \succ_1 L_2 \Leftrightarrow \{C(L_1) \leq C_g\} \ \wedge \ \{C(L_2) > C_g\} \\ &L_1 \sim_1 L_2 \Leftrightarrow \{C(L_1) \leq C_g\} \ \wedge \ \{C(L_2) \leq C_g\} \\ &L_1 \succeq_2 L_2 \Leftrightarrow R(L_1) \leq R(L_2). \end{aligned} \tag{36}$$

The MRA-preferences $\succ$ are the lexicographic composition of $\succ_1$ and $\succ_2$, i.e.

$$\begin{aligned} &L_1 \succ L_2 \ \Leftrightarrow \{L_1 \succ_1 L_2\} \vee \{(L_1 \sim_1 L_2) \ \wedge \ (L_1 \succ_2 L_2)\} \\ &L_1 \sim L_2 \ \Leftrightarrow \{L_1 \sim_1 L_2\} \ \wedge \ \{L_1 \sim_2 L_2\}. \end{aligned} \tag{37}$$

As for CEA, MRA is incomplete and transitive.

To check continuity and independence, we need to define MRA preferences over compound lotteries. As done for CEA in section 2.3 with risk, we specify the mitigation cost of a compound lottery by thinking about how MRA would treat this lottery without learning. In fact, the deterministic constraint in MRA is the special case of a probabilistic constraint with 100% compliance probability. MRA seeks to hold mitigation cost below C_g under any circumstance, hence we define for $p \in (0,1)$

$$C(pL_1 + (1-p)L_3) = \mathrm{Max}\{C(L_1), C(L_2)\}. \tag{38}$$

Continuity would be satisfied if for any $L_1, L_2, L_3 \in \mathbf{L}$ we had

$$L_3 \succeq L_2 \succeq L_1 \Rightarrow \exists p \in [0,1] : pL_1 + (1-p)L_3 \sim L_2. \tag{39}$$

However, for $L_1, L_2, L_3 \in \mathbf{L}$ with $C(L_2) < C(L_3) \leq C_g < C(L_1)$ and $R(L_2) > R(L_3)$ this does not hold. MRA preferences imply $L_3 \succ L_2 \succ L_1$. As soon as $p \in (0,1)$, the mitigation cost of the compound lottery $C(pL_1 + (1-p)L_3) = C(L_1)$ transgress the cost guard rail C_g which implies $pL_1 + (1-p)L_3 \prec L_2$. Hence, MRA violates continuity.

Independence would require for any L_1, L_2, L_3 and $p \in [0,1]$:

$$L_1 \succeq L_2 \Leftrightarrow pL_1 + (1-p)L_3 \succeq pL_2 + (1-p)L_3. \tag{40}$$

The same example as above shows that independence is violated. MRA preferences on the simple lotteries are $L_3 \succ L_2$. However, for any $p \in (0,1)$ preferences over $pL_3 + (1-p)L_1$ and $pL_2 + (1-p)L_1$ are not defined since $C(pL_3 + (1-p)L_1) = C(L_1) > C_g$ and $C(pL_2 + (1-p)L_1) = C(L_1) > C_g$. Hence, MRA violates independence, too.

References

Frank Ackerman, Stephen DeCanio, Richard Howarth, and Kristen Sheeran. Limitations of Integrated Assessment Models of Climate Change. *Climatic Change*, 95(3-4), 2009.

Matthew Adler and Nicolas Treich. Prioritarianism and Climate Change. *Environmental and Resource Economics*, 62(2), 2015.

Maurice Allais. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, 21 (4), 1953.

Myles Allen, David Frame, Chris Huntingford, Chris Jones, Jason Lowe, Malte Meinshausen, and Nicolai Meinshausen. Warming Caused by Cumulative Carbon Emissions towards the Trillionth Tonne. *Nature*, 458(7242), 2009.

Stefan Baumgärtner and Martin Quaas. Ecological-economic Viability as a Criterion of Strong Sustainability under Uncertainty. *Ecological Economics*, 68(7), 2009.

Roger Blau. Stochastic Programming and Decision Analysis : An Apparent Dilemma. *Management Science*, 21(3), 1974.

Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic Probabilities and Choice under Uncertainty. *Econometrica*, 59(1), 1991.

Sandrine Bony, Robert Colman, Vladimir Kattsov, Richard Allan, Christopher Bretherton, Jean-Louis Dufresne, Alex Hall, Stephane Hallegatte, Marika Holland, William Ingram, David Randall, Brian Soden, George Tselioudis, and Mark Webb. How well do we Understand and Evaluate Climate Change Feedback Processes? *Journal of Climate*, 19(15), 2006.

Richard Bradley and Orri Stefansson. Counterfactual Desirability. *British Journal for the Philosophy of Science*, 2016.

Simon Caney. Human Rights, Climate Change, and Discounting. *Environmental Politics*, 17(4), 2008.

Mark Charlesworth and Chukwumerije Okereke. Policy Responses to Rapid Climate Change: An Epistemological Critique of Dominant Approaches. *Global Environmental Change*, 20(1), 2010.

Marc Davidson. Climate Change and the Ethics of Discounting. *Wiley Interdisciplinary Reviews: Climate Change*, 6(4), 2015.

Michel den Elzen and Detlef Van Vuuren. Peaking Profiles for Achieving Long-term Temperature Targets with more Likelihood at Lower Costs. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (46), 2007.

Ottmar Edenhofer, Nico Bauer, and Elmar Kriegler. The Impact of Technological Change on Climate Protection and Welfare: Insights from the Model MIND. *Ecological Economics*, 54(2-3), 2005.

Itzhak Gilboa. *Theory of Decision under Uncertainty.* Cambridge University Press, New York, 2009.

Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler. Is it Always Rational to Satisfy Savage's Axioms? *Economics and Philosophy*, 25, 2009.

Christian Gollier. *The Economics of Risk and Time.* Massachusetts Institute of Technology, 2001.

Hermann Held, Elmar Kriegler, Kai Lessmann, and Ottmar Edenhofer. Efficient Climate Policies under Technology and Climate Uncertainty. *Energy Economics*, 31, 2009.

Chris Hope. Optimal Carbon Emissions and the Social Cost of Carbon over Time under Uncertainty. *The Integrated Assessment Journal*, 8(1), 2008.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013a.

IPCC. Annex III: Glossary. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013b.

IPCC. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.*

Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014a.

IPCC. *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014b.

Carlo Jaeger and Julia Jaeger. Three Views of Two Degrees. *Regional Environmental Change*, 11, 2011.

Daniel Kahneman and Amos Tversky. Prospect theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 1979.

Ralph Keeney. Decision Analysis: An Overview. *Operations Research*, 30(5), 1982.

Irving Lavalle. On Information-Augmented Chance-Constrained Programs. *Operations Research Letters*, 4(5), 1986.

Graham Loomes and Robert Sugden. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *Economic journal*, 92(368), 1982.

Alexander Lorenz, Matthias Schmidt, Elmar Kriegler, and Hermann Held. Anticipating Climate Threshold Damages. *Environmental Modeling and Assessment*, 17(1-2), 2012.

Gunnar Luderer, Robert Pietzcker, Christoph Bertram, Elmar Kriegler, Malte Meinshausen, and Ottmar Edenhofer. Economic Mitigation Challenges: How Further Delay Closes the Door for Achieving Climate Targets. *Environmental Research Letters*, 8(3), 2013.

Mark Machina. Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty. *Journal of Economic Literature*, 27(4), 1989.

Michael Mandler. Incomplete Preferences and Rational Intransitivity of Choice. *Games and Economic Behavior*, 50(2), 2005.

Alan Manne and Richard Richels. MERGE: An Integrated Assessment Model for Global Climate Change. In Richard Loulou, Jean-Philippe Waaub, and Georges Zaccour, editors, *Energy and Environment*. Springer US, Boston, 2005.

Delf Neubersch, Hermann Held, and Alexander Otto. Operationalizing Climate Targets under Learning: An Application of Cost-risk Analysis. *Climatic Change*, 126(3-4), 2014.

Eric Neumayer. *Weak versus Strong Sustainability.* Edward Elgar Publishing Limited, fourth edition, 2013.

William Nordhaus. *A Question of Balance: Weighing the Options on Global Warming Policies.* Yale University Press, 2008.

William Nordhaus. *The Climate Casino.* Yale University Press, 2013.

Roger Perman, Yue Ma, and James McGilvray. *Natural Resource and Environmental Economics.* Longman Publishing, first edition, 1996.

Robert Pindyck. The Climate Policy Dilemma. *Review of Environmental Economics and Policy*, 7(2), 2013.

Gerard Roe and Marcia Baker. Why is Climate Sensitivity so Unpredictable? *Science*, 318, 2007.

Robert Roth, Delf Neubersch, and Hermann Held. Evaluating Delayed Climate Policy by Cost-Risk Analysis Evaluating Delayed Climate Policy by Cost-Risk Analysis. *EAERE Paper*, 2015.

Matthias Schmidt, Alexander Lorenz, Hermann Held, and Elmar Kriegler. Climate Targets in an Uncertain World. *Working Paper, Potsdam Institute for Climate Impact Research*, 2009.

Matthias Schmidt, Alexander Lorenz, Hermann Held, and Elmar Kriegler. Climate Targets under Uncertainty: Challenges and Remedies. *Climatic Change*, 104(3-4), 2011.

Stephen Schneider and Michael Mastrandrea. Probabilistic Assessment of ”Dangerous” Climate Change and Emission Pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44), 2005.

Amartya Sen. Internal Consistency of Choice. *Econometrica*, 61(3), 1993.

Klaus Steigleder. Climate Risks, Climate Economics and the Foundations of a Rights-based Risk Ethics. *Journal of Human Rights*, 15(1), 2016.

Richard Tol. The Economic Change Effects of Climate. *The Journal of Economic Perspectives*, 23(2), 2009.

Richard Tol. The Economic Impact of Climate Change in the 20th and 21st Centuries. *Climatic Change*, 117(4), 2013.

UNFCCC: United Nations Framework Convention on Climate Change. Report of the Conference of the Parties on its Seventeenth Session held in Durban from 28 November to 11 December 2011. *UNFCCC/CP/2011/9/Add.1*, 2012.

UNFCCC: United Nations Framework Convention on Climate Change. Adoption of the Paris Agreement. *Conference of the Parties on its twenty-first session*, 21932, 2015.

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1944.

Peter Wakker. Nonexpected Utility as Aversion of Information. *Journal of Behavioral Decision Making*, 1, 1988.

Peter Wakker. Justifying Bayesianism by Dynamic Decision Principles. *Working paper, Leiden University Medical Center*, 1999.

WBGU: German Advisory Council on Global Change. Scenario for the Derivation of Global CO2 Reduction Targets and Implementation Strategies. *Statement on the Occasion of the First Conference of the Parties to the Framework Convention on Climate Change in Berlin*, 1995.

WBGU: German Advisory Council on Global Change. Human Progress Within Planetary Guard Rails. A Contribution to the SDG Debate. *Policy Paper No. 8*, 2014.

Mort Webster, Lisa Jakobovits, and James Norton. Learning about Climate Change and Implications for Near-term Policy. *Climatic Change*, 89(1-2), 2008.

Acknowledgments

I am grateful for the dedicated support I received from Prof. Dr. Hermann Held throughout my work on the thesis. Without the many helpful discussions and his interdisciplinary spirit this project would not have been possible. Also I am grateful for the insightful conversations with Prof. Dr. Andreas Lange. Moreover, a special thanks to Barbara Hof, Sandra Scholz, Lukas Stein and Aman Malik for illuminating discussions and helpful comments on the text. Finally, I would like to thank the members of the Research Unit on Sustainability and Global Change for their support.

Selbstständigkeitserklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Studiengang Master of Integrated Climate System Sciences selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, 15. Dezember 2016, Felix Schreyer