Multiple-Objective Decision Making

by

Damian Wassel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2016

Doctoral Committee:

Professor Richmond H. Thomason, Chair
Professor Sarah Buss
Professor Allan F. Gibbard
Professor Sekhar Chandra Sripada
Professor Brian J. Weatherson

For every parcel I stoop down to seize
I lose some other off my arms and knees,
And the whole pile is slipping, bottles, buns—
Extremes too hard to comprehend at once,
Yet nothing I should care to leave behind.
With all I have to hold with, I will do my best
To keep their building balanced at my breast.
I crouch down to prevent them as they fall;
Then sit down in the middle of them all.
I had to drop the armful in the road.
And try to stack them in a better load.
—Robert Frost, "The Armful"

For my grandmother, Marion Emily Wassel.

Acknowledgements

I might never have written a word of this dissertation without the support, intellectual generosity, and boundless patience of my advisor, Professor Rich Thomason. His steadfast insistence that I just *work harder* saw me through the rough patches.

I also received invaluable guidance from my committee. Early on, Professor Allan Gibbard's interest and encouragement helped restore my faith in the project. Professor Sarah Buss made sure I didn't lose sight of the humanity in philosophy, and urged me not to confuse a heap of mathematical formalism for an argument. And Professor Brian Weatherson, who graciously volunteered to help when deadlines loomed, pressed me to figure out which pieces of the project were genuinely important.

Earlier versions of this work were improved with feedback from my colleagues and the faculty of the Philosophy Department, especially the participants in the 2014 and 2015 Candidacy Seminars. I am grateful for conversations with and comments from Daniel Drucker, Professor James Joyce, Jeremy Lent, Cat Saint-Croix, Patrick Shirreff, and Rohan Sud.

The wonderful office staff of the Philosophy Department steered me around countless bureaucratic obstacles, and ensured that I was paid and provisioned with everything I needed. Even more, they were a beacon of humanity, warmth, and humor, shining at the end of the corridor. In particular, I am grateful to Jude Beck and Linda Shultes. I could always count on Jude Beck to put a smile on my face, even while she helped me reserve classrooms, fix copiers, and wrangle troublesome students. And Linda Shultes made me feel as though she was personally invested in my success. She shepherded me through

the program from first day to last, saving me from procrastination and inattentiveness more times than I could count.

Professor Soazig Le Bihan, of the Department of Philosophy at the University of Montana, generously leant me her office for a summer. I did some of my most enjoyable work in that sunny room beneath Mount Sentinel.

My friends, Cat Saint-Croix, Patrick Shirreff, and especially Sara Nitz, helped me to find some joy in the Michigan winters. It would have been a colder time without them.

Above all, I am deeply grateful for the support of my family. My partner, Kimberly Chuang, was there beside me every day, working with me in the trenches. My parents, Amanda Tasi and Damian Wassel, maintained my confidence, my sanity, and at times my bank account, and whisked me off for some much needed vacations from Ann Arbor. My cousin, Nathan Gooden, never gave up on me, even when I had almost given up on myself, and he reminded me to have a little fun now and then. And my brother, Adrian Wassel, helped in too many ways for words: he was my strength when I had none.

Contents

Abstract

Decision makers often approach decisions with a divided mind. Rather than having clear, overall preferences between options, they evaluate them according to many criteria. Worse still, these criteria often conflict in their rankings of options. Such decisions are hard, but it is clear that they can be resolved in better and worse ways. However, decision theoretic accounts of rational decision making have little traction on these realistic cases. This is because a complete, conflict-free preference order over the available actions, representable by a formal structure at least as robust as a weak order is typically understood to be the essential prerequisite for rational action.

There has been very little work on this problem within the philosophical literature; here, I lay the groundwork for an account of good multiple-objective decision-making. The first step is to characterize acceptable methods for constructing overall preference orders from the objective-specific rankings which are accessible to the decision maker. Here, I consider two such methods.

While there has been little work on multiple-objective decision making within philosophy, the problem has received considerable attention in the decision analysis literature. In the first chapter, I argue that decision analytic methods for constructing overall preferences are philosophically well-motivated, and explore how they can be applied to some simple examples of multiple-objective decisions.

In the second chapter, I consider an altogether different approach, which takes at face value the analogy between an individual decision maker trying to reconcile several objectives in her decision and a group of several individuals trying to reach a joint decision.

The thought is that multiple-objective decisions can be modeled as social choices—in the sense of Social Choice Theory. The challenge is that such an approach seems to run headlong into the limiting result of Arrow's Theorem. Against earlier work on this approach, I argue that Arrow's Theorem does not apply to individual decisions.

Introduction

Actual decision makers frequently face decisions that require them to reconcile conflicting commitments, values, or desires, in order to decide between alternatives each of which is better with respect to some commitments and worse with respect to others.

I refer to such decisions as *multiple-objective decisions.*

Facing a multiple-objective decision can be disheartening. Often, it can seem to the decision maker that there are no good ways to reconcile their conflicting commitments, and thus no good choices. Of course, some of our choices are just hard. Sometimes there really are no good ways to move forward. That might be all there is to say on the subject.

But perhaps, instead, the decision maker faced with a seemingly intractable multiple-objective decision is similar to a naïve gambler trying her hand at the tables for the first time. To the naïve gambler, which bets are good and which are bad may be mystifying. The games are intricate and risky. Even as she learns the rules, it may not be obvious to her which features of the game are relevant to the choices she should make, or how she should consider those features when she makes her choices. It may seem to her that she has no good choices. After all, decisions under uncertainty are hard.

However, for the naïve gambler, that is certainly not all there is to say on the subject. Decision theory can help the naïve gambler. Decision theory gives account of which features are relevant in her decisions and of how these features should be considered when she makes a choice. Of course it is not decision theory that makes a given choice good: the goodness of the choice consists in the way it corresponds to her preferences and her beliefs. Nonetheless, decision theory illuminates the complex, uncertain decisions of the

naïve gambler, by providing an account of which choices are good by her own lights.

Perhaps decision theory can be deployed similarly to illuminate multiple-objective decisions, and to give an account of better and worse choices in these complicated cases. I am optimistic that it can be. But there is scarcely any philosophical work on how we might do so. This dissertation explores how we might give a philosophical account of multiple-objective decision making.

If we are to apply decision theoretic accounts of good decision making to multiple-objective cases, there is a considerable obstacle to surmount. Decision theoretic accounts require that a decision maker have at least complete weak preferences over outcomes, or that her choices meet certain consistency conditions, in order to characterize good decision making. The theoretical apparatus has no purchase on decisions where such preferences cannot be identified. The challenge is that decision makers faced with multiple-objective decisions often lack introspective access to such overall preferences between outcomes. Sometimes a decision maker may have partial preferences over outcomes. But often, the best a decision maker can do is say whether a given outcome is better or worse with respect to a given objective.

To gain purchase on multiple-objective decisions, then, we have to give an account of how the necessary overall preferences can be characterized in these cases. Resolving this puzzle is the central task of this dissertation. The two chapters here diverge sharply with respect to the potential solutions they explore.

In the first chapter, I consider the gap between realistic decisions and the sort of formal decision problems that are within the scope of axiomatic decision theory. If axiomatic decision theory is to help us get traction on realistic decisions, we need to be able to specify decision problems that aptly represent the realistic cases. The task of specifying such a representation varies enormously in difficulty. For some realistic decisions there are obvious, ready-made representative decision problems. But in other cases, specifying

a decision problem is frustratingly hard. To illustrate the real scope of this problem, I sketch a rough taxonomy of human decisions, identifying features of realistic decisions that affect the difficulty of solving the specification problem. I argue that multiple objectives pose the biggest challenge to specifying representative decision problems.

Where philosophy is nearly silent on multiple-objective decisions, considerable progress has been made on this topic in the field of decision analysis. I give an overview of decision analysis, and explore its philosophical foundations, arguing that decision analytic methods for specifying decision problems are philosophically well-motivated. In the remainder of the paper, I give a detailed exposition of how decision analytic methods can be used to specify representative decision problems for multiple-objective decisions. I focus principally on a method for analyzing a decision maker's preferences for particular attributes of her available outcomes and tradeoffs between them, through which a weak preference order over outcomes themselves can be determined.

The primary goal of the second chapter is to argue for the possibility of a social choice theoretic approach to modeling multiple-objective decisions. There is an obvious analogy between a decision maker facing a multiple-objective decision, whose objective-specific rankings of outcomes take the form of weak orders, and a group of individuals making a single objective decision, each of whose preferences over outcome take the form of weak orders. In both cases, the goal is to go from several weak orders over outcomes, to a single overall or aggregate weak order over outcomes. However, in cases of group decisions, there is a familiar way to aggregate individual preferences into a single weak order: voting. Suppose we take this analogy at face value. We can conceive of the individual decision maker's various objectives as voters, whose votes are expressed by weak orders they give over the available outcomes. Then we can model an individual multiple-objective decision as a social choice—in the sense of Social Choice Theory. Let a *social choice function* be a function from a tuple of weak orders to a single weak order. And we can determine

an individual decision maker's overall preferences as a social choice function evaluated on her objective-specific weak orders. This approach faces an obvious challenge from Arrow's Theorem. The theorem establishes that no social choice function jointly satisfies four conditions, which are generally understood as necessary conditions on rationally acceptable social choice functions. In other words, Arrow's Theorem entails that there are no rationally acceptable social choice functions. At first glance, this limiting result seems to scuttle any social choice theoretic approach to multiple-objective decisions. Consequently, this modeling approach has been flatly rejected wherever it has been considered. But I argue that this is mistaken. While Arrow's conditions are well-motivated with respect to group decisions, I argue that the so-called *independence condition* does not apply with respect to individual decisions. Thus, I conclude, social choice theoretic models of multiple-objective decisions should be given more careful consideration. To close, I consider some questions for future work on this modeling approach.

Chapter 1
The Specification Problem: Lessons from Decision Analysis on how to Formally Represent Real-world Decisions

1.1 The specification problem

Let's say a *decision* is the sort of occasion for choice we encounter in our daily lives. We face decisions whenever its up to us what to do. Decisions range in complexity from the quotidian—what to wear to work, or where to eat dinner—to the momentous—whom to marry, or which career to pursue.

Let's say a *decision problem* is a formal representation of a decision replete with enough structure that it falls within the scope of Axiomatic Decision Theory (ADT).[1] Typically this formal structure will include at least a well-specified domain of events, a well-specified range of consequences, a well-specified set of actions—understood as functions from events to consequences, and a transitive and complete order over consequences from which a family of utility functions over consequences can be derived.[2]

While decisions are ubiquitous and familiar in the wild, decision problems are mostly hidden away in textbooks and academic papers, and are rarely encountered except by students and professional academics. In rare contexts like casino gaming, we bump into decisions for which aptly representative decision problems are fairly obvious. When we encounter such ready-made decision problems, ADT has obvious and specific practical

[1] Throughout, I'll ignore debates about which particular formulation of ADT is correct, and write as though there were a single, univocal formulation of ADT. This is because nothing in this chapter depends on which formulation of ADT one prefers. Also, I strenuously doubt that any convincing argument could be made that some particular formulation of ADT is the single, correct formulation.

[2] See: Lin 2014, 661–662 for discussion of the standard formal elements of a decision problem in ADT.

value: It recommends a set of optimal choices. Alas, for the most part, the world we navigate isn't furnished with ready-made decision problems.

Instead, our real-world decisions are messy and complicated, having both too little and too much structure—more on this later—to be aptly or easily represented as decision problems. Seemingly, the more momentous a decision, the messier and more complicated it typically is. I'd be underselling ADT if I said it offers no guidance with respect to our real-world decisions. But if there is a general, practicable lesson to be gleaned from ADT, it is disappointingly vague. Raiffa 1968 suggests the upshot of ADT is that we should decide in a manner consistent with our basic preferences for consequences taken together with our basic probabilistic beliefs about the world (xxxiii). Outside the casino, the decisions that can be settled by that maxim are few and far between.

Davidson, Suppes, and Siegel 1957 contend that an ultimate goal of decision theory is to throw light on our everyday decisions (7); Weatherson 2012 echoes this sentiment, charging that "we want decision theory . . . to be applicable to real-life situations." I agree. If we are to throw light on our everyday decisions with ADT, we need to be able to specify decision problems to represent our everyday decisions. Yet, as Lin 2014 notes, "standard decision theory is silent about which specification is the best one for the agent to adopt on a particular occasion (661)." In the contemporary philosophical literature, there is a vast gap between the sophisticated formal apparatus of ADT, and our informal everyday decisions. In this paper, I address how we might close this gap by making progress on the *specification problem*.

My objectives in this paper are fairly modest: First, I outline the gap between our everyday decisions and formal decision problems in more detail. I sketch a coarse taxonomy of our everyday decisions, and indicate where we should expect to find ready-made decision problems, and where things get trickier. Second, I argue that the our preferences in real-world decisions pose the biggest challenge to specification. The way we

evaluate consequences in our real-world decisions—in particular in our most significant decisions—seldom yields an obvious formal representation with a single preference order. Instead, we often evaluate consequences along multiple, potentially conflicting dimensions. Third, borrowing heavily from work in decision analysis, especially Raiffa 1969, Winterfeldt and Fischer 1973b, Keeney and Raiffa 1976, and Keeney 1992, I explore some methods for structuring formal representations of complex, real-world decisions with multiple objectives, discuss the philosophical foundations of these methods, and examine how the resultant representations can be brought in line with the axiomatic constraints of ADT.

The work is here, admittedly, exploratory. The paper raises at least as many questions as it answers, and I don't pretend to offer a general, ideal method for formalizing real-world decisions. To make the chapter accessible to the general philosophical reader, I will keep the formal details in the background. In the main body of the paper, when jargon is unavoidable, I use footnotes to clarify the ideas. I revisit some of the formal aspects in the appendix.

With that said, there are some concepts it's hard to talk about clearly without a regimented formalism, in particular, the difference between value functions and utility functions, conditional preferences, preferential independence, and risk attitudes. If my discussion of these matters is somewhat muddy in the main body of the paper, I strive to clarify it in the appendix.

1.2 The shape of the problem

The gap between our best formal accounts of ideal decision making embodied in ADT and our real-world decisions is a long-standing worry for decision theorists. In the first noteworthy monograph on ADT, von Neumann and Morgenstern 1944 recognize that "the process of mathematization" of any theory, let alone one of human decision, is "not

at all obvious (5)." They are optimistic, though, that factors which initially appear challenging or impossible to measure or formalize will become more tractable as the theory is developed. Drawing an analogy to the theory of heat within physics, they note it was the mathematization of the theory that showed the way to measuring the relevant quantities (3). Given this optimism, it makes sense that von Neumann and Morgenstern 1944 concern themselves principally with developing ADT, and give little consideration to how we should specify decision problems to represent our real-world decisions. Their hope is that theory building will help to show the way to theory application.

But by the time of Savage 1954, the analogy of von Neumann and Morgenstern 1944 to the theory of heat had not altogether panned out. Much work has been done on the measuring of utility functions—perhaps most important among it Mosteller and Nogee 1951, but the formal theory had not enabled much progress on the specification problem. Savage 1954 recognizes the difficulty, noting that our everyday decisions take place in a world too grand to fit easily within his formal account of decision making (83). To attack decision problems, Savage suggests, we must begin by "artificially confining attention to so small a world" that it fits within our formalism. But he doesn't share the optimism of von Neumann and Morgenstern 1944. He laments, "he is unable to formulate criteria for selecting these small worlds," and speculates that no complete and sharply defined general principles governing how this ought to be done can be discovered (16).

Subsequently, the specification problem is scarcely addressed anywhere else in the philosophical literature.[3] For example, the issue isn't addressed at all in the next semi-

[3]Davidson, Suppes, and Siegel 1957 takes up the task of experimentally verifying ADT, or at least laying the groundwork for such experimental verification. The authors express serious concern with the empirical adequacy of ADT, citing two key reasons: first, that people often seem to simply no meet the conditions of the models of ADT, and second that ADT has been given no empirical interpretation on the basis of which its adequacy can be tested. However, the work makes no serious effort to address the specification problem.

I also feel compelled to note that the issue of how to accommodate our every day decisions in both decision theoretic and logical systems has been taken up enthusiastically in the computer science community, in particular within the nonmonotonic logic community and formal artificial intelligence communities. Indeed, there is considerable work on formulating a qualitative decision theory more capable of capturing our everyday reasoning. For more on qualitative decision theory see: Brafman and Tennenholtz 1996 and

nal work of philosophical decision theory, Jeffrey 1965, except for an oblique suggestion that decision theory has a "central heartland" where things are tidy, beyond which complexities lurk (20). Resnik 1987 does discuss the specification problem, principally to suggest it's not a problem of genuine philosophical interest. He also briefly takes up the question of how we should specify states of the world, and our probability function over them, and quickly dismisses the regress problem—that is, whether rationality requires that we always decide how to decide, on pain of vicious regress (6–12). The specification problem also gets some attention in a few recent articles, notably, Weatherson 2012, Lin 2013, and Lin 2014. Weatherson 2012 argues for some constraints on how we capture our real-life probabilistic beliefs with states and a probability function when structuring decision problems; Lin 2013 develops a qualitative decision theory meant to capture everyday decisions; and Lin 2014 addresses the regress problem.

I suspect Savage 1954 is right that we won't find any complete, and perfectly general way to resolve the specification problem; I certainly don't have any general solutions to offer here. In fact, I think the problem is, if not harder, at least bigger even than Savage 1954 suggests, and that its scope and difficulty are not well-understood, even in the recent work discussed above. But for this reason alone, Resnik 1987 is dead wrong that the problem is not of philosophical interest. Whether or not we can identify general principles for solving the specification problem, recognizing the magnitude of the gap between ADT and real-world decisions is of independent philosophical interest. I turn now to sketching a rough taxonomy of human decisions, which will serve to highlight the real complexity of the specification problem.

Doyle and Thomason 1999. For a general introduction to the role of logic in formalizing everyday reasoning within the artificial intelligence literature, see: Thomason 2009 and Thomason Forthcoming. The latter paper is available upon request from the author.

1.2.1 A taxonomy of decisions

I adapt this taxonomy, with additions and simplifications, from work in Wendt and Vlek 1973, Winterfeldt and Fischer 1973b, and Triantaphyllou 2000.[4]

Decisions are often written about as though they are things decision makers (DMs) just happen upon, with particular structural features, on which DMs bring their propositional attitudes to bear. One of the things that becomes obvious, after careful consideration of the specification problem, is that there is a tremendous amount of feedback and interaction between structural and attitudinal features, and the two are not obviously or easily separable. Therefore, I take a decision to be the whole package of structural features constrained by the decision context, and attitudes brought to bear by the DM. Where appropriate, I'll flag features of a decision as either attitudinal or structural. Some features of a decision do not fit neatly into either category.

We can distinguish our real world decisions along at least the following dimensions: character of beliefs, plasticity, uncertainty, time variability, number of stages, and number of objectives. I discuss each in turn.

DMs face decisions in real-world contexts. In some cases of decision, the DM has explicitly quantitative beliefs about how things might turn out in her decision context. For example, if the DM is considering how to bet in a game of craps, she may believe that the dice each have six sides, allowing for a total of thirty six possible combinations, with eighteen distinct rolls under the rules of the game, two of which will constitute a win for some particular bet. In other cases, the DM has only purely qualitative beliefs about how things might turn out in her decision context. Lin 2013, 832, discusses the example of a DM considering whether to purchase groceries today or tomorrow. The DM is likely to have some basic, purely qualitative beliefs about when the store is open , say, that it is

[4]See: Wendt and Vlek 1973, 4; Winterfeldt and Fischer 1973b, 50–53; and Triantaphyllou 2000, Ch. 1. Winterfeldt and Fischer 1973b was originally published as Winterfeldt and Fischer 1973a and is available in that format upon request from the Library of University of Michigan, Ann Arbor.

likely open today, and probably not open tomorrow. In many cases, the DM has relevant beliefs of both kinds about her decision context. Of course, Bayesian epistemologists are always keen to formalize qualitative beliefs with quantitative credence functions, but this has to be seen for what it is: a formalization of basic beliefs that were not obviously quantitative. So we can distinguish decisions with respect to character of beliefs as purely quantitative, purely qualitative, or mixed. Character of belief is obviously an attitudinal feature of decisions.

In some cases, the set of possible choices open to the DM is rigid. That is, the apparent options are the only options, and this is not open to change. When betting on a game of craps, the available actions are rigidly fixed by the rules of the game. But in other cases, the DM can discover or construct additional options by reflecting on the decision context and her attitudes more carefully. When deciding what to cook for dinner from the ingredients in the pantry, the set of choices is plastic. Reflection and imagination can generate new actions. Obviously, plasticity comes in degrees, and it can be observed even in highly structured games. Subjectively, the choice of strategies in poker is more plastic than the choice of bets in craps, and the choice of what to cook is more plastic than the choice of moves in any casino game. So we can distinguish decisions with respect to plasticity as perfectly rigid, or plastic to any degree, and then further with respect the degree of plasticity. It's unclear whether plasticity should be conceptualized as a structural feature or an attitudinal feature.

Decisions also vary with respect to their riskiness. Sometimes all actions have fixed consequences, no matter what the world is like. In other cases the consequence of actions are uncertain, and dependent on how the world turns out to be. So we can distinguish between decisions under certainty, and decisions under uncertainty. Crucially, uncertainty—often modeled as a structural feature of the decision—must instead be understood as an attitudinal feature of the decision, if ADT is to have any normative import

for actual DMs. In other words, certainty must be understood as a subjective, epistemic notion, not an objective, metaphysical notion. Thus a decision is uncertain as long as some relevant features of the decision context on which the consequences of her actions depend aren't determinately fixed by her beliefs, even if those features of the decision context are determinately fixed by the objective state of the world. This isn't meant to glibly dismiss the longstanding debate between Bayesians and objectivists about probability. I have no dog in that fight. Rather, it stems from the same intuitions about what it means for a model to be normatively relevant that motivate the overall agenda of this paper. To be normatively relevant, the model must capture the key features of the DM's subjective experience of decision making. So it must not treat as certain a decision that is subjectively uncertain, or treat as uncertain a decision that is subjectively certain.

Sometimes consequences are time-invariant. That is, the consequences of all actions are received by the DM, or at least determinately fixed, at the same time. For example, on a bet determined by flip of a coin, the consequences of all betting actions are paid out—or at very least fixed—when the coin lands. Note that the consequences of the actions need not be received immediately, just all at some uniform, possibly future time, in order for the decision to count as time-invariant. But in other cases, the consequences of some actions are received at different times than others. For example, in a choice between job offers, some jobs may have different start dates. Time variability occurs even in highly-structured games. In craps, some bets are won or lost on the next roll of the dice, but others are won or lost only when the round comes to a close, so some actions' consequences are determined at different points than others. So we can distinguish decisions with respect to time variability as either time-invariant, or time-variable.

We can also distinguish single-stage decisions from multi-stage decisions. In a single-stage decision, each of the initially available actions leads directly, though perhaps uncertainly, to consequences without requiring a further choice on the part of the DM. In a

multi-stage decision, some actions themselves lead to further occasions for choice. Savage 1954, of course, suggests ways that we can conceive of multi-stage decisions as single-stage decisions between more complicated actions (13–17). But the aim of this taxonomy is to distinguish decisions in terms of the features they seem to have from the perspective of the real-world DM.

Finally, in some cases the DM has a single objective guiding her choice and in other cases, she has many objectives. Throughout, I'll adopt a very permissive definition of *objective*; an *objective* is anything the DM understands to be at stake in, or hopes to achieve through, her decision. This definition obviously includes the sort of values we typically think of as in-play in our decisions—say, moral, aesthetic, or epistemic value—while also allowing for more mundane and practically directed objectives—say, cost, or comfort, or wait time. Objectives, then, correspond to features of consequences of actions—Kantian nonconsequentialist ethical theories notwithstanding.

Objectives typically have an *orientation*. Roughly, for positively oriented attributes, the DM aims to be maximize some feature of the consequences of her actions. For negatively oriented attributes, the DM aims to minimize some feature of the consequences of her actions. She may aim, for example, to be maximally moral, or to minimize wait time. So, we can distinguish single-objective decisions from multiple-objective decisions, and we can further distinguish multiple-objective decisions by the number of objectives in play. Crucially, in multiple-objective cases, these objectives may *conflict* with one another, in the sense that realizing one may require compromising another.

Thus, we have the following grid of possibilities for a decision: beliefs about the decision context may be purely quantitative, purely qualitative, or mixed; the decision may be rigid or plastic; the actions may be certain or uncertain; the consequences may be time-invariant or time-variable; the decision may be single-stage or multi-stage; and the decision may be single-objective or multi-objective.

1.2.2 Specifying decision problems: From ready-made decision problems to trickier cases

Within this grid, we can locate cases which seem to come with ready-made decision problems, and cases for which specifying decision problems is frustratingly hard.

It is obvious that decisions are easiest to model when they involve purely quantitative beliefs, and are rigid, certain, time-invariant, single-stage, and single-objective. In these cases, formal decision problems which aptly represent the decision are so obvious as to be indisputable. Because it takes place under certainty, all that seems to be required to formally model the decision, and characterize rational decision making behavior, is a set of options and a weak preference order over those options. From these, we can quickly specify the rational choice as any of the most preferable options from the set. Indeed, a complete, conflict-free preference order over the available actions, representable by a formal structure at least as robust as a weak order is typically understood to be the essential prerequisite for rational action.[5] There is an obvious argument motivating this prerequisite: These seem to be the simplest sorts of decisions we can encounter. In order for there to be a set of best options in the eyes of the DM, she has to have at least a weak preference order over her options, therefore having a weak preference order over her options is a necessary condition for rational action on the part of the DM. Thus, it is typically taken for granted that a DM will have such a weak preference order over her options. Some decision theorists like early Savage, simply assume that rational agents will have

[5]For the formally disinclined reader: A weak order, $\prec$, is a binary relation—that is, a two-term relation—over a set S, that is *reflexive*, *transitive*, and *total* (or, *complete*, or *connected*.) A relation is reflexive if and only if, for every element a of S, $a \prec a$. A relation is transitive if and only if, for every triple of elements, a, b, c, of S, if $a \prec b$, and $b \prec c$, then $a \prec c$. A relation is total if and only if, for every pair of elements a, b of S, $a \prec b$ or $b \prec a$. An obvious example of a weak order is the relation of *less than or equal to* over the natural numbers; a more colloquial example is the relation of *at least as tall as* over the set of people.

Weak orders are especially helpful when constructing mathematical models, because whenever we have a weak order over a set, we can generate a function between elements of the set, and the real numbers, which respects this weak order. (In fact, we can generate many such mappings.)

such preferences.[6] Others, like Sen or Hammond, derive this from other assumptions about rational choice functions, or consistency in choices, together with an account of the relationship between choices and preferences.[7]

It is equally obvious that the the majority of our interesting, real-world decisions are not so tidy. Real-world decisions are almost always taken under conditions of uncertainty. However, provided beliefs remain quantitative, and the decision remains rigid, time-invariant, single-stage, and single-objective, obvious formal representations of the decision still suggest themselves. In these cases, we begin with a set of mutually exclusive states and a probability function over these states determined by the DM's quantitative beliefs about the decision context, a set of actions understood as functions from states to consequences, and a weak order over the set of possible consequences determined by the DM's weak preferences over those consequences. Each of these components is an obvious and indisputably apt representation of the real-world decision. Then, using techniques like those initially suggested in von Neumann and Morgenstern 1944 and later refined in works like Mosteller and Nogee 1951, we elicit a *utility function* over the options available to the decision maker. While the weak order over consequences constrains the space of admissible utility functions, it is not itself sufficient to determine a specific utility function. Eliciting utility functions requires collection of further information from the DM, in particular about her preferences for certain two-value lotteries of known probability and particular consequences.[8] Here, a *two-value lottery* is understood as an action that returns one consequence value with some probability, and the other consequence value with the complementary probability. Though eliciting a utility function is nuanced and tricky, it is a familiar and well-understood technique within the philosophical litera-

[6] See, for example, Savage 1954, Chs.1–2

[7] See, for example, Sen 1970, P. J. Hammond 1988, or Sen 1993.

[8] I discuss utility functions further in the appendix, including the difference between a simple value function over a space of consequences, and a utility function. I also discuss how utility functions can be elicited.

ture, so I don't go into details here.[9] Given a utility function, we can then identify the set of rational choices for the decision maker.

Even in cases where consequences are time-variable, and the decision is multi-stage, fairly obvious representations can be structured, by identifying a rate of exchange between consequences realized at different times,[10]and then applying techniques of averaging out and folding back[11] to represent the decision as a time-invariant, single-stage decision, for which rational choice behavior can again be characterized by eliciting a utility function corresponding to the DM's weak preferences over consequences.

Cases in which the decision maker has purely qualitative beliefs, and the decision is plastic, are somewhat trickier. Still, the formal representations feel familiar once one has a handle on the formal representations just discussed. Quantitative beliefs representing qualitative beliefs can be identified using familiar techniques again involving the DM's preferences with respect to certain two-value lotteries over consequences of known utility.[12] Plasticity can be dealt with either by representing the decision as multi-stage, with the first one or more stages understood as choices between distinct sets of actions, or by modeling the set of actions as a fuzzy set. Roughly, fuzzy sets are sets in which the elements vary with respect to degrees of membership: some elements can be clearly in the set, while others are only fuzzily in the set.[13] To model plasticity, we can represent the set of actions with a fuzzy set in which obvious actions have a high degree of membership, and less obvious actions have a lower degree of membership.[14] From there, we can again elicit a utility function for the DM, and—notwithstanding the additional subtleties that emerge from the double layer of uncertainty introduced if we elect to model plasticity

[9]For further discussion of this see: Savage 1954, Ch. 5, Jeffrey 1965, Resnik 1987.

[10]See: Keeney and Raiffa 1976, Ch. 9 for a detailed discussion of discounting techniques over so-called btime streams of consequences.

[11]See: Raiffa 1968, Chs. 1–2 for a detailed discussion of averaging out and folding back applied in several very tractable example cases.

[12]For further discussion of this see: Savage 1954, Ch.3, and Resnik 1987, Ch.5.

[13]See: Klir and Folger 1988, Ch. 1.

[14]See: Tzeng and Huang 2011, Chs. 1–3.

using fuzzy sets—define rational choice behavior in the usual way.

In uncertain decisions, structuring a representation of the decision sufficient to characterize rational choice behavior depends on being able to describe a utility function for the DM over her set of available consequences. This, in turn, depends on the DM having clear weak preferences representable as a weak order. Weak preferences are equally essential in aptly formally representing any decision under certainty. But such weak preferences can only be assumed to be clear because, thus far, we have considered only single-objective decisions.

1.2.3 The heart of the problem: Decisions with multiple objectives

The majority of real-world decisions, in particular, our momentous life-changing decisions, involve qualitative beliefs, and are plastic, uncertain, time-variable, and multi-stage. Thus, they are generally toward the trickier end of the spectrum. More importantly, though, they are typically multiple-objective decisions. For example, when choosing between careers, or between automobiles, or even between caterers for a wedding, there are always many things at stake in the decision. For that matter, many decisions that are typically thought of as single-objective—say betting decisions while playing a casino game—may be better conceived as multiple-objective decisions. Not every casino gamer plays with the sole objective of winning money. Some players also hope to have fun, take out-of-character risks, get free drinks, or catch the attention of the attractive stranger across the table.

Wherever there are multiple objectives, there is the potential for conflict. The job with the highest pay is unlikely to be the job with the least responsibility. The fastest car might not be the most fuel efficient. The best reviewed caterer might not offer the most interesting menu. The boldest, most exciting bet may not be the bet with the highest expected payout. The potential for conflict once multiple objectives are introduced

complicates things enormously. When objectives conflict, even minimally, a DM may not have clear overall preferences over her available options. Thus, her preferences may not be obviously representable as a weak order. In good cases, she will at least have objective-specific weak preferences over her options. That is, in good cases, holding some particular objective fixed, she will be able to determine whether any of her options is at least as good at achieving that objective as any other. Even so, her evaluations over her options will be representable not as a single weak order, but instead as a tuple of weak orders, each corresponding to one of her objectives. It is for this reason that I said above that in many of our everyday decisions we have both less and more structure than is required to specify a clear, representative decision problem. In multiple-objective cases we sometimes lack obvious overall weak-preferences—there is too little structure—but we instead have many objective-based weak orders over our options—there is too much structure.

As we saw above, formally modeling even the simplest decisions taken under certainty in such a way that the decision problem is resolvable with ADT requires that we can represent the DM's preferences with a weak order. And representing the DM's preferences as a weak order remains essential to modeling increasingly complex cases involving uncertainty, since the weak order is required to determine a utility function for the DM. Thus, without some way of moving either from a tuple of objective-specific weak orders to a single overall weak order over options, or directly from a tuple of objective-specific weak orders to a utility function over options, we find ourselves unable to get any purchase on multiple-objective decisions with the formal apparatus of ADT. Decisions with multiple objectives, then, are at the very heart of the specification problem.

Given the ubiquity of multiple-objective decisions in our everyday lives, it is somewhat surprising that they are discussed scarcely anywhere in the philosophical literature. Ellis 2006 singles out multiple-objective decisions as an altogether "neglected problem" in the philosophy of decision making, but himself offers no positive suggestions for how

it might be resolved (313).[15] The problem is discussed somewhat in the essays collected in Elster 1985b—in particular in Steedman and Krause 1985—and Kavka 1991. Steedman and Krause 1985 and Kavka 1991 both explore the possibility of using social choice theoretic techniques to aggregate a tuple of weak orders into a single weak order. Ultimately, both dismiss these techniques as dead ends because of the limiting result of Arrow's Theorem. I disagree, and take up this issue in my second chapter. Otherwise, the philosophical literature on decision making is silent on the issue.

Outside philosophy, however, the problem has been addressed extensively within the decision analysis literature. In the remainder of the paper, I will explore and defend the philosophical legitimacy of using decision analytic techniques to get formal purchase on multiple-objective decisions.

1.3 What is decision anaylsis?

It is safe to say that philosophy, as a discipline, has had very little interaction with decision analysis. In terms of academic geography, decision analysis is a research subfield principally of interest to researchers in operations research, decision and control, management sciences, behavioral economics, artificial intelligence research, and cybernetics. Decision analysis emerged in the 1960s, spurred along by the pioneering work of Robert Schlaifer, Ronald Howard, and most centrally Howard Raiffa who, to the best of my knowledge, first coined the name 'decision analysis.'[16] Raiffa, who began his career in operations research, then completed a Ph.D. in pure math, sought to apply the mathematical framework of ADT to the complex and practical problems he had worked on as

[15] He does explore two potential avenues of progress, the decision analytic methods discussed in this chapter, and the social choice theoretic methods discussed in the third chapter of this dissertation. But he dismisses both out of hand without much in the way of argument.

[16] Their key early works on the subject include Schlaifer 1959, Raiffa and Schlaifer 2001, Pratt, Raiffa, and Schlaifer 1964,Howard 1966, Raiffa 1968, andRaiffa 1969. For more on the origins of decision theory as a discipline: Raiffa 2002 and Keeney 2006.

an operations researcher.[17]

It is not surprising that there is no single, widely accepted, clear, and non-trivial creed of decision analysis. But one gets the sense that decision analysts are quite used to being asked to explain themselves. Researchers in decision analysis are unusually concerned with and candid about the aims and scope of their research. Monographs, handbooks, and introductory texts alike typically include lengthy discussions of these themes.[18] From these we can distill a fairly brief characterization of the field. Research in decision analysis focuses on the paradigm of choice in uncertain environments, and is generally concerned with prescribing practicable methods for making *good* choices in complicated, realistic cases. The research is often grounded in empirical science, and is driven by case-studies of troublesome, large-scale institutional decisions.

Decision analysts generally recognize three kinds of accounts of decision making.[19] The first are *normative* accounts of decision making, which characterize *ideally rational* decision making behavior. ADT is the standard normative account of decision making within philosophy. But normative accounts only retain obvious normative force with respect to those cases in which the gap between a real-world decision and its formal representation as a decision problem is so narrow as to be disregarded. The second are *descriptive* accounts, which characterize the behavior of actual decision makers. The third are *prescriptive* accounts, which characterize *good* decision methods for resolving real-world decisions. Decision analysis aims to generate a prescriptive account of real-world decision making under uncertainty.

[17] In particular, he was inspired by the work of the mathematician Abraham Wald, who in Wald 1947 had developed the statistical technique known as *sequential analysis*, in which data is analyzed as it is gathered, rather than after it has been collected in its entirety. In Wald 1950 he then attempted to apply these techniques to decision theoretic problems. No doubt Wald would have continued to be at the leading edge of decision analysis research had he not died in a plane crash in 1950.

[18] See, for example: Raiffa 1968, ix–xii; Keeney and Raiffa 1976, vii–xi; Clemen 1996, Ch. 1; Tzeng and Huang 2011, Ch. 1; or Parnell et al. 2013, Chs. 1–3.

[19] See Becker and McClintock 1967, 239–241, Raiffa 1968, x–xi; Fischer 1973; Keeney and Raiffa 1976, vii; and J. S. Hammond, Keeney, and Raiffa 1999, Ch. 1.

Prescriptive accounts are, by definition, not ideal accounts. Thus they lack the obvious philosophical luster of normative accounts. They lack the precise formal systematicity and coherence of ADT, and do not reduce cleanly to a few axiomatic principles. Decision analytic accounts of decision making fall short of being ideal accounts, because decision analytic methods do not uniquely generate a single best formal representation of any given decision. For that matter, it's somewhat misleading to suggest that there is a singular, univocal decision analytic method. Rather, there are a variety of techniques that can be applied to a particular decision in a variety of ways to generate multiple distinct formal representations, or multiple distinct regimentations of the same formal representation. Given that there are many different admissible formal regimentations of the decision, it is no longer reasonable to assume that a single set of best decisions will be identified. To be clear, though, each model generated by the decision analytic methods below will identify an indifference class of best actions according to that model. The issue is that the models themselves, or their precise regimentations, are non-unique.

Despite this, decision analytic accounts of decision making remain philosophically interesting for four key reasons. First, decision analytic accounts of good decision making embrace the central idea of ADT, that we should decide in a manner consistent with our basic preferences for consequences, taken together with our basic probabilistic beliefs about the world. The challenge, from the perspective of decision analysis, is to describe these preferences and beliefs in a clear and robust enough way that this belief-preference consistency requirement actually has some traction on our decisions. Thus, decision analysts are especially interested in methods for resolving the specification problem, and this is philosophically interesting in its own right. This amounts to the second reason. Third, decision analytic methods can, at bottom, be thought of as techniques for investigating the DM's beliefs and preferences. In other words, they are not tools for imposing structure on unstructured beliefs and preferences, but rather are tools for discovering

preexisting structure in the DM's subjective attitudes about the decision and extending it so that ADT has purchase on the decision. Thus, decision analytic accounts remain essentially subjective accounts of good decision making, where a good decision is picked out by the relation it bears to the DM's attitudes. Fourth, as we'll see below, case studies in decision analysis reveal feedback between a DM's preferences and decision analytic attempts to formally represent them. This suggests that the loosely Humean idea that our preferences are unassailable or unalterable through acts of reasoning sometimes advanced in contemporary philosophy is false.[20] As it turns out, the heart may not simply want what the heart wants. While we may find ourselves nudged this way or that by brute desires, urges, or appetites, these seem not to be the same thing as all-things-considered preferences. It is through the latter that our dispositions to act are often mediated, and where the former may be unalterable through acts of reasoning, the latter are not. After all, if there are such things as all-things-considered preferences, then consideration enters into the picture somehow, and there may be better and worse ways of carrying out said consideration. I'll say a bit more about these matters below, when I reply to some objections.

Decision analytic methods for formally characterizing decisions typically involve three phases: option analysis, uncertainty analysis, and preference (and utility) analysis. In the option analysis phase, the analyst or the decision maker (sometimes the two are the same) carefully inventories the options available. In the uncertainty analysis phase, the analyst gathers data about how uncertainty affects the relationship between the choices available to the DM and their consequences. In the preference analysis phase, the analyst investigates the DM's reactions to potential outcomes, identifying those features of the outcomes

[20]This idea is sometimes formulated thus: All reasons issue forth from desires, and these desires themselves are not within the scope of reason. See, for example, Schroeder 2007. I don't take issue with this point, per say. Rather, the claim is that the way desires give rise to reasons for action is frequently mediated by acts of reasoning. Roughly, which reasons we are given by our desires depends on how we reason about them to arrive at all-things-considered preferences.

that are relevant to the decision, and exploring how they affect the DM's preferences over them. Naturally, there may be feedback between the phases.

Here I focus on methods for preference analysis in multiple-objective cases. Before delving into that further, I reply to some worries for this approach.

1.4 Three worries: the end of rationality, conflict means inconsistency, the problem of regress

Three foundational objections arise at this juncture: We might worry that if decision analytic methods do not uniquely select formal representations of decisions, then the choices they recommended can no longer be thought of as rational. Independently, we might worry that no good account of decision making under conflict can be given, because such conflicts are indicative of problematic underlying inconsistencies in the DM's attitudes. Then there is the problem of regress: if there are several ways we might formally represent a decision, then we need to make a choice between those representations, but this choice is itself a decision with many possible representations, and so on down the rabbit hole of regress. I elaborate on and reply to each of these objections below.

1.4.1 Worry No. 1: The end of rationality

Above I noted that decision analytic methods do not uniquely generate a single best formal representation of a decision, thus, it is a mistake to suggest that they select a single set of best decisions. Because of this, when applying decision analytic methods, we can no longer talk in terms of rational choices or optimality. In a sense, once we entertain methods with non-unique solutions, we've come to the end of rationality.

The only reply here is to bite the bullet. We do, in fact, lose access to the concepts of ideal rationality and optimality when we consider methods with non-unique solutions. But if the worry is that we somehow fall off a theoretical cliff, and are left with nothing

interesting to say about decision making, the worry is misplaced. Many scholars who find themselves thinking carefully about realistic decision problems come to think that fixation on the ideal of rationality is misplaced and unhelpful. This is evident in the work of Herbert Simon, Jon Elster, George Ainslie, Christopher Cherniak, Stuart Russell and Eric Wefald, Gerd Gigerenzer and Reinhard Selten, and Marvin Minsky.[21] Nonetheless, they all insist and their work stands as testament to the fact that we can still talk cogently about good and bad decision making.

Decision analysts are the first to admit that applying ADT to complex, real-world cases is more art than science.[22] But this does not strip us of the ability to talk about better and worse applications of the theory, or better and worse resolutions of the decision. Elsewhere, philosophy does not balk at taking up questions of what makes for good art. And it should not in this domain. In fact, we may be better equipped to talk in terms of good and bad with respect to decision making, than in most other areas. Here we have a clear ideal theory: ADT. We also have clear ideal cases: those decisions representable with ready-made decision problems. We also have a clear general principle to accord with: decide in a manner consistent with our beliefs and preferences. Further, we can identify other more specific principles that constrain the methods we apply to analyze multiple-objective cases. Triantaphyllou 2000, 178, suggests two: First, any method for resolving a multiple-objective decision should concur with the ideal theory when applied to a single-objective case. Second, any method for resolving a multiple-objective decision should still select the same choice when some non-best option is replaced by another option that is strictly worse in every respect.

If the best we can aspire to is an aesthetics rather than a science or a logic of decision, these are unusually sturdy foundations for an aesthetics. Losing access to the concept of

[21] See, for example: Simon 1982, Elster 1985a, Ainslie 1985, Cherniak 1986, Russell and Wefald 1991, Ainslie 2001, Gigerenzer and Selten 2002, and Minsky 2006.

[22] See: Raiffa 1969, 239–242.

ideal rationality doesn't scuttle further investigations of good and bad decision making; we can put this objection to rest.

1.4.2 Worry No. 2: Conflict means inconsistency

A cursory review of Gowans 1987, Mason 1996, or Baumann and Betzler 2004 reveals that philosophers interested in decision under conflict have focused overwhelmingly on cases patterned after the now-classic example from Sartre 1946. Sartre presents a case of a young man, torn between patriotic loyalty and filial duty, who must choose between leaving home to join the resistance, and staying behind to tend to his ailing mother. Decisions like this are hard. They are so hard that many philosophers—among them Socrates in Plato's *Euthyphro*,[23] Jean-Paul Satre,[24] W.D. Ross,[25] R.M. Hare,[26] and Kurt Baier[27] have suggested that the only way for an agent to resolve such decisions is to qualify, relax, or get rid of one or other of the value commitments that led her to conflict. Not only is there no ideal way to resolve such decisions, our best philosophy seems to indicate that there aren't even any good ways. Any decision will result in tremendous loss of value and unavoidable regret.

This intuition that decisions under conflict can only be resolved by relaxing or abandoning commitments to one or more of the conflicting objectives is no doubt related to the belief that systems of commitments (whether to objectives or values) that can give rise to conflict are problematically inconsistent, cannot serve as bases for reasonable decision

[23] See *Euthyphro* 7E to 8E (Plato 1997, 7–8.)

[24] Nussbaum 1985 and Davidson 2001b characterize Sartre as suggesting that we avoid such conflicts by improvising our decisions, rather than allowing ourselves to be bound by principle, and plagued by regret.

[25] See Ross 2002.

[26] As Williams 1987, 121, notes, Hare 1952, 50, suggests that when faced with conflicting principles, the agent is to revise or modify as many principles as it takes to eliminate conflict in the case at hand.

[27] As Davidson 2001a, 34, notes, Baier 1958 allows "only one ultimate moral principle" and "holds that in cases of conflict between principles, there are higher-order principles that tell which principles take precedence." In other words, Baier 1958 argues that whenever we are faced with an apparent conflict in (moral) values, one of these values is to be respected over and above others. Thereby, he suggests that our objectives are always arranged in a lexicographic order.

making, and hence are to be avoided. For example, when discussing moral dilemmas Davidson 2001a suggests that if we allow systems of moral principles which can come into conflict, we must altogether give up our ordinary conception of practical reason.[28] Brink 1996 argues similarly that given certain deontic principles we cannot countenance conflicts of all-things-considered obligations on pain of paradox. Both authors suggest that if we are to proceed with practical reason as usual, we must have a system of commitments that allows at worst prima facie conflicts.

We might worry, then, that there simply can't be any account of good multiple-objective decision making in any case in which the objectives come into conflict with one another. But I think we can dismiss this worry for four reasons. First, and most important, even if conflicts spring only from deficient systems of commitment, this doesn't throw much light on what a DM should do when she's faced with such conflicts. Suppose conflicts only come from deficient systems of commitment, and a reasonable agent should respond to these conflicts by revising her system of commitments. Surely some ways of qualifying, relaxing, or giving up commitments are better than others. Surely some tradeoffs are reasonable, and some are just crazy. For example, compromise options are frequently available that respect many of the conflicting commitments to some degree. When such options are present it's clearly bad practice to ignore them, but this is not ruled out by the directive to revise one's commitments. Even when faced with a simple dilemmatic conflict like Sartre's case, it doesn't seem reasonable for an agent to simply give up and do nothing at all. If Sartre's young man simply throws up his hands, and wanders off into the sunset he's done something wrong, by his own lights. Yet this is the most obvious way out of the conflict. Merely counseling a DM to avoid conflict, and where she encounters it to eliminate her conflicting commitments, doesn't steer her away from this extreme option, or give her any guidance regarding good decision making behavior. Philosophy can

[28] See Davidson 2001a, 24; see Marcus 1980 for discussion.

clearly do better in its age-old advice-giving role.

Second, conflicts come in considerably greater variety than the philosophical literature allows. Cases in which exactly two commitments pull in opposite directions with respect to exactly two choices are not the norm. Far more common are more complex cases in which the decision maker has several objectives, and several options. Perhaps there is a background assumption that the two-commitment, two-option case is the simplest, and that if we cannot resolve this case well, then surely more complex cases will be out of reach. But this assumption is naïve. Part of what makes cases like Sartre's challenging is that there are only two options, each of which is militated against by one of the objectives. But another part of what makes the case challenging is that there are only two objectives; the addition of a third objective might help settle the case if the two options were not equivalent with respect to that objective.

In an often discussed letter directed to a vexed Joseph Priestly, Benjamin Franklin advocates that tough decisions can be resolved by marshaling our reasons for and against each option, and canceling out reasons of equal weight, until, if we are lucky, one clear choice remains.[29] Suppose Sartre's young man also had a young daughter for whom his mother couldn't provide adequate care while he was away. Then going off to war would have only one mark in favor and two against, while staying home would have two marks in favor and one against. It's safe to assume his commitment to his filial and patriotic duties are of comparable strength, otherwise Sartre's case wouldn't have been much of a bind to begin with. Following Franklin's method of canceling out reasons of comparable weight, we're left with one reason in favor of staying at home—care of his daughter, and no reasons against, and one reason against going off to war—the same reason, and no reasons in favor. Thus, the choice seems clear: he should stay at home. This is not to say that we can or should resolve all of our tricky decisions with Franklin's method.

[29]See Franklin 1772. The letter is discussed in detail in J. S. Hammond, Keeney, and Raiffa 1999, 84, and also Horty 2012, Introduction.

Rather, the point is that, despite becoming more complex, the decision actually becomes easier to resolve when a third objective is introduced. As we will see in more detail below, additional complexity, rather than impeding decision making, instead sometimes provides enough structure to identify a good choice.

Third, I think that the tendency to blame conflicts on problematically inconsistent systems of commitments is in part due to the lack of any decent philosophical account of good decision making under conflict. The reasoning seems to run like this: The supposition that the underlying systems of commitment are flawed explains and excuses the lack of progress toward an account of good decision making under conflict. Conflict reveals problematic underlying inconsistencies. These inconsistencies preclude good decision making in these cases. Therefore, there can be no account of good decision making in these cases, so it's no wonder we haven't found one. But this chapter, and this dissertation more broadly, lays the groundwork for such an account. Thus, I'm unwilling to take for granted that conflicts stem from deficient systems of commitment.

Finally, I think Marcus 1980 argues successfully against Davidson 2001a that the occurrence of conflicts does not indicate underlying inconsistency, and Horty 2012 argues successfully against Brink 1996 that allowing conflicts between all-things-considered commitments does not give rise to paradox. Getting into the specifics of the Davidson-Marcus and Brink-Horty debates would take me rather far afield, and in light of the other reasons just adduced, I don't think such a digression is necessary in order to disarm this objection.

1.4.3 Worry No. 3: The problem of regress

Much like the problem of multiple-objective decisions, the regress problem has not received much attention within the academic literature on decision theory. Indeed, just as Ellis 2006 flagged the topic of multiple-objective decisions as a neglected problem

within the philosophy of human action, Conlisk 1996 complains that regress problem of deciding how to decide has not been adequately addressed.[30]

Lin 2014 gives a clear and pointed presentation of the problem, that is especially relevant for my purposes in this chapter. When confronted with a decision, to settle on a formal representation of a decision is to settle on a way of deciding. In any case in which a DM is not satisfied with simply fixing the elements of a formal representation without further consideration or deliberation, it's open to consider how she should settle on a formal representation of that decision. Lin writes:

> And she may deal with the problem as a second-order decision problem, in which one decides among many various fixations of the elements in order to address the first decision problem. This opens the door to higher-order decision problems, leading to a regress. In daily life we stop the regress. The question is what would make it rational for us to stop the regress. This is what I call the regress problem of deciding how to decide. (662)

Lin argues that the problem of regress is serious enough to completely undercut Bayesian models of rationality, because no way of choosing between decision problems will count as rational by the Bayesian's own lights. He argues similarly against the sorts of adaptive rationality accounts advanced within the bounded rationality literature, though he argues that adaptive rationality comes closer to providing a satisfactory solution to the problem.[31] Ultimately, Lin only suggests the beginning of a solution to the problem, cashed out in terms of the notion of *goal-conduciveness* invoked in work on adaptive rationality. Call a method of resolving a decision *robustly goal-conducive* if that method would actually help the DM "achieve her actual goals in every situation similar to the actual situation (666)." Lin argues that for it to be rational for the DM to stop the regress of deciding how to decide, it is necessary that she *not* believe that the method she elects is

[30] The problem is discussed only in a handful of places that I'm aware of, among them: Johansen 1977, Resnik 1987, Smith 1991, Conlisk 1996, and Lin 2014. Lin 2014 offers helpful summaries of the views in the other articles.

[31] For an overview of the idea of adaptive rationality see: Gigerenzer and Selten 2002, Introduction.

not robustly goal conducive. Beyond this, though, he admits its unclear how to specify further conditions that would be jointly sufficient to rationally stop the regress.

As it turns out, biting the bullet above and backing away from the notion of ideal rationality eases the burden of responding to this objection. When we shift our interest away from ideal decision making to better and worse decision making, the question then becomes, not what would make it rational for us to stop the regress, but which ways of stopping the regress are better, and which are worse. Thus, we no longer need to achieve some perfect recursive equilibrium where the decision method recommends itself, in order to forestall the regress. Instead, it is enough to say of some particular method that it is a good way to stop the regress. From there, it seems that a belief that the decision making method is robustly goal-conducive is sufficient to forestall the regress. The key point is that when we're no longer worried about ideal rationality, we only need good enough reasons to stop the regress, and good enough reasons are far easier to come by than ideal reasons. For example, we can invoke some or other heuristic to forestall the regress. Most obviously, we could employ a satisficing heuristic constrained by the foundation principles principles discussed in §4.1 above, according to which we'd stop the search for a decision method once we'd hit on one that met these constraints.[32] So, we can also close the door on this objection.

1.5 Decision analytic methods for structuring multiple-objective decision problems

Fischer 1973 writes:

> . . . almost all decisions in fact involve multiple criteria, and these criteria are often subjective in nature, eluding easy quantification. The essence of good decision making in such circumstances lies in trading off one goal against another. Mathematical decision making models can be properly applied in such situations only if these trade-offs can be expressed in quantitative form. (19)

[32] The notion of satisficing as a non-optimizing means to resolve decisions originates in Herb Simon's work on bounded rationality. See: Simon 1982. Gigerenzer and Selten 2002 discuss a variety of so-called "fast and frugal" heuristics we might employ to forestall regress here.

I turn now to methods for structuring such models. Recall that decision analysis typically divides into three phases: option analysis, uncertainty analysis, and preference analysis. Here I focus principally on methods of preference analysis, since these are obviously the techniques by which multiple objectives will be accommodated. As noted above, there is considerable feedback between the preference analysis phase and other phases of analysis, and I'll call attention to some of that below.

The key method for modeling preferences in a multiple-objective decision has its first thorough presentation in Raiffa 1969, but originates in the independent work of Miller 1966a and Manheim and Hall 1968. I'll refer to the method as *hierarchical decomposition*.[33] Since Raiffa 1969, hierarchical decomposition is the dominant method used to model multiple-objective decisions within decision analysis. The goal of hierarchical decomposition is to represent each possible decision outcome as a vector of values, each element of which is associated with a single objective. In some work on multiple-objective decision making, these vectors are simply assumed as a starting point, and then subjected to a value function or utility function analysis.[34] But, since my goal in this chapter is to get some traction on the specification problem for multiple-objective decisions, it is necessary that I discuss how we might arrive at such a vectorial representation of consequences.[35]

[33]Raiffa 1969, 122, refers it as a hierarchical method; Fischer 1973, 22 refers to it as a decompositional method. I think the combined term is apt, for reasons that will become apparent. The method is discussed in detail in Raiffa 1969, and Keeney and Raiffa 1976, but because what follows is an informal and condensed presentation of ideas distributed throughout these the whole of these works, it is not generally possible for me to give page-specific citations of this material. Keeney 2006, 171, presents a concise overview of the method.

[34]See: Winterfeldt and Fischer 1973a, 51, or Tzeng and Huang 2011.

[35]For the formally disinclined reader: In single-objective decisions, consequences are typically represented with a single, *scalar*, real number values. The real numbers come with a ready-to-hand weak-order (see my earlier note on weak orders), the relation of less-than-or-equal-to, which is perfect for comparison of value levels. A *vector*, on the other hand, is a tuple of values, and there is no single obvious metric over these tuples that stands in for the less-than-or-equal relation over the real numbers. Consider a toy example. Suppose we have a room full of one hundred individuals. If we want to sort them individually by height from shortest to tallest, it is clear how the task should be completed. Each student can be represented by a real-value scalar—her height in inches—and we can then order these scalars with the less-than-or-equal to relation. But suppose instead, that the students are arranged in groups of ten, and the

1.5.1 Identifying the decision

The first step in the hierarchical decomposition method is, not unexpectedly, to identify the decision being made. The essence of the hierarchical decomposition method is to work downward from this overall characterization of the decision—often somewhat misleadingly referred to as a "goal" in the decision analysis literature—to lower and lower level objectives, until quantifiable structure emerges. Thus, the overall characterization of the decision need not be especially concrete, structured, or quantifiable. These features will emerge as the analysis continues. However, the overall characterization of the decision is meant to provide some guidance in how to carry out the rest of the process. Thus, it should be specific enough that we can ask practically oriented questions about how to resolve the decision. Thus, decision analysts urge that we avoid characterizations like, "deciding on what's best," or "deciding what to do," and instead characterize the particular decision that the DM faces more precisely, like "deciding which bet to place on the next round of craps."

Throughout, I'll illustrate the hierarchical dependence method by reference to two example cases. In the first case, let us suppose that Chandeep's old jalopy has finally died, and she is considering buying one of several distinct new cars that fall within her price range. In the second case, let us suppose that Dayo is a recent professional school graduate who is considering accepting one of several initial job offers that have come from his professional school's hiring fair. With respect to these cases, we'll assume that these obvious characterizations are also correct characterizations of the decisions that

task is to sort the groups by height from shortest to tallest. There is no longer any single, obvious way to sort the groups. We could sort them by the sum of their heights, or the average of their heights (since the groups are all the same size these methods will be identical), or by tallest member, or by shortest member, or by modal height, or by variance in height, and so on. How we elect to sort the groups will depend on our purposes in sorting them. If we're picking a basketball team, we might want the group with the tallest individual. But if we're picking a crew team, we might want the group with the lowest variation in height and an average height above a certain threshold. Thus, as we transition from single-objective decisions, in which outcomes can be represented with scalar values, to multiple-objective decisions which require vectorial representation, the task of comparing consequences becomes dramatically more complex.

Chandeep and Dayo face.

Keeney 1992 and J. S. Hammond, Keeney, and Raiffa 1999 suggest it is a mistake to take for granted the obvious characterization of any decision, and stress that we should not trivially identify the decision before a DM as a choice between the apparent options. Reflection on which decision the DM actually faces sometimes reveals that instead of an apparently rigid decision with obvious options, she faces an altogether different, plastic decision for which she can generate a considerably wider variety of options. For example, the obvious characterization of Chandeep's decision is as a choice between new cars. But some reflection on her situation may reveal that, now that she has retired her old jalopy, what she really needs isn't a new car, but simply a way to get to and from work and occasional recreational destinations. So, perhaps she could consider used cars in addition to new cars, or she could consider using public transportation and ride sharing services, or simply bicycling, rather than simply defaulting to buying a new car. And we've suggested that Dayo's overall goal is to select the right job offer. But some reflection on his situation may reveal that he should consider applying to additional jobs, or trying to renegotiate some of his additional offers, or applying for additional schooling, or taking some time off.

As we discuss these cases further, it will be obvious how characterizing the decision differently can lead hierarchical decomposition method to produce entirely different results. For simplicity of presentation, we'll continue to characterize Chandeep and Dayo's decisions in the obvious way, as choices between new cars and job offers, respectively.

1.5.2 Identifying the objectives

Once an overall characterization of the decision is identified, we proceed by identifying the objectives the DM has for her decision. Recall that we're thinking of an objective as anything the DM understands to be at stake in, or hopes to achieve, through her deci-

sion. Keeney 1992, suggests we think of identifying the objectives as making explicit "the values that are of concern in a given decision situation (55)." As noted in §2.1 above, objectives may involve the sort of things we traditionally think of as philosophically interesting values (moral, aesthetic, epistemic), but they may also be more mundane and practically directed. Recall also that objectives have an orientation. We can distinguish positively oriented objectives, according to which the DM aims to maximize some feature of the consequences of her actions, from negatively oriented objectives, according to which the DM aims to minimize some feature of the consequences of her actions.

Sometimes objectives will themselves be fairly high-level—that is, they will involve fairly high-level features of the outcomes. Then the process becomes iterative. Holding a higher level objective fixed, we can identify further sub-objectives achievement of which constitutes achievement of the higher-level objective.

From this, an *objectives hierarchy* emerges. Sometimes, especially when hierarchical decomposition is applied to large scale institutional decisions, it may seem that we can delve almost arbitrarily deeper and deeper into the objectives hierarchy, without any obvious end. This problem of where to draw the line when constructing an objectives hierarchy parallels the regress problem mentioned above. I discuss this and related concerns further in §5.4, after I introduce the notion of attributes.

In the meantime, let's return to the example cases just introduced. To identify Chandeep's objectives in realizing her decision between cars, we need to identify what she sees as at stake in her decision, and which features of her options are relevant for realizing her goal. We can elicit objectives in any of the many ways we would ordinarily try to identify what someone cares about. One way of generating objectives is to consider means for realizing the overall goal; another is to consider an imagined ideal option, and identify its goal-relevant features; still a third is to consider directly how to describe the relevant

features of the consequences of the decision.[36] Suppose, for example, we ask Chandeep to describe her ideal car, and she suggests it is affordable, environmentally friendly, safe, and comfortable. Then we can identify four distinct objectives in Chandeep's decision: to minimize cost of ownership, to maximize fuel-efficiency, to maximize vehicle safety, and to maximize comfort. It is obvious that these objectives can conflict. The cheapest car simply won't be the safest car; safety features cost money. And the most comfortable car is unlikely to be the most fuel efficient, since vehicle size tends to make cars more comfortable and less fuel efficient.

It may seem something of a stretch to move from Chandeep's description of her ideal car as affordable to an objective to minimize cost of ownership. But it is a safe characterization, because if affordability is a goal, then whenever other things are equal—say, Chandeep is considering the same make and model vehicle at two distinct dealerships—obvious dominance principles come into play, and it's clearly the case that Chandeep should choose the car at the lower price. In general, this is why we understand positively oriented objectives as objectives to maximize some feature, and negatively oriented objectives as objectives to minimize some feature of the consequences. Its also likely that in a real-world car-buying decision, more objectives would come into play. For example, it's likely that she would care about the color of the car, the quality of the stereo, the presence of absence of certain features, and so on. But in the interest of keeping the presentation here tractable, let's assume that Chandeep's objectives are limited to the preceding list.

And let's suppose that Dayo's objectives are to maximize his compensation, minimize his commute time, and maximize his prestige in his field. Again, it is likely that more objectives would come into play in a realistic decision between job offers, but this is a difference in complexity, not a difference in kind. So, again, to keep things tractable, we'll limit Dayo's list of objectives.

[36] Keeney 1992, 56–65 discusses the process of eliciting a list of objectives, introducing a variety of practically applicable techniques.

1.5.3 Identifying attributes

Corresponding to each lowest-level objective we then identify one or more *attributes*. We can think of an objectives hierarchy as bottoming out in the attributes. These are the ground level features of consequences that will be formally represented in the decision problem. Choice of attributes will be highly sensitive to the values of the individual DM subject to the analysis.

Good attributes will be both *comprehensive* and *measurable* features of decision outcomes. An attribute is comprehensive if, by knowing the level of the attribute in a particular decision context, the DM can gauge the extent to which the associated objective is achieved. An attribute is measurable if it is reasonable to obtain a probability distribution over the possible levels of the attribute for each of the DM's available actions, and we can assess the DM's preferences for levels of that attribute, holding all other attributes fixed. In decisions under certainty, of course, the task of obtaining a probability distribution over levels of the attribute for each action reduces to simply determining that attribute's level for the certain consequences of each action. Once the set of attributes is fixed, we can represent each possible decision outcome as a vector of those attribute values.

Keeney and Raiffa 1976 and Keeney 1992 suggest that we can distinguish three kinds of attributes: *natural* attributes, *constructed* attributes, and *proxy* attributes.[37]

Natural attributes are salient features of outcomes that are obviously apt representations of the extent to which the associated objective is achieved. Natural attributes are likely to be familiar and to have a common interpretation to any DM faced with a decision involving the associated objective. For example, if an individual has the objective to maximize her returns in a choice between a variety of investment packages, then net returns on investment in dollars is a natural attribute for this objective. Or if an individual has

[37] See, Keeney and Raiffa 1976, 33–35,55–61, and Keeney 1992, 100–112, for discussion of the various types of attributers

the objective to minimize her travel time in in her choice between air travel arrangements, then destination-to-destination time in minutes is a natural attribute for this objective. Obviously, naturalness of attributes comes in degrees, and value judgments emerge even in the choice of natural attributes. For example, if maximizing storage space is an objective in a DM's choice of homes, a natural attribute is area of storage space in square feet, but the choice of that attribute involves the value judgment that every square foot of storage space should be treated equally.

For some objectives, it is difficult to find natural attributes. This is most obvious with respect to the sort of objectives we think of as qualitative or subjective in nature, including objectives like maximizing visual appeal in a branding decision, or minimizing patient discomfort in a medical decision. In these cases, we can sometimes generate *constructed* attributes. One common way to generate a constructed attribute is through a procedure we might call *scoring out*, in which the DM directly evaluates each of the options, and assigns to each a subjectively generated numerical score on the basis of her qualitative experience of the option. There are, of course, more complicated ways of generating constructed attributes, and the task of generating a constructed attribute may itself reduce to a small multi-criteria problem. It is crucial to note that the DM need not care directly about the level of the constructed attribute. If, for example, a DM is choosing between childcare providers, and maximizing felt rapport with the provider is one of her objectives, she might score this out to generate a constructed attribute. It goes without saying, in such a case, that it is not the score, but the rapport itself that she cares about. The score is merely used as a way to incorporate her qualitative evaluations into the model.

In other cases, we can sometimes identify a *proxy* attribute. A *proxy* attribute is one that indicates the extent to which the associated objective is achieved, but in some sense, does not measure this achievement directly. Typically the level of a proxy attribute will

bear a means-to-end relationship with the associated objective. Suppose a DM has the objective to minimize degradation of an old painting in her choice of display options. Degradation itself may be challenging or impossible to measure without thereby damaging the painting, and a constructed attribute may be unreliable or hard to generate. In this, case, we could use a proxy attribute like UV light exposure. Note that exposure to UV light is itself a means to degrade the painting, and thus minimizing UV light exposure is a means to minimizing degradation. So a measurement of UV light exposure, while not a direct measurement of the extent to which the painting will degrade, is indicative of that extent.

In some cases, more than one attribute may be required to aptly represent the extent to which the associated objective is achieved. Here, though, we'll assume that we can identify a single comprehensive attribute for each objective. We can make this assumption without loss of generality. Suppose we are dealing with an objective the achievement of which is most obviously measured by a tuple of more than one attribute. Then the problem of distilling from these a single, overall attribute we can use to track the achievement of the objective is just a multiple-objective decision problem, writ small. Thus the techniques discussed below for distilling a value function from a tuple of attributes can be applied to construct a single, composite attribute from some tuple of sub-attributes. We can then treat that value function evaluated at each of the options as a single attribute for the associated objective.

Like objectives themselves, attributes can also have an orientation. For positively oriented attributes a higher attribute level indicates a higher level of achievement with respect to the associated objective. For negatively oriented attributes, a lower attribute level indicates a higher level of achievement with respect to the associated objective. In general, it simplifies an analysis and the resulting model if all attributes are kept positively oriented. One way to do that is to press for only positively oriented objectives,

which may be easier to associate with positively oriented attributes. (This is not always the case: consider the prestige ranking in Dayo's case below.) But another way is to begin with an obvious negatively oriented attribute, and transform it into a positively oriented attribute. Often, the simplest way to do this is to set a upper threshold value, and track the difference between this threshold value and the actual value of the natural attribute for each option.

Let's return to our example cases. Recall that Chandeep had objectives to minimize cost of ownership, minimize environmental impact, maximize vehicle safety, and maximize comfort. Let's suppose, to keep things simple, that she plans to pay cash for the car, without negotiating. Then list price of the vehicle is a natural, comprehensive, and measureable attribute to associate with her objective to minimize cost of ownership. Unfortunately, it is a negatively oriented attribute: the higher the list price, the worse she is doing with respect to the associated objective. But, as just noted, we can easily generate a natural, positively oriented attribute. We need only identify the maximum list price of options under consideration, and then track the attribute of savings in dollars beneath this maximum price.

Minimizing environmental impact, on the other hand, is not obviously associated with any natural attributes. And, unless Chandeep is herself an environmental scientist with expertise in automobiles, it's unlikely that she will be positioned to generate a constructed attribute. So we should consider proxy attributes. Independently estimated estimated highway miles-per-gallon is a strong candidate for a proxy attribute here. Since data about the environmental impact of manufacturing processes of automobiles are not widely available, it's unlikely Chandeep can learn much more about the environmental impact of the vehicles she is considering. So estimated highway miles-per-gallon stands out as both a measurable and comprehensive proxy attribute to associate with this objective.

Maximizing vehicle safety is also not associated with any natural attributes. And again, if Chandeep is not an expert highly proficient in the subject, it's unlikely that she herself will be able to generate a constructed attribute. So again, we should consider proxy attributes. In this case, the most plausible proxy attributes are constructed attributes generated by experts in the field, like independently assessed vehicle safety ratings. Let's suppose that Chandeep has special confidence in the safety ratings of a particular agency; then these ratings serve as a measurable and comprehensive constructed proxy attribute to associate with this objective.

That leaves maximizing vehicle comfort. This, like many realistic decision objectives, is a qualitative, and highly subjective objective. It seems we have two obvious choices here. If Chandeep can actually test-drive all of the vehicles she's considering—it's not a stretch to make this assumption—then she can score out the vehicles with respect to comfort on some arbitrary scale to generate a constructed attribute. If she is unable to test-drive the cars, perhaps because she is buying remotely, then she may have to resort to a constructed proxy attribute as above, like the comfort score assigned to each car by a trusted independent review of automobiles. Let's assume that she is able to test-drive the vehicles, and generates a comfort scale as a constructed attribute.

We now have four attributes associated with her four objectives: savings in dollars beneath the maximum price, independently estimated highway miles-per-gallon, independently assessed safety score, and subjectively scored comfort score. Then we can represent each possible decision outcome—that is, each car she is choosing between—as a vector of values of these four attribute values.

Recall that Dayo had objectives to maximize his compensation, minimize his commute time, and maximize his prestige in his field. For simplicity, let's assume that all of Dayo's job offers feature all-salary compensation packages. Annual salary in dollars is an obvious natural, measurable, and comprehensive attribute for this objective. There are no natural

and comprehensive attributes for minimizing commute, but good constructed attributes are fairly obvious. While it's unlikely Dayo could simply score out his various options, he could, for example, research the estimated peak-traffic travel time in minutes for each possible route to each of his job options, and average these for each option, and then consider the difference between these values and the maximum estimated time in traffic. This seems like an ideal constructed attribute that is both measurable and comprehensive. With respect to prestige, we might consider this a purely qualitative objective, for which Dayo has to generate a constructed attribute. But there might be an independently established prestige rating of firms which he could use as a ready-made constructed attribute. (Such ratings actually exist for law firms, and investment banks.) For Dayo we have three attributes associated with his three objectives, and we can represent each possible decision outcome as a vector of values of these three attributes.

1.5.4 Properties of good objectives and attributes

Eliciting an objectives and attribute hierarchy from a DM is a sensitive and tricky process, and for many decisions there is no single, obviously correct objectives hierarchy. Keeney and Raiffa 1976 put the point thus:

> The objectives hierarchy for a particular problem is not unique. It can be varied simply by changing the degree to which the hierarchy is formalized. However, even if the degree of formalization remains unchanged (in the sense that the number of lowest-level objectives is the same), the objectives hierarchy can be significantly varied. Whether one arrangement is better than another is mainly a matter of the particular points the decision maker wish to make. . . . With different hierarchies, different tradeoffs facing the decision maker can be more easily identified and illustrated. (47)

Nonetheless, we are not adrift without direction here. Hierarchical decomposition is ultimately a method deployed to a clear practical end: to structure a formal representation of a real-world decision that simultaneously aptly represents the DM's attitudes, and involves the appropriate formal structures to be resolved using ADT. Thus, Raiffa 1969,

125, suggests that since the elaboration of an objectives hierarchy is not unique, we would be well advised to choose an objectives hierarchy that enables further analysis. With that in mind, we can identify some desirable properties of the attributes corresponding to the lowest-level objectives, and can reasonably stop delving deeper once we've reached a set of attributes with these properties.[38]

A good set of attributes will be *complete*, *operational*, *decomposable*, *non-redundant*, and *minimal*.

A set of attributes is *complete* if it covers all important aspects of the outcomes. In other words, given a complete set of attributes, knowledge of all the attribute values for a given outcome provides the DM with a full description of those features of the outcome she views as relevant in her decision.

A set of attributes is *operational* if it can actually be meaningfully applied to the decision at hand. The attributes must be intelligible to the DM, and it must be possible to actually collect the necessary information about the options to assess either attribute levels, or probability distributions over attribute levels.

A set of attributes is *decomposable* if we can break it into parts to simplify aspects of the analysis. This is especially crucial when the set of attributes is of even moderately large size. Typically, independence conditions are leveraged to decompose the set of attributes. I defer discussion of these notions to §5.8 below, since they require the introduction of some further concepts.

A set of attributes is *non-redundant* when it avoids double-counting in the evaluation of outcomes. Double-counting occurs when the same feature of the outcomes are measured (perhaps in a different way) by two distinct attributes. Suppose, for example, that in a decision between investment packages, expected return from securities and expected return from stocks are both used as attributes. Since stocks are themselves a kind of se-

[38] For discussion of desirable properties of lowest level attributes, see: Raiffa 1969, 122–124 Keeney and Raiffa 1976, 50–53; and Keeney 1992, 82–86.

curity, their impact on the consequences will be problematically counted twice. This will falsely exaggerate the difference between an investment package which contains some stocks and one which contains none. Most importantly, redundancies can compromise the independence properties that can be leveraged to decompose the attribute set.

Provided we have identified a set of attributes with the above properties, it is desirable to keep it to a *minimum* size.

Let's assume that the sets of attributes described for Chandeep's and Dayo's cases are complete. They are also clearly operational, and non-redundant, and seemingly minimal, since there is no way we could prune them down. I'll revisit whether they are decomposable below in §5.8.

Above I noted that weak preferences are typically assumed to be a necessary condition for characterizing rational action. In considering how to model multiple-objective decisions we have backed away from this assumption considerably. It would, of course, be a mistake to say that representability by a set of attributes that is complete, operational, decomposable and non-redundant is a necessary condition for characterizing good decision making in a multiple-objective case. Suppose we have a multiple-objective decision for which no set of attributes satisfying these desiderata can be described. There may still be clear ways to identify good choices. One option may simply dominate all others. Or we might be able to identify a lexicographic order over the attributes according to which a best indifference class could be selected. Or we might be able to leverage fragmentary holistic preferences—more on this below—to rough-in a value function, and then refine this approach with feedback from the DM. Nonetheless, when we cannot identify a set of attributes with these desiderata, we must concede that we are pushed to the limits of our theoretical capacity to model decisions using decision analytic techniques.

1.5.5 The consequence space, outcomes, states, and actions

Once we have determined an adequate set of attributes, we have thereby described what we can call the *consequence space* for the decision. Roughly, the consequence space is the set of mathematically possible combinations of attribute values.

More precisely, since each attribute is a real-valued measure over some feature of the decision outcomes, we can represent these attributes in our model as bounded intervals of real numbers. Then the consequence space is the set of vectors the first element of which takes a possible value for the first attribute, the second element of which takes a possible value for the second attribute, and so on. We can then identify any possible decision outcome as a particular vector of attribute values, or *point* within this consequence space. It is important to stress that the consequence space is the set of *mathematically* possible combinations of attribute values, not the set of *practically* or *realistically* possible combinations of attribute values. In multiple-objective decisions, just as in single-objective decisions, it is sometimes important to be able to consider imaginary outcomes.

We opt for intervals of real numbers to facilitate modeling of continuous trade-offs of arbitrary levels of the attribute in question. And we opt for bounded intervals in order to keep the consequence space both manageable, and more importantly, intelligible to the DM. For some attributes the maximum and minimum values will be obvious or natural. In other cases, there may be no natural maximum or minimum value. In these cases we can bound the intervals at values beyond which levels of that attribute will no longer be intelligible to the DM. If we've adopted dollar-value attributes for a small-scale personal financial decision faced by a middle-income DM, then we can reasonably ignore values of these attributes that are a factor of one hundred greater than her annual income. Values that large will simply be practically unintelligible to her.

Since the focus here is on the preference analysis required to represent decisions with multiple objectives, I'll assume that we already have an adequate representation of the

uncertainties that impinge on the decision. So I'll assume that we have aptly represented the DM's relevant probabilistic beliefs with a probability function[39] over a set of mutually exclusive states of the world. Here, we should think of a state of the world as some way the world might turn out to be that can relevantly affect the consequences of the DM's decision. That is, we can ignore differences in how things could turn out that make no distinctions between any of her available actions.

Informally, we can think of *actions* as ways for the decision maker to get to outcomes. More precisely, we can think of actions as functions from states to outcomes.[40] Using the probability function over states, we can determine a probability distribution over outcomes associated with each action. *Certain* actions always lead to the same outcome, no matter the state of the world; if actions are ways to get to outcomes, a certain action is a sure way to get to a particular outcome. When we're dealing with a decision under certainty, we can simply identify each action with its associated outcome. And we will have identified good decision making behavior when we have developed the model enough to characterize a best outcome according to the model. When dealing with decisions under uncertainty, we have to go a step further, and characterize a best action according to the model.

Let's return again to our example cases: Let's call her attributes **dollars saved**, or D for short, **miles-per-gallon** or M, **safety rating** or S, and **comfort score** or C, for ease of reference. We can represent each of these attributes as a range of possible values. We'll set these ranges arbitrarily, as nothing hangs on the particular values the attributes can take. In a realistic analysis, these ranges will obviously be determined, or at least loosely constrained by the results of the analysis. Let's assume values of D range from 0

[39] For the formally disinclined reader: A probability function is a mapping over a set of inputs that pairs each input with a value from zero to one, inclusive, such that the sum of values for all inputs is one, and the values are assigned in a manner that corresponds to some axiomatization or othe of the laws of probability.

[40] For the formally disinclined reader, each action can be thought of as a rule which associates exactly one particular outcome to each state.

to 25,000, values of M range from 15 to 50, values of S range from 0 to 100, and values of C range from 0 to 10. Then any tuple of the form $\langle d, m, s, c\rangle$ where d, m, s, and c fall within the ranges of D, M, S, and C respectively is an outcome in the consequence space of her decision.

And let's call Dayo's attributes **wages in dollars** or W, **time not in traffic** or T, and **prestige rating** or P. Let's suppose that values of W range from 50,000 to 250,000, values of T range from 0 to 100, and values of P range from 0 to 25. Then any tuple of the form $\langle w, t, p\rangle$ where w, t, and p fall within the ranges of W, T, and P respectively is an outcome in the consequence space of her decision.

Let's assume for now that Chandeep's and Dayo's decisions are made under certainty. Then we can identify the actions available to each of them with points in their respective consequence spaces. Suppose one of the actions available to Chandeep is to buy a Volvo station wagon, which will result in a savings of \$5,000, gets an estimated 25.2 highway miles-per-gallon, has a safety rating of 98, and to which she assigned a comfort score of 8.5. The we can identify the action of **buying the Volvo** with tuple $\langle 5{,}000, 25.2, 98, 8.5\rangle$. Suppose one of the actions available to Dayo is to accept an offer from Moneymaker and Rich Partners, a firm that has offered him a salary of 125,000 dollars, the commute to which is 25 minutes less than the maximum commute, and which has a prestige rating of 18. Then we can identify the action of **working for M & R** with the tuple $\langle 125{,}000, 25, 18\rangle$.

1.5.6 Preference structures, value functions, holistic preferences, and partial preference structures

Ultimately, the goal of the analysis is to identify a *value function* over the consequence space, which will determine a *preference structure* over the same space sufficient to resolve decisions under certainty, and with which we can elicit a *utility function* over the same

space sufficient to resolve decisions under uncertainty. I elaborate on the notions of value functions and preference structures here, and defer further discussion of utility functions to the appendix.

In keeping with the standard use of the term, let's say that two outcomes are *comparable* in the eyes of a DM, if she can say of them whether one is preferable to the other, or she's indifferent between them.[41]

We'll say a DM has a *preference structure* over the consequence space of her decision, if and only if, by her lights, any two outcomes are comparable without intransitivities. In other words, she has a preference structure whenever she has weak preferences over outcomes. A *value function* is a function from outcomes to real numbers associated with a preference structure in the following ways: Preferable outcomes are assigned higher values than less preferable outcomes, and outcomes between which the DM is indifferent are assigned equal values. It is crucial to note that a value function is not necessarily a utility function, in the sense of von Neumann and Morgenstern 1944 or any other standard axiomatization of ADT, though every utility function is a by definition a value function which conforms to additional axioms.

Whereas a preference structure determines a family of value functions unique only up to positive linear transformation, a value function uniquely determines a preference structure.[42] Thus, if we can leverage the DM's available preference information to determine an acceptable value function, we have thereby determined a preference structure

[41] See the essays of Chang 1997a, in particular Chang 1997b, for a good overview of comparability and incomparability.

[42] For the formally disinclined reader: Suppose we have just three items, a, b, and c, and a is strictly preferred to b which is indifferent to c. Then the function that maps b and c to 0, and a to 1 is a value function over those options that conforms to the preferences. But so is the function that maps b and c to $1,000$ and a to $1,000,000$. When we say that value functions are unique only up to positive linear transformation, we mean that if we take the outputs of some value function, and multiply them all uniformly by the same positive real number, and add the same real number to them all uniformly, the result is another acceptable value function.

On the other hand, if we start with a value function that maps a to 3, b to 5, and c to 7, this uniquely determines a preference structure over a, b, and c, according to which c is strictly preferred to b which is strictly preferred to a.

over the consequence space.

Multiple-objective decisions would be no different from single-objective decisions if we could simply assume that the DM had a preference structure over the consequence space of her decision. However, the DM may have preferences over small bundles of outcomes, which may overlap, but do not cover the whole of the consequence space. That is, there may be subsets of the consequence space (sets of outcomes) over which the DM has *holistic preferences*. I refer to these as holistic preferences because they are determined by a direct assessment of the outcome as a whole, not in terms of its representation as a vector of attributes. These should be understood as primitive features of the DM's attitudes, to be represented in our formal model, and leveraged in our efforts to assess a value function.

If we are to have any hope of generating an overall preference structure that aptly represents the DM's attitudes, then we'll need to assume that these holistic preferences are consistent and free from transitivities such that they can be represented as a partial order over the consequence space.[43] We'll call the partial order determined by the DM's

[43]For the formally disinclined reader: Recall form the earlier note that a weak order is a binary relation that is reflexive, transitive, and total. A partial order is a binary relation that is reflexive and transitive, but is not necessarily total. Every weak order is, ipso facto, a partial order. Every partial order can be extended to a weak order in a multitude of ways.

Here's a toy example to illustrate: Suppose 52 people are dealt cards face-down from a standard deck of playing cards, to be sorted by into groups by the binary relation of winning-or-tying the others at a game of High Card, which is won by the card of higher rank, regardless of suit. Suppose that individuals turn over their cards one by one. When the first individual turns over his card, a partial order is determined over the set of people: she is related to herself, and no one else is related to anyone. At this juncture, there is an incomprehensibly enormous, but finite number ($51 \times 50 \times 49 \times \cdots \times 1$) of ways the partial order can be extended to a weak order over the individuals that are consistent with the possible values the cards can take. As each additional individual turns over her card, the partial order becomes more complete, and the number of distinct ways it can be extended to a weak order is considerably reduced. Suppose only Ashima and Badri have yet to turn over their cards, and the two remaining cards are an ace and a deuce. Then either Ashima will be ranked higher than everyone but the other aces, and Badri will be ranked lower than everyone but the other aces, or vice versa. There are no other ways to extend the partial order to a weak order.

Now imagine that one of the 52 standard cards is removed, and replaced with a joker, which beats a king, but loses to a deuce, and ties anything else. As soon as the joker, a king, and a deuce have been turned over, we no longer have a partial order, and this can no longer be extended to a weak order over the individuals, since these individuals are intransitively related.

holistic preferences a partial preference structure. On a first pass at analysis, we'll restrict our consideration to value functions that respect this partial preference structure (in the sense that they determine preference structures which are extensions of the partial preference structure). No doubt eyebrows have been raised at the qualification in the preceding sentence. I'll offer further explanation in §5.11 below.

1.5.7 Dominance and lexicographic orders

In some cases, once a consequence space has been described, and the available actions have been represented as functions from states to outcomes, the decision can be settled by recourse to a dominance principle, or a lexicographic order over the attributes.

We can generalize the familiar idea of *dominance* to multiple-objective outcomes and actions in the obvious way. We'll say one outcome *dominates* another if and only if it is strictly better with respect to every attribute. And we'll say that one action *dominates* another if and only if the former action's outcomes dominate the latter actions outcomes in every state of the world.

In rare cases, one available action will dominate all others, and will clearly stand out as the best choice according to the model. When this is the case, if our only ambition is to identify good decision making behavior, we have carried the analysis as far as it needs to go. On the other hand, if we hope to describe a complete preference structure over the consequence space, or to assess a utility function, we have to press on.

Even where there is no single dominant action, dominance can be used to extend any partial preference structure that results from the DM's holistic preferences.

Equally rarely, the DM's attitudes may reflect a lexicographic ordering over attributes. This arises when the DM can sort the attributes strictly by priority. Then we can quite easily describe a complete preference structure over the consequence space. We begin by assessing outcomes according to the highest-priority attribute; outcomes with higher

values of this outcome will be strictly preferred to outcomes with lower values. Then, among outcomes that are equivalent with respect to the highest-priority outcome, we'll consult the next highest attribute, and so on, until a complete preference structure has been described.

I reiterate, realistic cases are rarely settled by dominance, and decisions for which lexicographic emerge are few and far between. However, dominance relations help to extend any partial preference structure determined by the agents holistic preferences, and can thus be used to help in the assessment of a value function.

1.5.8 Conditional preferences, preferential independence, and additive value functions

Among the desiderata for a set of attributes listed is that the set be decomposable. Sets of attributes can be decomposed when we can establish that they satisfy certain independence properties. These properties are challenging to discuss in an informal way; I direct the reader to the appendix for clarification.

To characterize these properties, we first need the notion of *conditional preference*. Consider an arbitrary set of attributes, and select one attribute from it. If we hold the values of all other attributes fixed, we can then elicit from the DM preferences for values of the selected attribute conditional on the fixed values of the non-selected attributes. We could instead select two attributes, and hold the values of all other attributes fixed, and then elicit from the DM preferences for tuples of values of the two selected attributes, conditional on the fixed values of the non-selected attributes. And so on. Call these preferences for tuples of values of some set of selected attributes determined while values of the set of non-selected attributes are held fixed, *conditional preferences*.

From this we can define a notion of *preferential independence*. A selected subset of attributes is *preferentially independent* of the complementary subset of attributes, if the

conditional preferences for tuples of values of the selected attributes don't change, no matter which values we choose for the remaining attributes.

Whenever a subset of attributes and its complement are preferentially independent of one another, we can decompose the original set of attributes into these two pieces, and continue our analysis on each separately. We can think of this as breaking the consequence space into two simpler subspaces. The crucial fact here—entailed by well-established results—is that we can independently assess value functions over each of these subspaces, from which we can construct an overall value function over the original consequence space.

When every subset of attributes we can select is preferentially independent of its complementary subset of attributes, then the set of attributes is *mutually preferentially independent*. The original set of attributes can then be decomposed down to its atoms. Or, if you like, the consequence space can be split into individual attributes. Then we can independently assess a value function for each attribute. Further, if we have three or more attributes, it is a well-established result that the overall value function for the whole consequence space will be an *additive* function which simply sums the (scaled) values of each attribute-specific value function. Above, I suggested that a higher number of objectives can sometimes make a decision easier to resolve. We cannot leverage pairwise preferential independence to establish the existence of an additive value function in a case where we have only two attributes, instead stricter conditions apply. Thus, the addition of a third objective, and thereby a third attribute, can sometimes considerably simplify the analysis of the decision. However, for even a small set of attributes, directly establishing pairwise preferential independence requires establishing mutual preferential independence a disconcerting number of times. Fortunately, there are additional widely-established results that dramatically reduce the number of comparisons required.

Obviously, use of these independence results to decompose a set of attributes into

smaller and smaller subsets simplifies the ensuing analysis enormously. Decisions involving unmanageably large numbers of attributes can be made tractable by carefully identifying a decomposable set of attributes.

Let's return to our example cases. Recall that Chandeep's attributes were **dollars saved** (D), **miles-per-gallon** (M), **safety rating** (S), and **comfort score** (C). Consider **dollars saved** to one side, and the remaining attributes to the other. It is intuitively obvious that **dollars saved** is preferentially independent of the remaining attributes. No matter where we hold the other attribute fixed, Chandeep will prefer to save more, conditional on those fixed values. Indeed, each individual attribute is clearly preferentially independent of the three remaining attributes. However, it is not obvious that each triple of attributes is preferentially independent of the remaining single attribute. Again, consider **dollars saved** and the remaining triple of attributes. Suppose we hold D fixed at a low value, so that Chandeep is considering an expensive car. She may have particular expectations for an expensive car, say that it be especially safe and comfortable. Indeed, this seems likely. So, with D held fixed at that value, she may weight values of S and C more heavily in her preferences than she does values of M. If she does, we will not be able to identify an additive value function in Chandeep's case.

Still, it may be possible to decompose her set of attributes to simplify the problem. For example, the pairs of attributes **dollars saved** and **miles-per-gallon**, and **safety rating** and **comfort score**, it seems likely these pairs will be preferentially independent of one another. Let's suppose that they are. Then we can decompose Chandeep's attribute set into those two pairs of attributes, and independently assess value functions over each pair. Then, we are effectively left with a two-value characterization of outcomes, from which we can assess an overall value function.

Recall that Dayo's attributes were **wages in dollars** or (W), **time not in traffic** (T), and **prestige rating** or (P). Again, each individual attribute seems preferentially independent

of the remaining attributes. For example, if we hold levels of W and T fixed, Dayo will obviously prefer higher levels of P, and this will be the case no matter where we hold W and T. Here, though, each pair may be preferentially independent of the remaining individual attribute. For example, it seems unlikely that Dayo will weight values of W and P differently at different values of T. Suppose, that the pair W and P is mutually independent of T. Then at worst, we can decompose Dayo's attribute set into those two parts, and simplify the resulting analysis to assessing a value function over T, and a value function over W and P, and then constructing an overall value function from these. It is crucial to note that assessing the value function over T is not necessarily trivial; the value function may not simply map values of T directly to themselves. For example, Dayo may more strongly prefer going from saving 60 minutes of his commute to saving 40 minutes of his commute, than he prefers saving 80 minutes off his commute to saving 60 minutes.

In Dayo's case, let's suppose further that each pair of attributes is independent of the remaining individual attributes. Then Dayo's attributes are mutually preferentially independent, and we can decompose his consequence space down to individual attributes, assess a value function over each attribute, and his overall value function over the consequence space will simply be the (scaled) sum of each attribute-specific value function evaluated for each individual element of an outcome.

1.5.9 Assessing value functions

Thus far, we have explored how to use the hierarchical decomposition method to construct a set of attributes which represents outcomes of a multiple-objective decision as points in a consequence space. And we've discussed how a value function over that consequence space can be used to determine a preference structure over that consequence space, and suggested that it can also be used to determine a utility function over the consequence space. We've also seen how a set of attributes can be decomposed to simplify

the ensuing analysis of the value function. To fit everything together, we need to see how to elicit value functions.

Methods for eliciting single-attribute value functions are familiar within the philosophical literature on ADT. Savage 1954 and Resnik 1987 both discuss such techniques. Consequently, my presentation here is somewhat superficial. The standard procedure is roughly as follows: We begin by considering the upper and lower bound for levels of the attribute. Call these u and l respectively. We map these attribute levels to arbitrary values, say 100 and 0, respectively. Then we identify the subjective midvalue point between u and l. Call this m. The midvalue point is the level of the attribute such that the change in value from l to m is identical to the change in value from m to u. In other words, the DM considers m just as much better than l as u is than m. Then it follows that m should have a value exactly in between the value of u and the value of l, so m should be mapped to the value 50. Then we consider the subjective midvalue points between l and m, and m and u. These should be mapped to values exactly in between the values of l and m, and the values of m and u, so to 25 and 75 respectively. We continue this procedure until we have enough data to determine a precise mathematical function, using standard techniques for fitting curves to data. Then we check this function against the DM's subjective evaluations of attribute levels, to verify consistency. And thus we get a single-attribute value function.

In some cases, it is actually easier and more intelligible to assess a single-attribute value function against the backdrop of other attributes, than in isolation. Consider a two-attribute case, in which the attributes are, say, cost in dollars, and volume in cubic feet, and suppose that the additivity conditions have been met for these attributes, such that we can independently assess single-attribute value functions. Suppose we are trying to assess a value function for volume in cubic feet. Suppose the least volume is 50 cubic feet, and the greatest is 500. Then, to find the midvalue point we can formulate things

in terms of the other attribute, cost in dollars. We then elicit from the DM the level of volume between 50 and 500 such that she'd be willing to pay the same amount in dollars to increase the volume from 50 to that level, as she would to increase from that level to 500. Pricing out the attribute under analysis, in terms of levels of the other attribute(s) can give the DM some traction when making value comparisons.

Eliciting a value function over two attributes can be tricky. Where the additivity conditions are met, of course, we can follow the procedure just outlined. But where these conditions are not met, things become considerably more complicated. Typically, this process involves a protracted cycle of guesswork, consistency checks, refined guesses, and more consistency checks. The process can be made easier if the DM has some partial preference structure of holistic preferences over the consequence space determined by those two attributes. We can also exploit dominance relations to help narrow our focus. Further, the process can made a great deal more tractable if the pair of attributes under consideration is mutually preferentially independent of another set of attributes over which a value function has already been assessed. In its broad strokes, the procedure for assessing two-attribute value functions is similar to the one used for single-attribute value functions, just a great deal messier. I defer discussion thereof to the appendix.

When we are confronted with a set of three or more attributes that cannot be decomposed into smaller sets, over which a value function must be assessed directly, the process explodes in terms of complexity, and the task cannot be reasonably managed without computer assistance.

Then, even when an additive overall valuation function exists, we must confront the further task of assigning *scaling factors* to each lower-level valuation function over which it sums. Crucially, these must not be understood as weights—at least not in the sense generally understood in philosophy, according to which a heavier weight communicates

a greater degree of importance.[44] Keeney 2006, writes:

> . . . [S]caling factors are often misinterpreted as indicating the relative importance of the objectives. Scaling factors do not indicate the relative importance of the objectives, but rather they indicate the relative importance of changing the level of performance on the respective objectives from their worst to their best levels as specified for the decision under consideration. (172)

Again, we can leverage the DM's partial preference structure over outcomes to aid in the task. By considering groups of outcomes over which the DM is indifferent according to her partial preference structure, we can identify points of necessarily equal value. These indifferences will constrain the scaling factors; any set of scaling factors that generates a value function which is consistent with these indifferences will be consistent with the DM's partial preference structure.

Let's return to the example cases. Recall that Chandeep's attributes could be decomposed no further by way of preferential independence conditions than to the pair **dollars saved** and **miles per gallon**, and the mutually preferentially independent pair **safety rating** and **comfort score**. (It remains possible that these pairs of attributes may be further decomposed if they meet some additional conditions discussed in the appendix.) Assessing an overall value function in this case will likely be tricky. First, assessing the two-attribute value functions over each pair of attributes could get quite messy, and then we more or less have to repeat the process a third time to generate the overall valuation function. (This is not to say that an acceptable value function cannot be assessed; approaches to assessing value functions like this are discussed in the appendix.)

In Dayo's case, on the other hand, the prospects for assessing an overall value function are much better. Since each of his individual attributes is mutually preferentially independent of the complementary pair of attributes, an additive overall value function exists, and we can independently assess value functions over each attribute using the methods for assessing single-attribute value functions above. From there, we need

[44] See, for example, Dancy 2004, Ch. 1; Schroeder 2007, Ch. 7; and Horty 2012, Introduction.

only identify scaling factors, and we have an overall value function for Dayo. Recall that Dayo's attributes are **wages in dollars** (W), **time not in traffic** (T), and **prestige rating** (P). For simplicity of presentation, let's assume the value functions for each attribute are just the identity functions—that is, they map each attribute value to itself—and the scaling factors are $\frac{1}{2{,}500}$, 1, and 4, respectively. Then given any point in the consequence space of Dayo's decision of the form $\langle w, t, p \rangle$, the value function will map that point to the real value $\frac{w}{2{,}500} + t + 4p$. Above, we identified the action of **working for M & R** with the outcome $\langle 125{,}000, 25, 18 \rangle$. Suppose another available action for Dayo is **working for the Feds**, which is identified with the outcome $\langle 80{,}000, 60, 22 \rangle$. Then, using the value function, we can describe each of these actions with a single value. Call the real value to which an outcome is mapped by the value function its *evaluation*. The evaluation of **M & R** is 147, while the evaluation of **Feds** is 180. Because the evaluation of working for the Feds is higher than that of working for Moneymaker and Rich, that working for the Feds is a better choice for Dayo. In the next section, we will get a better understanding of why this outcome should be preferable to Dayo.

1.5.10 Enough for certainty; from certainty to uncertainty

Just as with single-objective decisions, it is clear that multiple-objective decisions vary considerably in difficulty when it comes to structuring a representative decision problem. When we can identify a complete, operational, non-redundant set of attributes that exhibits mutual preferential independence for a given decision, fully specifying an aptly representative decision problem can be quite easy. When we cannot identify such a set of attributes, and in particular when the set of attributes resists decomposition, the decision may slip through our formal net. In the range of cases in between, the gap between real-world multiple-objective decisions and representative decision problems varies considerably in width, and we may have to really turn the crank of our analytical machinery

to churn out a representative decision problem.

Still, suppose we have deployed the hierarchical decomposition method to elicit from the DM a complete set of attributes, and a scaled value function over the associated consequence space that is consistent with any partial preference structure we could determine from her holistic preferences. Hierarchical decomposition is a process of identifying the low-level features of outcomes that the DM cares about—either directly, or as proxies—and determining the structure of her preferences for tradeoffs between these attributes. The resultant value function implicitly encodes the DM's preferences for tradeoffs between these attributes, in a manner consistent with any unrevised preexisting partial preference structure we could elicit from the DM. Recall that this value function uniquely determines a complete preference structure over the consequence space. Then we can think of the value function as using the DM's preferences for tradeoffs between attributes to extend her preexisting partial preference structure into regions of the consequence space that were initially preferentially obscure. Crucially, the preferences between outcomes fixed by the value function are determined by *her* preferences for tradeoffs between features of outcomes. The value function does not impose a preference structure on her; rather it can be thought of either as revealing her heretofore obscure preferences, or as a way for her to construct them that is consistent with her preferences for tradeoffs between attributes. Thus the preferences between outcomes determined by the value function should either be thought of as *her* preferences, or preferences she could consistently adopt.

The preference structure determined by the value function is representable as a weak order over outcomes. Then we have arrived at a standard formal representation of a decision under certainty: we have a set of options, and an associated weak order over those options. According to ADT, this is sufficient to characterize good choices under certainty; any option weakly preferred to all others in the resulting preference structure

is a good choice.

So, let's take one last look at our example cases. We identified **working for M & R** with the outcome $\langle 125,000, 25, 18\rangle$ which had an evaluation of 147 under Dayo's value function. And we identified **working for the Feds** with the outcome $\langle 80,000, 60, 22\rangle$, with an evaluation of 180. Then, according to the preference structure determined by that value function, **Feds** should be strictly preferable to **M & R**. This value function should be understood to implicitly encode Dayo's preferences for tradeoffs between levels of these attributes. Dayo's particular value function indicates a preference for a conjoint improvement in **time not in traffic** and **prestige rating**, over an increase in **wages in dollars** at the expense of these other attributes. According to his value function, Dayo should be indifferent between **Feds** and the outcome $\langle 207,500, 25, 18\rangle$. In other words, even at the reduced wage level of working for the Feds, he values the extra 35 minutes he saves in traffic and the 22% far more than he does a pay bump of 45,000 dollars at the expense of his time and career prestige. Since its unlikely Moneymaker and Rich could move their offices or improve traffic conditions, they'd have to offer him a tremendous increase in compensation to offset those features. No matter whether we understand Dayo's value function to reveal his heretofore obscure preferences between outcomes, or to describe a way of consistently constructing those preferences, his value function reveals that he should prefer **Feds** to **M & R**.

Further, hierarchical decomposition characterizes a decision problem in such a way that we can readily enrich it with the structure necessary to resolve decisions under uncertainty, in the standard way. To do so, we need to elicit from the DM a set of states, and a probability function over those states, and we need to characterize her available actions as functions from states to outcomes in the consequence space. From there, we need only assess a utility function over outcomes. At first glance, it might seem that we need a method for assessing a utility function over a tuple of attributes, but this not the

case. The situation is far tidier than this. Since we have already assessed a value function, each outcome is mapped to a single real value.

Then, by composing the action functions with the value function—that is, linking them up so that actions are mapped directly to their outcome's evaluation—we can represent all actions as functions from states to single-dimensional evaluations, rather than to multi-dimensional points in the consequence space. Then, we can use standard techniques for assessing a single-attribute utility function, like those of Mosteller and Nogee 1951—not over outcomes, but over evaluations. And we can then apply ADT to characterize good choices under uncertainty; any option which maximizes subjective expected utility is a good choice.

I hedge here and talk in terms of good choices, rather than rational or ideal choices because in §4.1 above, I conceded that since the hierarchical decomposition method can result in many distinct formal representations of a given decision, or many distinct regimentations of a particular representation, the choices recommended by these models cannot be thought of as ideally rational. But I also argued that the choices recommended by these models can still be understood as good choices by the DM's own lights, that is, choices the DM has strong reasons to make. I revisit this topic in §6 below, but before pressing on, I address one final detail of hierarchical decomposition.

1.5.11 Resolving inconsistencies between assessed value functions and partial preference structures

Recent research in psychology suggests that there is a difference between *inherent* preferences and *constructed* preferences.[45] The former, also sometimes called *retrieved* preferences, are preferences that can be retrieved by a simple act of introspection. The latter are

[45] See, for example: Simonson 1989, Simonson 1990, Slovic 1995, Nowlis and Simonson 1997, Johnson, Steffel, and Goldstein 2005a, Schwarz 2007, Bettman, M. F. Luce, and Payne 2008. Much of this research is helpfully surveyed in Warren, McGraw, and Boven 2011 and Lichtenstein and Slovic 2006.

preferences that are somehow calculated or formulated in the decision making process.[46] According to Warren, McGraw, and Boven 2011, research indicates that "decision makers often retrieve existing underlying preferences in familiar situations (200)." On the other hand, constructed preferences typically emerge when the DM faces unfamiliar or complex decisions, or when the DM is put in a justificatory context.

This distinction helps to frame discussion of preferential inconsistencies that can arise under application of the hierarchical decomposition method. Whatever preexisting partial preference structure that can be elicited from the DM should clearly be thought of as comprised of inherent preferences. On the other hand, a preference structure determined by an assessed value function seems best thought of as comprised of constructed preferences. When the constructed preferences determined by the assessed value function disagree with the preexisting inherent preferences initially elicited, we might think that the only acceptable course of action is to revise the value function so that the preference structure that it determines is consistent with DM's preexisting preferences. That is, we might think that constructed preferences should always yield to inherent preferences. Yet Keeney and Raiffa 1976 suggest that in real cases of analysis, DMs sometimes accommodate such inconsistencies by instead reconsidering or revising their preexisting preferences.[47]

The decision analysis literature is, itself, somewhat lean on such example cases, but good examples can be found within the literature on preference construction. Warren, McGraw, and Boven 2011 gives an good overview of research on reconciling inherent and constructed preferences, though several of the cases there are too complex for easy discussion here. Sanbonmatsu and Fazio 1990 provide an especially compelling and simple case. In their study, participants were induced to form a preference between

[46] The term "constructed preference" is also sometimes used to refer to context-sensitive preferences, but that sense of the term isn't relevant for my purposes here.

[47] See Keeney and Raiffa 1976, Chs. 3, 7, and also Anderson and Clemen 2013.

two department stores, so that it was certain they had an inherent preference between the stores. They were then presented with information about several attributes of each store—selection, customer service, product expertise, etc.—and asked at which store they would prefer to buy a camera. At this stage, participants typically retrieved their inherent preference. But when participants were asked to engage in some rudimentary analysis, by simply weighting the various attributes according to importance with respect to the goal of buying a camera, many participants reached a constructed preference inconsistent with their inherent preference. When this occurred, participants were asked to state whether the inherent or the constructed preference best reflected their sincere preference for the store at which to by a camera; participants typically stood by the constructed preference.

If DMs are willing to revise inherent preferences in light of such a rudimentary analysis, it is easy to imagine that they might do the same in more complex cases. If we're inclined to doggedly stick to what I above called the loosely Humean idea that our preferences are unassailable or unalterable through acts of reasoning, then we might simply brand this sort of behavior as patently irrational. But stamping these cases as irrational seems hasty to me. It is important to note that DMs only carry out or subject themselves to decision analysis when they are keenly interested in making good choices that genuinely reflect their objectives and their attitude toward consequences, or in justifying their choices to some auditor. It's simply too much work to undertake for any other reasons. That is, DMs only subject themselves to these methods when they are explicitly trying to make good choices, or to justify their choices in light of their objectives. Further, research on preference construction indicates that when DMs are faced with highly complex decisions, especially when they are given ample time and asked to justify their eventual choice, inherent preferences are either nonexistent or considerably harder to retrieve, and DMs depend increasingly on constructed preferences.

So, if we brand as irrational any revision of inherent preferences resulting in constructed preferences, then we are forced to adopt some sort of error theory to cover the apparently numerous cases in which this occurs. We have to explain how an agent who is explicitly trying to make or justify her choices with respect to her objectives and her attitudes toward consequences instead directly runs afoul of them. To be sure, agents can be mistaken about their own preferences. But it seems strange that they should turn out to be mistaken precisely when they're taking tremendous care to sort their preferences out. Thus, it seems unlikely that we could give a plausible error theory here. Without some sort of error theory, though, it seems like we're simply interpreting cases to fit our preconceived account of the relationship between human agents' preferences and acts of reasoning, instead of giving an account that explains the actual cases.

On the other hand, if we take a step back and suppose that the DM who, through a process of analysis, revises her inherent preferences in the direction of the constructed preferences may be behaving reasonably, this reveals a fascinating sort of interaction between our preferences and our faculty of reason. The straightforward explanation for what's going on is that the DM generates or discovers, through an act of reasoning, reasons to revise her preexisting preferences. But in the course of the analysis, the DM is not necessarily discovering anything new about the objects of her preferences. Instead, she is primarily analyzing and reasoning through the structural relationships between her preexisting preferences for outcomes, her preferences for tradeoffs of features of her outcomes, and her objectives. Thus, it seems that the DM can reasonably modify her preferences between outcomes through deliberate acts of reason, without first learning any new non-relational properties of those outcomes.

Here, I have only sketched a rough theory of what might be going on in these cases. Giving a full account is a task for future work. Suffice to say, I think this sort of interaction between preferences and reason deserves more consideration.

1.6 Conclusion

The bulk of this chapter has been devoted to exploring decision analytic methods for modeling realistic, multiple-objective decisions as formal decision problems. But I began by exploring the question of how we can throw light on our everyday decisions. I argued that the way forward was to make progress on the specification problem: to find ways to specify formal decision problems that were aptly representative of realistic decisions. Perhaps the most important claim in this chapter is that if we apply the decision analytic methods discussed herein to model a realistic decision, we can indeed throw some light on it with ADT. I have argued that we should consider the choice recommended by such a model, together with ADT, to be a good choice for us to make, by our own lights. I have conceded that the recommendations of these models together with ADT do not characterize rational decision making in any ideal sense, but nonetheless they still inform us about which ways of deciding are better for us, and which are worse.

Allan Gibbard has cautioned me against stating this claim in a way that implies that we should care about conforming with recommendations of such models, rather than about satisfying our own preferences. We care directly about the consequences of our decisions, not about the outputs of a value function evaluated on representative points in a consequence space. So the models should not be thought of as tools for helping us sort out what we care about, because we don't care about features of the models, we care about features of our choices.

On the one hand, I agree. To select an option because it is the choice recommended by such a model and ADT is, in a sense, to choose for the wrong reasons. The choice is good for us if it satisfies our preferences, and we should choose thus because our preferences are satisfied by that option.

But on the other hand, I disagree. I think there is something important about the

models, and the choices they recommend, which makes them worth caring about more directly. We have to remember that these are models of multiple-objective decisions. A key feature of multiple-objective decisions is that it's enormously hard for a DM to sort out what she cares about, because the outcomes are tremendously complex, and the objectives can come into conflict with one another. Indeed, it is common that a DM cannot easily reconcile her raw preferences for features of outcomes with one another, nor can she see her way clear to a decision from these alone. Really, the key insight of Sartre's classic case discussed above is that when an agent is faced with conflicting objectives, it is hard to maintain her integrity and consistency as a decision maker.

We might think of describing a model through hierarchical decomposition as a way of discovering the structure of our preferences. Or we might think of it as a way of constructing and refining our preferences directly. Either way, the resultant model, and the decision it recommends represent a way for a decision maker to retain integrity and consistency. For this reason then, these models and their recommendations are of direct value to decision makers.

1.7 Appendix

In this appendix, I revisit the key concepts from §5 above in a more formal and regimented way; I do not reiterate informal characterizations of these concepts already given in the main body of the chapter. Content is adapted from Raiffa 1969, Winterfeldt and Fischer 1973a, Keeney and Raiffa 1976, Keeney 1992.

1.7.1 Preliminaries

States and probability function

Let S be a set of mutually exclusive *states*, and $p : S \longrightarrow [0,1]$ is an probability function over S.

When discussing a particular decision frame, or decision, we will assume that $\mathcal{S}$ and p represent the decision maker's subjective probability function over states of the world.

Objectives, attributes, outcomes, actions

Let $\{O_1, O_2, \ldots, O_n\}$ be a set of *objectives* with associated set of *attributes* $\mathbb{X} = \{X_1, X_2, \ldots, X_n\}$, such that, for all i, $X_i \subset \mathbb{R}$ and X_i has both a least and a greatest element. We can assume without loss of generality that X_i is positively oriented, for all i.

Then $\mathcal{X} = X_1 \times X_2 \times \cdots \times X_n$ describes a *consequence space*. An *outcome* or *point* is an ordered n-tuple of the form $\langle x_1, x_2, \ldots, x_n \rangle \in \mathcal{X}$.

An *action* $a : S \longrightarrow \mathcal{X}$ is a function from states to outcomes. An action a is certain, if and only if, $a(s) = \mathbf{x}'$ for all $s \in S$, and some $\mathbf{x}' \in \mathcal{X}$.

1.7.2 Preference structures, partial preference structures, and dominance

A binary relation $\succ$ over $\mathcal{X}$ is a *preference structure* if and only if it is a weak order over $\mathcal{X}$—that is, it is reflexive, transitive, and complete over $\mathcal{X}$.A binary relation $\gg$ is a *partial preference structure* if and only if it is a partial order over $\mathcal{X}$—that is, it is reflexive and transitive over $\mathcal{X}$. Every preference structure is a partial preference structure, but the converse is not true.

An outcome $\mathbf{x}'$ dominates the outcome $\mathbf{x}''$, if and only if, $x_i' > x_i''$, for all i. We can extend the notion of dominance to actions in the natural way: an action a' dominates an action a'', if and only if, $a'(s)$ dominates $a''(s)$ for all $s \in S$.

A partial preference structure $\gg$ is *admissible*, if and only if, if an outcome $\mathbf{x}'$ dominates the outcome $\mathbf{x}''$, then $\mathbf{x}' \gg \mathbf{x}''$.

1.7.3 Decision frames, pre-decisions, and decisions

A *decision frame* is a tuple of the form $\mathscr{D} = \langle S, p, \mathcal{X}, A \rangle$, where S is a set of mutually exclusive states, p is probability function over S, $\mathcal{X}$ is a consequence space, and A is a set of actions from S to $\mathcal{X}$. A *pre-decision* is a tuple of the form $\mathfrak{D} = \langle \mathscr{D}, \gg \rangle$, where $\mathscr{D}$ is a decision frame, and a $\gg$ is a partial preference structure over $\mathcal{X}$. A *decision* is a tuple of the form $D = \langle \mathscr{D}, \succ \rangle$, where $\mathfrak{D}$ is a decision frame, and a $\succ$ is a preference structure over $\mathcal{X}$.

Any pre-decision can be extended to a decision by extending the associated partial order $\gg$ to a weak order.

A decision frame, pre-decision, or decision is *certain* if and only if S contains only 1 element, or all actions in A are certain. For a certain decision D, the set of best choices is the equivalence class of $\succ$-maximal elements.

1.7.4 Value functions

A *value function* $v : \mathcal{X} \longrightarrow \mathbb{R}$ is a real-valued function over a consequence space, with an associated preference structure $\succ_v$, which satisfies the following conditions for any $\mathbf{x}', \mathbf{x}'' \in \mathcal{X}$, (i) $v(\mathbf{x}') > v(\mathbf{x}'')$ if and only if if $\mathbf{x}' \succ_v \mathbf{x}''$, and (ii) $v(\mathbf{x}') = v(\mathbf{x}'')$ if and only if

neither $\mathbf{x}' \succ_v \mathbf{x}''$ nor $\mathbf{x}'' \succ_v \mathbf{x}'$.

Note that a value function uniquely determines a preference structure, but a preference structure determines a value function only up to positive linear transformation. That is, if v is a value function determined by a particular weak order, then so is $a \cdot v + b$ for any positive real value of a and non-negative real value of b.

1.7.5 Decomposing attribute sets

The objective, in decomposing an attribute set, is to identify a partition over that set, such that value functions can be assessed independently over each cell in that partition.

The corresponding tradeoffs condition

Let $\{X, Y\}$ be the set of attributes. Choose four arbitrary points in the consequence space determined by X and Y of the form $\langle x_1, y_1 \rangle$, $\langle x_1, y_2 \rangle$, $\langle x_2 y_1 \rangle$, and $\langle x_2, y_2 \rangle$.

Assume that (i) at $\langle x_1, y_1 \rangle$ an increase of b in Y is worth a payment of a in X; (ii) at $\langle x_1, y_2 \rangle$ an increase of c in Y is worth a payment of a in X; (iii) at $\langle x_2 y_1 \rangle$ an increase of b in Y is worth a payment of d in X. Then, if at $\langle x_2, y_2 \rangle$ an increase of c in worth a payment of d in X, and this holds no matter which values we choose for x_1, x_2, y_1, y_2, then X and Y satisfy the *corresponding tradeoffs condition.* (See Keeney and Raiffa 1976, 90, for more on the corresponding tradeoffs condition.)

If X and Y satisfy the corresponding tradeoffs condition, then we can consider preferences for levels of X independent of preferences for levels of Y. It follows that we can assess single-attribute value functions v_X and v_Y respectively.

More importantly, given two attributes, X and Y, an additive value function of the form

$v = v_X + v_Y$, where v_X and v_Y are value functions over X and Y respectively, exists if and only if X and Y satisfy the corresponding tradeoffs condition.

Conditional preference relation

Let Y, Z form a two-cell partition over a set of attributes $\mathbb{X} = \{X_1, X_2, \ldots, X_n\}$. Since we can arbitrarily permute the indices over $\mathbb{X}$, we can assume without loss of generality, that $Y = \{X_1, X_2, \ldots, X_s\}$ and $Z = \{X_{s+1}, X_{s+2}, \ldots, X_n\}$. Recall that $\mathcal{X}$ is the consequence space associated with the set of attributes $\mathbb{X}$. Then we can represent any $\mathbf{x} \in \mathcal{X}$ as $\mathbf{x} = \langle \mathbf{y}, \mathbf{z} \rangle$, where $\mathbf{y} = \langle x_1, x_2, \ldots, x_s \rangle$ and $\mathbf{z} = \langle x_{s+1}, x_{s+2}, \ldots, x_n \rangle$.

Given such a partition, Y, Z, we can say that $\mathbf{y}'$ is *conditionally preferred* to $\mathbf{y}''$ given $\mathbf{z}'$ if and only if $\langle \mathbf{y}', \mathbf{z}' \rangle$ is preferred to $\langle \mathbf{y}', \mathbf{z}' \rangle$, and $\mathbf{y}'$ is *conditionally indifferent* to $\mathbf{y}''$ given $\mathbf{z}'$ if and only if $\langle \mathbf{y}', \mathbf{z}' \rangle$ is indifferent to $\langle \mathbf{y}', \mathbf{z}' \rangle$. Let a *conditional preference relation* $\succ^{Y}_{\mathbf{z}'}$ be the relation of conditional preference over Y given $\mathbf{z}'$.

Preferential independence, and mutual preferential independence

For some two-cell partition Y, Z over $\mathbf{X}$, Y is *preferentially independent* of Z if and only if $\succ^{Y}_{\mathbf{z}'}$ is invariant over values of $\mathbf{z}'$. The elements of $\mathbf{X}$ are *mutually preferentially independent* of one another if and only if, for every two-cell partition Y, Z of $\mathbb{X}$, Y is preferentially independent of Z.

1.7.6 Simplifying value functions, additive value functions

Where Y, Z are mutually preferentially independent, we can independently assess value functions v_Y and v_Z over the $\mathbf{y}$-space and the $\mathbf{z}$-space, and there will be an acceptable

valuation function of the form $v = f(\langle v_Y, x_Z \rangle)$, where f is some real-valued function over $v_Y \times v_Z$.

Consider a n-element set of attributes $\mathbb{X}$, with associated consequence space $\mathcal{X}$. Let v_i be an acceptable single-attribute value function over the attribute $X_i \in \mathbb{X}$. A *additive value function* over $\mathcal{X}$ is a value function of the form

$$v(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i v_i(x_i)$$

where λ_i is a non-zero real number. λ_i is referred to as a *scaling factor*.

An acceptable additive value function over $\mathcal{X}$ exists if and only if either:

(i) $n = 2$ and the corresponding tradeoffs condition is met by the two elements of $\mathbf{X}$; or

(ii) $n > 2$ and the elements of $\mathbf{X}$ are mutually preferentially independent.

Crucially, when there are only 2 attributes, mutual preferential independence is *not sufficient* for the existence of an additive value function, and the stronger corresponding tradeoffs condition must be met.

For $n = 2$ the implication from left to right is obvious; clearly if v is an additive value function the corresponding tradeoffs condition will be met. The implication from right to left is much trickier. That result is proven in R. D. Luce and Tukey 1964. Proofs for $n > 2$ can be found in Debreu 1960, Pruzan and Jackson 1963, Fishburn 1970, and Krantz et al. 1971.

Establishing the existence of an additive value function

It is obvious that for even modest values of n, in order to establish mutual preferential independence, the number of two-cell partitions for which we need to establish preferential independence is exponential in the number of attributes. To be precise, for n attributes, we have to consider $2^{(n-1)} - 1$ two-cell partitions.

But Keeney and Raiffa 1976, 112 call our attention to a collection of results from Leontief 1947b, Leontief 1947a, Gorman 1968b, and Gorman 1968a which greatly simplifies things.

Building on the other work just cited, Gorman 1968b proves the following: Suppose Y and Z are subsets of $\mathbb{X}$ such that $Y \cap Z \neq \emptyset$, $Y \cup Z \neq \mathbb{X}$, $Y \not\subset Z$ and $Z \not\subset Y$. Then, if Y and Z are each preferentially independent of their respective complements, so are $Y \cup Z$, $Y \cap Z$, $Y \setminus Z$, $Z \setminus Y$, and $(Y \setminus Z) \cup (Z \setminus Y)$

From this it follows that we can reduce the number of two-cell partitions we have to consider to at most $n - 1$. For example, Keeney and Raiffa 1976, 114 notes that if each pair of attributes $\{X_i, X_{i+1}\}$ for $i < n$ is preferentially independent of its complement, then it follows that the elements of $\mathbb{X}$ are mutually preferentially independent.

In a four attribute case, like Chandeep's above, if we can establish that both $\{X_1, X_2\}$, $\{X_2, X_3\}$, or both $\{X_2, X_3\}$, $\{X_3, X_4\}$, or both $\{X_1, X_3\}$, $\{X_3, X_4\}$ are each preferentially independent of their complementary sets of attributes then it follows from Gorman's result that all four attributes are mutually preferentially independent. So, perhaps we could have assessed an additive value function in Chandeep's case after all.

1.7.7 Assessing value functions

Techniques for assessing value functions are familiar in the literature on decision theory and decision analysis. I present some common techniques here. Throughout, we'll suppose we are assessing a value function for a pre-decision, $\mathfrak{D}$, since realistic decision makers often have only partial preference over the consequence space of their decision.

Assessing value functions over a single attribute

Let X be an attribute with greatest element x^+ and least element x^-. Since, value functions correspond to interval scales[48] over outcomes, we can begin by arbitrarily assigning $v(x^+) = 1$ and $v(x^-) = 0$. Then we elicit from the DM the *midvalue point* x^*, between x^- and x^+. This is the point such that the DM thinks there is the same change in value to increase from x^- to x^*, as from x^* to x^+. We assign $v(x^*) = \frac{1}{2}$. We then elicit midvalue points between x^- to x^*, and x^* to x^-, and assign these values of $\frac{1}{4}$ and $\frac{3}{4}$ respectively. We iterate until we have sufficient data to use curve fitting techniques to describe a value function over the data. Then we check the assessed function for consistency against the partial preference order $\gg$. If there are inconsistencies, these are generally accommodated by adjusting the value function, and the process is repeated until the function v is consistent with $\gg$.

Here it can be extremely helpful if there is another attribute Y, which is preferentially independent of X and vice versa, over which a value function has already been assessed. Then, we can describe the point x^* as the point for which the decision maker would be willing to pay the same value in Y to go from x^- to x^* as from x^* to x^+.

[48] See Stevens 1946 for more on scales of measure.

Assessing value functions directly over two attributes

The process of assessing a value function *directly* over even two attributes is considerably more difficult. In its broad strokes, it resembles the processed just described, but is considerably more computationally complex.

Suppose we have attributes X, Y with greatest elements x^+, y^+ and least elements x^-, y^- respectively. Then we begin by setting $v(\langle x^+, y^+\rangle) = 1$ and $v(\langle x^-, y^-\rangle) = 0$. We then elicit midvalue points in the consequence space. Note here there may be *many* such points. These are mapped to values of $\frac{1}{2}$. Here, unlike in the single attribute case, we can immediately leverage the partial order $\gg$, to guide the analysis. If $\mathbf{m}'$ is a midvalue point, then so too is any point $\mathbf{m}''$ such that $\mathbf{m}' \gg \mathbf{m}''$ and $\mathbf{m}'' \gg \mathbf{m}'$

From there we iterate, looking for midvalue points between these points and the pairs of greatest and least elements.

Suppose we have elicited a set of midvalue points M. Then we can sometimes leverage the partial order $\gg$ to simplify further analysis, by restricting our search for upper midvalue points to the set $\{\langle x, y\rangle \mid \exists \mathbf{m} \in M, \langle x, y\rangle \gg \mathbf{m}\}$, and our search for lower midvalue points to the set $\{\langle x, y\rangle \mid \exists \mathbf{m} \in M, \mathbf{m} \gg \langle x, y\rangle\}$. There is no guarantee we will find midvalue points in this way, especially when $\gg$ is sparse. But when $\gg$ is more complete, the method can be of considerable help. At very least, it can give the analyst some neighborhoods of points to direct the DMs attention to in search of midvalue points.

Once we have sufficient data we use curve fitting techniques and check for consistency, revising to accommodate inconsistency as needed.

Determining scaling factors for additive functions

Suppose an additive value function

$$v(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i v_i(x_i)$$

exists and we have assessed acceptable single-attribute value functions $v_1, v_2, \ldots, v_n$.

It remains to determine the scaling factors $\lambda_1, \lambda_2, \ldots, \lambda_n$.

If $\gg$ determines at least one indifference class with n elements, then acceptable scaling factors can be identified by choosing any n such points, and mapping those points to an arbitrary positive value. The result is a system of n linear equations with n unknowns and can be solved using standard techniques of linear algebra.

When $\gg$ does not determine such an indifference class, we can nonetheless use it to identify inequalities between the scaling factors, and elicit information from the DM to narrow in on acceptable scaling factors. This process can get quite messy.

From pre-decision to decision

Suppose we start with a certain pre-decision, $\mathfrak{D}$, for which we have assessed the value function v, consistent with $\gg$. Then v extends $\mathfrak{D}$ to the decision $D_v = \langle \mathfrak{D}, \succ_v \rangle$, where $\succ_v$ is the weak order over $\mathcal{X}$ determined by v. We then have all the structure required to identify a set of best actions under the framework of ADT.

In other words, under conditions of certainty, by assessing an acceptable value function over a pre-decision, we can thereby move to a resolvable decision.

1.7.8 Assessing utility functions

Suppose we began with an uncertain pre-decision, $\mathfrak{D}$, for which we have assessed the value function v, and which we have extended to the decision $D_v = \langle \mathfrak{D}, \succ_v \rangle$.

Then, to resolve this decision under the framework of ADT, we have to further assess a *utility function* over $\mathcal{X}$. A *utility function* is a value function that conforms to additional axioms, like those of von Neumann and Morgenstern 1944.

Conveniently, because we have already assessed a value function v, this reduces to the familiar task of assessing a utility function over a single attribute—namely, levels of v.

One simple way to carry out this assessment is as follows:[49] Let v^+ be the upper bound on values of v evaluated on elements of $\mathcal{X}$ and v^- the lower bound. Then we arbitrarily assign $u(v^+) = 1$ and $u(v^-) = 0$. We then consider the 50/50 lottery (v^-, v^+) and identify the *certainty equivalent* v^* such that the DM is indifferent between v^* and the lottery (v^+, v^-). Then we assign $u(v^*) = \frac{1}{2}$, consider the 50/50 lotteries (v^-, v^*) and (v^*, v^+), and identify their certainty equivalents, assigning these utilities of $\frac{1}{4}$ and $\frac{3}{4}$ respectively. We iterate until we have sufficient data to fit a curve, and then check for consistency with any direct preferences over lotteries we can assess from the DM.

The resultant utility function u, together with the associated decision D_v, is sufficient to identify a set of best actions under uncertainty under the framework of ADT.

[49] See Raiffa 1968, Ch. 4 for more detail on the nuances of assessing utility functions.

Chapter 2
Individual Decisions and Arrow's Theorem

Introduction

Decision makers often approach decisions with a divided mind. Rather than having clear, overall preferences between options, they evaluate them according to many criteria. Worse still, these criteria often conflict in their rankings of options. Rational choice theorists broadly agree that overall preferences are required for rational action.[1] If this requirement is correct—I will assume here that it is—then our task in such cases is to somehow aggregate our conflicting criteria into a single preference order. It is commonplace in modeling conflict, to treat conflicts as arising from a multiplicity of conflict-free preference orderings. I assume, then, that each criterion is associated with a *weak preference ordering* (or *weak order*, or *total pre-order*)[2]—a binary relation over options that is complete, reflexive, and transitive. This assumption immediately invites an analogy be-

[1] Some rational choice theorists, like Savage in his early work on the subject, simply assume that rational agents will have such preferences. (See, for example, Savage 1954.) Others, like Sen or Hammond derive this from other assumptions about choice functions, or consistency in choices, supplemented with some underlying account of the relationship between choices and preferences. (See, for example, Sen 1970, Sen 1993, and P. J. Hammond 1988. Presumably, Samuelson's Revealed Preference Theory should be understood similarly.)

[2] One might also say that we have at least complete *weak preferences* over options with respect to each criterion. We have total weak preferences over some collection of options, when for any pair of options, we can determine whether the first is at least as preferable as the second. I use the term 'weak preference' in the manner standard in the economics and rational choice literature. (See, for example: Sen 1970, 7–9.) Given some dimension of measurement, when we can judge of any two objects whether the first ranks at least as high as the second, we can determine an *ordinal scale* of measurement along that dimension. Weak preferences, then, determine an ordinal scale of preferability. That is, each option can be assigned a natural number, with preference corresponding to numerical magnitude. On an ordinal preference scale, neither ratios nor intervals between measurements carry any significance about degrees of preference. (For detailed discussion of scales of measure, see: Stevens 1946.)

tween conflict within an individual decision maker and social conflict.[3]

In fact, May 1954 suggests that the problem faced by an individual decision maker determining her overall preferences when she is "confronted with conflicting criteria applied to a set of alternatives" is formally identical to the problem of aggregating the conflicting preference rankings of "different individuals in a group (9)." Call the first problem *individual aggregation of preferences* and the second problem *social aggregation of preferences*. In general, call a problem of either sort a *preference aggregation problem*. At bottom, the problem in either case is to generate a single *aggregate weak order* from a tuple of weak orders.

May's point is that formal results about social aggregation apply equally well to individual aggregation. Let a *preference aggregation function* be a function that inputs a tuple of weak orders and outputs a single weak order. The General Possibility Theorem of Arrow 1950 establishes that no social aggregation function can satisfy all of four Arrow Conditions. I explain the Arrow Conditions in greater detail below. They are generally understood as constraints on rational social aggregation functions. Thus, Arrow's Theorem is often taken to establish that there can be no rational social aggregation function. In the literature on Arrow's Theorem, it is often further assumed that rational resolution of aggregation problems must be characterized by an aggregation function. Given this *functionality assumption*, Arrow's Theorem shows that there can be no rational resolution

[3]In contemporary research, May 1954 is often taken as the *locus classicus* for the analogy between individual decision makers and collectives, without any deeper discussion of the history of this idea. But the philosophical analogy between individuals and states predates May 1954 by more than two thousand years, dating back at least to Plato. Steedman and Krause 1985 note that in *The Republic*, Plato draws "an analogy between the conflicting aspects of the individual, and the citizens of a state and thus between that which integrates the person and the government of the state (199)." Plato introduces the city-soul analogy in Book II [386c] and develops the idea further in relation to the five constitutions beginning in Book VIII [approximately 544d] and continuing through his discussion of tyranny and the tyrant in Book IX. In his *Fifteen Sermons*, Butler gives perhaps the most sustained discussion of the individual-to-state analogy. (See, in particular, Sermons I and III.) He argues that, like the state, the individual decision maker must somehow reconcile a self divided by "reason, several appetites, passions, and affections prevailing in different degrees of strength" to act as one. (From Sermon III, "Upon the Natural Supremacy of Consciences"–Rom. ii. 14; Butler 2006, 14.)

of social aggregation problems.

Philosophers have argued that Arrow's Theorem is equally pessimistic for individual aggregation problems. For example, Steedman and Krause 1985, Kavka 1991, and Ellis 2006 all argue along these lines. These authors share the strategy of arguing that the Arrow conditions are also constraints on rational individual aggregation functions.[4] Thus, they conclude that there can be no rational individual aggregation function.

To the contrary, I argue that Arrow's Theorem does not rule out the possibility of rational individual aggregation functions. I argue further that individual aggregation functions which violate Arrow's independence condition can be substantively rational. May, Steedman and Krause, Kavka, and Ellis all take Arrow's Theorem to close the door on modeling individual decisions under conflicting criteria as aggregation problems. Here, I reopen the door to this modeling approach.

I begin by introducing the required notation and formalism. Next I introduce the Arrow conditions and present Arrow's Theorem. As it turns out, Arrow's independence condition is widely misstated, misunderstood, and poorly motivated. Fruitful discussion of the condition requires a more precise understanding of what the condition requires, and what a violation of the condition looks like. I address this in the third section. In the fourth and fifth sections of the chapter, I argue for my central claim that Arrow's independence condition is not a constraint on rational individual aggregation functions. In the fourth section, I review the standard arguments used to motivate the condition as a constraint on rational social aggregation functions, and I argue that none of these arguments can be extended to the case of individual aggregation. In the fifth section, I argue di-

[4]These authors embrace functionality to differing degrees. For example, Steedman and Krause seem to take functionality for granted. Thus, their chief concern is to characterize better and worse sets of conditions an individual aggregation function might meet, short of being rational. Whereas Ellis assumes that there must be rational means of resolving individual aggregation problems, and takes Arrow's Theorem to count against functionality. Kavka occupies an intermediate position. While he allows that non-functional means of resolving individual aggregation problems might be rational, he also argues that Arrow's Theorem might explain certain predictable irrationalities exhibited by human agents, like intransitive preferences.

rectly for rational individual aggregation functions which violate Arrow's independence condition. I conclude by discussing prospects for modeling individual decisions under conflicting criteria as aggregation problems.

2.1 Notation and formalism

Let $X = \{x, y, z, \dots\}$ denote the set of all options.[5] This definition of X is loose in the extreme, leaving room for construals under which X contains all logically possible options, all conceivable options, all initially available options, and so on. For now, I leave the precise scope of X unspecified, though later greater precision will prove important. Let $S \subseteq X$, $S \neq \emptyset$, where S denotes the set of feasible options. Call S the *agenda*. Unless context indicates otherwise, assume that $S = X$—that is, that all options are feasible. Let $N = \{1, 2, \dots, n\}$, denote a set either of individuals, or of some individual's decision criteria, as determined by context, with $n \geq 2$. Let R be a binary relation on X. That is, R is a set of ordered pairs of elements from X. R will be understood to be a preference relation. When written without an index, R will be understood to be the aggregate preference relation—the social preference relation in the case of social aggregation, and the individual's overall preference relation in the case of individual aggregation. Where context is ambiguous, superscripts will be used to differentiate aggregate orders. When written with an index, R_i will be understood to be either the preference relation of individual i, or the preference relation determined by criterion i. That $\langle x, y \rangle$ is an element of R will be denoted xRy; that $\langle x, y \rangle$ is not an element of R will be denoted $\neg xRy$.

R is reflexive if and only if, for all $x \in X$, xRx. R is complete if and only if, for all $x, y \in X$, xRy or yRx. R is transitive if and only if, for all $x, y, z \in X$, if xRy and yRz then xRz. When R is reflexive, complete, and transitive, we will say R is a weak preference relation. Suppose R is a weak preference relation. Let P denote the anti-symmetric part

[5] I adopt the formalism of Gaertner 2009, which is more-or-less standard, and exceptionally clearly presented. (See, in particular, section 1.3.)

of R. That is, xPy if and only if xRy and $\neg yRx$. P will be a strict (but not necessarily complete) preference relation over X. Let I denote the symmetric part of R. That is, xIy if and only if xRy and yRx. I will be a (not necessarily complete) indifference relation over X.

Let $\mathcal{P} = \langle R_1, R_2, \ldots, R_n \rangle$ be a tuple of weak preference orders over X. Call $\mathcal{P}$ a *profile* over X.

Where R is a binary order over X, and $A \subset X$, let $R\rfloor_A = \{\langle x, y \rangle \mid x, y \in A \text{ and } xRy\}$. Call $R\rfloor_A$ the restriction of R to A. It is trivial that if R is a weak order over X, then the restriction $R\rfloor_A$ is a weak order over A. Similarly, where $\mathcal{P}$ is a profile over X, let $\mathcal{P}\rfloor_A = \langle R_1\rfloor_A, R_2\rfloor_A, \ldots, R_n\rfloor_A \rangle$. Call $\mathcal{P}\rfloor_A$ the restriction of $\mathcal{P}$ to A.

Let $\mathfrak{P}$ be the set of all possible profiles over X. Let $\mathfrak{R}$ be the set of all weak preference orders over X. Let $f : \mathfrak{P} \longrightarrow \mathfrak{R}$ be a function from profiles to weak preference orders. Call f an *aggregation function*.[6] Let $f(\mathcal{P})$ refer to the image of profile $\mathcal{P}$ under f—that is, the weak preference order returned when f is evaluated at the profile $\mathcal{P}$. Since $f(\mathcal{P})$ is itself a weak order, it will sometimes be convenient to write $xf(\mathcal{P})y$ to indicate that x is related to y by the weak order f returns on profile $\mathcal{P}$.

Let $\mathcal{X}$ be the set of non-empty subsets of X. A choice function is a function $c : \mathcal{X} \longrightarrow \mathcal{X}$, subject to the restriction that $C(S) \subseteq S$.

An element $x \in S$ is a *best element* of S with respect to binary relation R if and only if for all $y \in S$, xRy. For short, call these the R-best elements. The R-best elements are at least as good as all other elements with respect to R. The set of all R-best elements in S is called the *choice set* of R on S, and is denoted $C(S, R)$. Every binary relation R on X thus determines a choice function on X in the obvious way. The converse is false. In general, a choice function does not uniquely determine a binary relation. For example, no binary

[6] For simplicity's sake, I have adopted a formalism that does not permit the set of individuals or alternatives to vary. To be clear, this is a perfectly standard formalism. Gärdenfors 1973 provides an excellent, and easy to follow example of a formalism which does allow these objects to vary.

relation corresponds to the choice function defined by $C(\{x,y,x\}) = \{x\}, C(\{x,y\}) = \{y\}$.[7] A choice function determined by a binary relation is called a *representable* choice function.[8]

2.2 Arrow's General Possibility Theorem

Arrow's Theorem demonstrates inconsistency between the following conditions.[9]

2.2.1 The Arrow Conditions

Let $\mathcal{P} = \langle R_1, R_2, \ldots, R_n \rangle$, $\mathcal{P}' = \langle R'_1, R'_2, \ldots, R'_n \rangle$ be profiles. Let f be some arbitrary aggregation function. Let $R = f(\mathcal{P})$, $R' = f(\mathcal{P}')$. That is, R and R' are the aggregate orders returned by aggregation function f evaluated at profiles $\mathcal{P}$ and $\mathcal{P}'$ respectively. Let P be the strict preference relation corresponding to R.

(U) Unrestricted domain For all $\mathcal{P} \in \mathfrak{P}$, $f(\mathcal{P})$ is defined.

(P) Weak Pareto principle If, for all $i \in N$, xP_iy, then xPy.

(D) Non-dictatorship There is no $i \in N$ such that, for all $\mathcal{P} \in \mathfrak{P}$, and for all $x, y \in X$, if xP_iy then xPy.

[7] Sen 1982, 17.

[8] See: Sen 1970, Sen 1982.

[9] The conditions I present here differ in strength and substance from those presented in Arrow 1950. They are both weaker and more generally stated. By now, though, they have become more or less standard in the literature on Arrow's Theorem, and are the conditions commonly used in primers on the subject. (See, for example: Kelly 1987, 80–87; Gaertner 2009, 20; and Sen 2014a, 34.) One can prove results highly similar to Arrow's Theorem from weaker sets of conditions. In particular, Hansson 1973 shows that one can get an almost equivalent theorem by replacing the weak Pareto principal with two conditions the conjunction of which is still weaker than weak Pareto. And Wilson 1972 shows that one can prove a somewhat weaker theorem without the use of the weak Pareto principle or any nearby replacement conditions. Wilson's theorem is weaker in the sense that an aggregation function which satisfies his conditions of unrestricted domain and independence of irrelevant alternatives won't necessarily be a dictatorial function, and may instead be what he terms a null-function. (See: Wilson 1972, 480.) All the same, I opt for these now-standard conditions for a simple reason: I don't care much about any condition other than independence of irrelevant alternatives, and this condition is indispensable to every result closely related to Arrow's Theorem.

(I) Independence of irrelevant alternatives For any $x, y \in S$, and for all $i \in N$, if xR_iy iff xR'_iy, then xRy iff $xR'y$.

Then Arrow's Theorem is as follows:

Arrow's Theorem Given at least three elements in S, and with N finite,[10] no aggregation function f can satisfy all of: (U), (P), (D), and (I).[11]

Or as Sen 2014a so elegantly puts it: "a social choice function that satisfies unrestricted domain, independence of irrelevant alternatives, and Pareto Princple has to be dictatorial (34–35)."

2.2.2 Why relation theory rather than choice theory?

Readers familiar with Arrow 1951 or Arrow 1963 will recognize that I depart from Arrow by giving a relation theoretic rather than choice theoretic definition of condition (I). More generally, I discuss aggregation functions which return binary relations rather than choice sets, and present a relation theoretic version of Arrow's Theorem rather than a choice theoretic version.[12] I have two simple reasons for this: Relation theoretic results can easily be extended to choice theoretic results, but the converse is not true. Also, I aim to make room for a modeling approach for rational individual decisions under conflicting criteria. The hope is that the output of such a model could be plugged into standard decision theoretic models of rational decision making under uncertainty, which generally require that the individual decision maker have a single overall weak order over her options.

[10] Fishburn 1970 shows that given an infinite number of voters, it is possible for a social choice function to satisfy all four conditions.

[11] I omit a proof of Arrow's Theorem here. The interested reader should see Sen 2014a, pp. 34–37, for an especially concise and elegant proof, or Gaertner 2009, 21–34 for a presentation of several different proofs. Arrow 1950 proves only the restriction of the theorem to the two-voter case.

[12] See Hansson 1973 for several choice theoretic versions of the theorem.

2.2.3 Motivating the first three Arrow conditions with respect to social aggregation

Condition (U) requires that an aggregation function be defined for all possible preference orders over the set of options. That is, it requires that no possible preferences be excluded in advance; individuals or decision criteria can rank options however they see fit. Any aggregation function which violates (U) rules out certain means of participation in the aggregation problem. Worse, as we shall see below, aggregation functions which violate (U) can fail to yield actionable or rational collective preferences, by returning empty orders, or orders that are problematically intransitive.

Condition (P) requires that consensus be respected wherever it arises. For this reason, this condition is also sometimes called the *consensus principle*.[13] The condition requires that unanimous individual strict preference entail aggregate strict preference. Any aggregation function that fails to satisfy (P) might give rise to situations in which, despite unanimous weak preferences for one candidate over another, the aggregate order is indifferent or strictly prefers the second candidate to the first. The appeal of (P) is as obvious as it is indisputable.[14]

Condition (D) requires that an aggregation function does not simply parrot the strict preferences of one individual or individual criterion. Maskin 2014 notes that (D) is typically justified as an entailment of a much stronger condition, generally called *anonymity* and in the context of elections *equal treatment of voters* (50).

Anonymity If $\mathcal{P}'$ is permutation of $\mathcal{P}$ then, $f(\mathcal{P}) = f(\mathcal{P}')$.

A thorough defense of anonymity could fill its own book, but the gist of the reasoning is

[13] See: Maskin and Sen 2014, 46.

[14] The Pareto principle is not entirely without criticism as a constraint on collective decision. Existing criticism divides quite neatly into two camps. To one side, some authors have questioned whether apparent unanimity encoded in agreement of pairwise preferences across individuals should always be taken at face value. (See, for example, Mongin 2005.) To the other side, some authors argue that the Pareto principle conflicts with principles of fairness and justice. (See, for example, Kaplow and Shavell 1999, Kaplow and Shavell 2000, and Kaplow and Shavell 2003.) However, these criticisms plainly have no bearing on the discussion here.

fairly simple. In cases of social aggregation, aggregation of preferences of *moral agents* is at stake. Moral agents are the natural subjects of considerations of fairness and justice. Standardly, considerations of fairness and justice require that individual moral agents are treated equally unless we have principled reasons to deny them equal standing. That is, how a moral agent should be treated is not dependent on who in particular the agent is. In other words, considerations of fairness and justice impose a defeasible presumption that moral agents are to be treated anonymously. In social aggregation problems, weak orders are the only information we have about the individual moral agents whose preferences are to be aggregated. On the basis of this information alone, there is generally no principled reason to deny the moral agents involved equal standing. Thus we should adhere to something like the anonymity condition, which entails non-dictatorship.

I defer further discussion of condition (I) to the next section.

2.2.4 Motivating the conditions with respect to individual aggregation

Conditions (U) and (P) can be motivated as constraints on rational individual aggregation functions by straightforward analogues of the arguments used to motivate them in the social case.

Condition (D) demands just a bit more thinking. Dictatorship is objectionable in the case of social aggregation because, as noted above, the individuals whose preferences are to be aggregated are moral agents subject to considerations of justice and fairness. On basis of these considerations we can argue for their equal treatment by aggregation functions, and argue against specific ways of treating them differently like dictatorship.

But the decision criteria in an individual aggregation problem are not agents, let alone moral agents. They are not even quasi-autonomous subagents. Instead, they are something like value considerations. There are no natural analogs of justice and fairness from which to argue for equal treatment of the decision criteria. Indeed, there's good reason to

think that some of our decision criteria—say our moral and justice-related evaluations of options—should be treated as distinctly more important than our other decision criteria.

Nonetheless, there's a fairly compelling reason to impose (D) as a constraint on individual aggregation functions. To condone dictatorial aggregation functions is to give up on saying anything interesting about decision under conflicting criteria. Suppose we concluded there was one rational aggregation function, namely the one which made the decision maker's moral evaluation of options the dictator. Conflict with other criteria might lead to all sorts of psychological fallout for the decision maker, like regret, shame, dissociation, perhaps even an eventual break from her previous moral commitments. But these purely formal conflicts are completely swept aside when it comes to the practical question of what to choose. On that question, the dictator's evaluation is final. And in that sense, all conflicts would be obviated. Similarly, conflicts would be obviated whenever the individual decision maker had any kind of fully worked-out weighting or priority scheme over her decision criteria. This point will prove important later.

Thus, if the aim is in fact to say something about decision under conflicting criteria in the discussion of individual aggregation problems, we must impose (D) as a constraint on individual aggregation functions. Note well, though, that (D) is not motivated by considerations of rationality. The motivation for (D) is altogether methodological. It remains possible that violations of (D) might be rational.

Again, I defer discussion of condition (I) to the next section.

2.2.5 Two familiar aggregation functions

To illustrate the significance of Arrow's Theorem, I briefly introduce two familiar aggregation functions, and indicate which of the Arrow Conditions they violate.

The Condorcet Rule

Let R^C be a binary relation over X, with associated strict preference relation P^C.[15] Define $C_y^x = \left|\{i \mid xR_iy\}\right|$. (Here, where A is a set, $|A|$ denotes the cardinality of A.) Then xR^Cy if and only if $C_y^x \geq C_x^y$. And xP^Cy if and only if xP_iy holds for a simple majority of $i \in N$.

Condorcet Rule $f_C : \mathfrak{P} \longrightarrow \mathfrak{R}$, such that $f_C(\mathcal{P}) = R^C$.

The Condorcet Rule plainly satisfies (P), (D), and (I). It violates (U). Condorcet noted that the rule fails to return a weak order over so-called *cyclic profiles*. This is easiest to illustrate in the three individual, three option case. Let $\{x, y, z\}$ be the set of candidates. Adopt xP_iyP_iz as shorthand for xP_iy, yP_iz and xP_iz. Suppose the strict preferences of the three individuals are xP_1yP_1z, yP_2zP_2x, and zP_3xP_3y. Then a majority of individuals prefers x to y, a majority prefers y to z, and a majority prefers z to x. Thus the binary relation R^C determined by this profile is not a weak order, since it is cyclic without also being flat, and is therefore intransitive.[16]

The Borda Rule

For notational convenience when discussing the Borda Rule, we will assume that $X = \{x_1, x_2, \ldots, x_c\}$.[17] Let $\mathbf{B}_i$ be a c-dimensional vector, such that $\mathbf{B}_i = \langle b_1, b_2, \ldots, b_c\rangle$ and define

[15] The Condorcet Rule takes its name from the Marquis de Condorcet, who discusses it at length in his monograph of 1785, *Essay on the Application of Analysis to the Probability of Majority Decisions*. However, identical or highly similar rules were previously discussed in the works of Ramon Llull (in the thirteenth century), Nicolas Cusanus (in the fifteenth century), and Pufendorf (in the seventeenth century.) (See: Gaertner 2009, 3–6.)

[16] That the Condorcet Rule violates (U) depends on the fact that I define the range of aggregation functions as the set of weak preference orders. Cyclic profiles reveal that R^C is sometimes not a weak preference order. If, however, one characterizes aggregation functions as functions from tuples of weak preference orders to mere binary relations (see, for example, Gaertner 2009), then the Condorcet Rule satisfies (U). Nonetheless, it's still deficient since it sometimes returns inconsistent aggregate orders, which are generally regarded as problematically inconsistent and inadequate for rational choice. However, I prefer my characterization of aggregation functions, for the simple reason that on this characterization, Arrow's Theorem covers all aggregation functions, instead of having exceptions in rules like Condorcet. Others, like Sen 2014b characterize (U) as I do here.

[17] The Borda Rule takes its name from Jean-Charles de Borda, who discusses it at length in his 1781 essay, *Memoir on the Election by Ballot*. Again, identical or highly similar rules were previously discussed by Llull,

$b_j = \left|\{y \in X \mid x_j P_i y\}\right|$. Let $\mathbf{B}^+ = \sum_{i=1}^{n} \mathbf{B}_i$. Then, $x_p R^B x_q$ if and only if $\mathbf{B}_p^+ \geq \mathbf{B}_q^+$. That is, $x_p R^B x_q$ if and only if the p-th component of $\mathbf{B}^+$ is at least equal to the q-th component.

Borda Rule $f_B : \mathfrak{P} \longrightarrow \mathfrak{R}$ such that $f_B(\mathcal{P}) = R^B$.

Put more expressively, the Borda Rule works like this: First, convert each preference order in the profile into a positional scoring, where each option's positional score is the number of options to which it is strictly preferred.[18] Then, for each option, sum the positional scores it receives from each individual or decision criterion. Call this the option's *aggregate positional score*. Then one option is weakly preferred to another in the aggregate order if and only if the first option's aggregate positional score is greater than or equal to the second's aggregate positional score.

The Borda Rule plainly satisfies (P) and (D). It also satisfies (U), since on cyclic profiles it returns the flat order,[19] rather than an intransitive, cyclic order. However, it violates (I), since aggregate preferences between x and y depend on the aggregate positional scores of x and y, which in turn depend on how other options are ranked relative to x and y. This will become clear from examples in the next section.

Cusanus, and Pufendorf.

[18] Gärdenfors 1973, B. Fine and K. Fine 1974, P. J. Hammond 1987, and Pattanaik 2002 all note that there is ambiguity regarding how to define an option's positional score when the profile is allowed to contain non-linear weak orders. A weak order R with associated strict order P is linear if and only if for all x, y, with $x \neq y$, either xPy or yPx. Given a linear order over the options, the natural positional score for each option is the number of options to which it is strictly preferred on that order. But when the order is non-linear there are at least two ways to generate positional scores: each option can be scored on basis of the number of options to which it is strictly preferred, or on basis of the number of disjoint indifference classes beneath it. Two options x, y are in the same indifference class if and only if xIy. These two scoring methods can return different scores. Let $X = \{x, y, z\}$. Suppose xPy and yIz. Then scored by the strict preference method, x receives a positional score of 2 since it is strictly preferred to both y and z. Scored by the indifference class method, x receives a positional score of 1, since it is strictly preferred to elements from the indifference classes $\{x, y\}$. It is trivial that these methods converge on the same score for linear orders. Here I follow P. J. Hammond 1987 and Pattanaik 2002 and opt for the strict preference scoring method, which more closely approximates how linear orders would be scored in cases where weak orders are very close to linear.

[19] An order is flat when nothing is strictly preferred to anything else. Everything is weakly preferred to everything; the options are all in the same indifference clance.

Violations of (P) and (D)

It seems that no aggregation function actually in use as a voting rule violates either (P) or (D). But such aggregation functions are easy to imagine. The function that maps every profile to its first element violates (D). And any *imposed* function will violate (P). An aggregation function is imposed, in Arrow's sense, if it the aggregate order it returns is not dependent on the profile on which it is evaluated.[20] So a function that maps every profile to the same fixed aggregate order is imposed, and thus violates (P).

2.3 The Independence Condition

Hansson 1973 writes, "Arrow's Theorem is really a theorem about the independence condition (25)." Condition (D) is quite reasonably taken as sacrosanct. Dasgupta and Maskin 2014 and others have shown that there are restricted domains such as the domain of single peaked preferences and the domain of single troughed preferences on which certain aggregation functions—in particular the Condorcet rule—satisfy all the other Arrow conditions. But we generally lack good reason to expect individual preferences to fall into such domains, thus we generally lack good reason to reject or relax condition (U). Condition (P) seems just as sacrosanct as (D). Further, Hansson proves versions of Arrow's Theorem that involve substantial weakenings of the weak Pareto principle. Relatedly, Wilson 1972 proves a nearby impossibility theorem that doesn't require any version of the Pareto principle. So, despite its sacred role, (P) isn't exactly essential to the theorem.

The appeal of condition (I) is not so obvious. In fact, Arrow 1983 admits that if one means to object to one of his conditions, (I) is the natural starting point. Despite this, the condition is widely taken for granted. Mackie 2003 notes that "justifications of the condition are typically thin and dogmatic, often no more than an assertion that its

[20] Arrow 1950, Arrow 1951, and Arrow 1963 actually contain an explicit *non-imposition* condition, though it later became clear that the weaker Pareto principle was sufficient for the impossibility result. See: R. D. Luce and Raiffa 1985, 329 for further discussion of imposition and dictatorship.

appeal is intuitively obvious (123)." For example, Arrow 1983 states that the "essential argument" in favor of this condition is its direct appeal to intuition (51).

Barry and Hardin 1982 suggest that it is perhaps because the condition is so subtle that it is so readily taken for granted. As evidence of its subtlety consider that Nash 1950, Goodman and Markowitz 1952, Radner and Marschak 1954, R. D. Luce 1956, R. D. Luce and Raiffa 1957, Vickrey 1960, Samuelson 1967, Dummett 1984, among others, have all been alleged to have confused condition (I) for another condition to which it is only indirectly related,[21] which has sometimes also been called independence of irrelevant alternatives. Allegations of these confusions are catalogued at length in Ray 1973, Hansson 1973, Barry and Hardin 1982, Kemp and Ng 1987, Bordes and Tideman 1991, McLean 1995, Denicolò 2000. Mackie 2003, and elsewhere.[22] And Maskin 2014 and Dasgupta and Maskin 2014 somewhat misleadingly treat Arrow's condition (I) as though it is interchangeable with Nash's condition of the same name.[23] This other condition is also known as contraction consistency, and is referred to in Sen 1970 and Sen 1982 as Property α. Let C be some choice function. Then the condition is:

(C) Contraction consistency If $S' \subset S$, $x \in C(S)$, and $x \in S'$, then $x \in C(S')$.

Indeed, in Arrow 1987, Arrow himself admits to having made the same confusion in Arrow 1950, Arrow 1951, and Arrow 1963, when he provided the oft-cited alleged example of a violation of condition (I) involving the death of a candidate in a club election.

Given the history of confusion surrounding the condition, it is somewhat surprising that anyone should claim the condition appeals obviously to the intuition. It would be one

[21] Denicolò 2000 shows that if an aggregation function is restricted in certain fairly weak ways, then it can satisfy at most one of Arrow's condition (I) and contraction consistency.

[22] Of historical interest, McLean 1995 suggests that this confusion is evident even in the work of Condorcet, and his most prominent intellectual heir, Pierre Claude François Daunou.

[23] See, in particular: Maskin 2014, 47, nt. 6; and Dasgupta and Maskin 2014, 103, nt. 6. In personal communication with Maskin, he admits that his presentation of Nash's condition as interchangeable with Arrow's was somewhat confusing. He had intended to communicate their interchangeability only with respect to the majority dominance theorem in discussion in those papers, and not their interchangeability in the larger context of Arrow's Theorem.

thing if the confusion was endemic to laypersons and non-experts, but the very content of the condition seems to have eluded sophisticated commentators who are trained experts in the field.

To better explore the motivations for condition (I), it is necessary to more carefully distinguish condition (I) from condition (C), and to discuss what violations of each condition might look like.

2.3.1 No aggregation function violates (C)

It is also somewhat surprising that anyone has ever confused conditions (I) and (C). Condition (I) is what Kelly 1987 has called an *interprofile* condition on aggregation functions. It states that if two profiles relate to one another in certain ways, then the aggregate orders generated from those profiles must relate to one another in certain ways.

Condition (C) is not even straightforwardly a condition on aggregation functions. Rather, it is a condition on much more general choice functions. Recall, though, that every binary relation determines a choice function. Let C^* be a choice function over X such that for for some aggregation function f and profile $\mathcal{P}$, and for any $S \subset X$, $C^*(S) = C(S, f(\mathcal{P}))$. Sharpening up condition (C) in the obvious way, we get something like the following condition:

(C′) Contraction Consistency on Aggregation Functions If $S' \subset S$, $x \in C^*(S)$, and $x \in S'$, then $x \in C^*(S')$.

Then we have what Denicolò 2000 calls an *interagenda* condition. (C′) states that if two agendas are related to one another in certain ways, then the choice sets from those agendas must relate to one another in a certain way. If we intend to demonstrate that an aggregation function violates (C), presumably we'll do so by showing that it violates the more specific condition (C′).

Here's Arrow's (in)famous example of an alleged violation of (I), now widely understood to be an example of a violation of (C):

> With a finite number of candidates, let each individual rank all the candidates, i.e., designate his first-choice candidate, second-choice candidate, etc. Let preassigned weights be given to the first, second, etc., choices, the higher weight to the higher choice, and then the the candidate with the highest weighted sum of votes be elected. In particular, suppose that there are three voters and four candidates, x, y, x, and w. Let the weights for the first, second, third, and fourth choices be $4, 3, 2$ and 1, respectively. Suppose that individuals 1 and 2 rank the candidates in the order x, y, z, and w, while individual 3 ranks them in the order z, w, x, and y. Under the given electoral system x is chosen. Then, certainly if y is deleted from the ranks of candidates, the system applied to the remaining candidates should yield the same result, especially since, in this case, y is inferior to x according to the tastes of every individual; but, if y is in fact deleted, the indicated electoral system would yield a tie between x and z.[24]

The voting system Arrow has in mind requires a bit more explanation: Arrow suggests that the way to respond to the removal of a candidate from the agenda is to take each individual order in the profile, blot out the name of the removed candidate from that order, and aggregate from there.

The example is plainly not a violation of (I), since to show a violation of an interprofile condition, we need an example that involves a change in profiles.[25] In light of this, commentators have often held that it is meant to show a violation of (C′). But the example is not even structured properly to demonstrate a violation of (C) or (C′). The only element in the initial choice set, x, remains in the choice set after contraction. Let $S = \{w, x, y, z\}$, and $S' = \{w, x, z\}$. For the time being, I will hedge and refer to the choice function at play in the example as C, rather than C^* with explanation to follow below. Then we have $C(S) = \{x\}$, and $C(S') = \{x, z\}$. There is no element in $C(S)$ that is not in $C(S')$. And this would be required to demonstrate a violation of (C).

[24] Arrow 1950, Arrow 1951, Arrow 1963, 25–26.

[25] Bordes and Tideman 1991 offer an ingenious, rather complicated, though eminently plausible reading of Arrow 1951 and Arrow 1963 according to which Arrow makes no such confusion. Kemp and Ng 1987 argue along similar lines. However, I take Arrow's own admission of the mistake to trump their clever interpretive efforts.

We can generate an example similar to Arrow's example with the right structure. Suppose there are five voters, rather than three. Suppose each of the individuals rank the candidates in the following orders, respectively: x, z, y, w; x, y, w, z; w, x, z, y; y, w, z, x; and y, w, x, z. When y remains a feasible candidate, x and y tie. When y is deleted from the ranks of candidates, w wins outright. $C(S) = \{y, z\}$ and $C(S') = \{w\}$. We now have a case in which $S' \subset S$, $x \in S'$ and $x \in C(S)$, but $x \notin C(S')$.

But, as, Ray 1973 explains, this example can't actually work. Sen 1970 proves it's impossible. Condition (C) is equivalent to Sen's Property α. Condition (C′) is just that condition reformulated in terms of C^*, which is the choice function determined by the weak order $f(\mathcal{P})$. Sen proves that every representable choice function necessarily satisfies Property α.[26] Recall that a representable choice function is one determined by a binary relation. $f(\mathcal{P})$ is a binary relation,[27] so C^* is a representable choice function. (C′) turns out to be an inviolable condition. The example simply can't be showing us what it appears to be showing us.

So what's going on? The answer lies in a subtlety of the voting method Arrow uses. When a candidate is removed from the agenda, Arrow suggests blotting them out of the individual orders in the profile, rather than blotting them out of the aggregate order generated by the aggregation function on the profile. Let f be the aggregation function for the voting method in the example. When the agenda contracts from S to S', he suggests we consider $f(\mathcal{P}\rfloor_{S'})$ rather than $f(\mathcal{P})\rfloor_{S'}$. Strictly speaking, I am abusing my notation a bit, since the domain of f is the set of profiles of weak orders over X, whereas $\mathcal{P}\rfloor_{S'}$ is a profile over S'. However, the meaning of the construction is obvious, and so I don't see any reason to add additional notation. Now we can see why I hedged earlier in referring to the choice function in Arrow's example as C^*. With respect to the initial agenda, S, Arrow

[26] See: Sen 1970, 17–18; and Sen 1982, 170–173.

[27] Even better, it's a weak order, so we can actually expect the choice function it determines to satisfy stricter conditions.

may very well consider C^*, where C^* is the choice function determined by $f(\mathcal{P})$. But after contraction to S', Arrow instead considers the choice function determined by $f(\mathcal{P}\rfloor_{S'})$. $f(\mathcal{P})$ and $f(\mathcal{P}\rfloor_{S'})$, are, quite obviously, different binary relations since the former is complete over S and the latter only over S'. But what is worse, for Arrow's voting method they disagree over S'. That is, for Arrow's voting method, $f(\mathcal{P}\rfloor_{S'}) \neq f(\mathcal{P})\rfloor_{S'}$. Thus the choice function determined by the latter over S' cannot be a subset of the choice function determined by the former over S. They must be two quite distinct choice functions.

Arrow's case, then, is not an example of an aggregation function violating (C) or (C′). The upshot of this is that I can safely ignore a whole class of arguments and examples intended to motivate condition (I). In particular, I can ignore any examples and arguments that treat (I) as though it is, or entails, an interagenda consistency condition like (C). I can restrict my focus to those arguments and examples that take condition (I) for what it is: an interprofile condition, violations of which depend on changes in profiles.

2.3.2 An example violation of (I)

Arrow's example above makes it clear that there is some ambiguity in how we define the Borda Rule for any case where $S \neq X$. Recall that f^B is the aggregation function associated with the Borda Rule, and let $\mathcal{P}$ be the profile under consideration. One way to generate a weak order over S is to evaluate $f^B(\mathcal{P})\rfloor_S$. That is, we can first restrict the profile to the set S, and then calculate the aggregate order from the restricted profile. Call this the *local Borda Rule.* But we might instead evaluate $f^B(\mathcal{P})\rfloor_S$. That is, we can evaluate the Borda Rule on the profile over X, and then restrict the resultant aggregate order to S. Call this the *global Borda Rule.* Arrow's example and my modified case show us that the local and global Borda rules will sometimes yield different results. The examples also illustrate a further divide between the rules. Whereas the local Borda Rule can be used to generate apparent inconsistencies under contraction, no such inconsistencies can be

generated from the global Borda Rule. However, the local Borda Rule quite obviously satisfies condition (I), whereas the global Borda Rule violates (I). Since I'm interested in violations of (I), rather than (C), hereafter I intend the global version of the rule.[28]

Here's a simple example of a violation of (I) adapted from Mackie 2003: Suppose there are five individuals, and $X = \{x, y, z\}$. Adopt the shorthand $x > y$ to indicate that on the order under discussion x is strictly preferred to y. Suppose two individuals rank the options in order of strict preference $x > y > z$, two rank them $y > z > x$, and one ranks them $z > x > y$. On this profile, the Condorcet Rule returns the cycle $x > y > z > x$, whereas the Borda Rule returns the aggregate order $y > x > z$. To show a violation of (I), we'll need to home in a on particular pair of options. Focus on x, y, and suppose that the two individuals who initially ranked the options $y > z > x$ change their ranking to $z > y > x$. Now the Condorcet Rule returns the aggregate order $z > x > y$, and the Borda Rule returns the order $z > x > y$. The aggregate order returned by the Borda Rule on the first profile ranks $y > x$. On the second profile, the aggregate order returned by the Borda rule ranks $x > y$. The Borda Rule has changed its pairwise ranking of x and y. Crucially, no individual's preferences between x and y have changed from the first profile to the second. Thus we have a violation of condition (I).

Were condition (I) to have some sort of obvious intuitive appeal, we should expect it to be evident in the example just discussed. But, at least to my mind, nothing seems obviously wrong with the behavior of the Borda Rule on the profiles in the example. Perhaps the appeal of the condition could be made more obvious by fleshing out the example with details. However, we are supposed to be troubled by violations of the condition in general, and not violations in some particular case. So we'd be moving in the wrong direction

[28] For more on the relationship between global and local versions of an aggregation function and condition (I), see Bordes and Tideman 1991. There, the authors introduce a more expressive formalism, in which they characterize a condition called *regularity*. Roughly, an aggregation function is regular whenever its global and local versions always agree. They show that any regular function must satisfy condition (I), but the converse is not true.

to exchange abstract and generic violation for a more concrete and particular case. There is supposed to be something so troubling about the ground-level formal structure of the example that we can readily grasp it with our intuition. If that's the case it should be evident no matter how spare we are with concrete details. Consider as an analogy the principle that intransitive preferences are irrational. This principle, I readily concede, appeals directly to one's intuition, and it can easily be motivated by the most abstract of examples. Condition (I) just doesn't seem to have the same degree of intuitive force.

Motivating condition (I) requires more substantive arguments. I turn now to addressing these.

2.4 Motivating condition (I): arguments and replies

Though condition (I) is widely motivated by appeal to intuition, there are several more substantive arguments on offer in the literature. While these arguments may motivate the condition with respect to social aggregation functions, the question at hand is whether they adequately motivate the condition with respect to individual aggregation functions.

Existing work on extending Arrow's theorem to individual aggregation problems either takes condition (I) as intuitively obvious (see Ellis 2006), or gestures to one of the arguments discussed below to motivate the condition without carefully exploring whether the argument actually extends to cases of individual aggregation (see Steedman and Krause 1985 and Kavka 1991.)

2.4.1 Argument 1: Condition (I) prohibits strategic misrepresentation of preferences

An individual misrepresents her preferences if she submits for aggregation a weak order that does not reflect her actual preference order over the options. Call the former an *insincere order* and the latter her *sincere order*. She misrepresents her preferences *strategically* if by way of submitting an insincere order, she actually increases the chances that an

options she prefers on her sincere order will win. This is also called *strategic voting*; for reasons that will become obvious, I avoid this term.

Riker 1961, Riker 1982, P. J. Hammond 1987, and Dasgupta and Maskin 2014 have argued that condition (I) rules out aggregation functions that are susceptible to manipulation by strategic mispresentation of preferences.[29] One has to parse the conclusion of this argument carefully, as there are several possible ways to manipulate the results of an aggregation function. In addition to strategic misrepresentation of preferences, individuals can also manipulate the agenda by causing options to be added or removed. When discussing this argument, both P. J. Hammond 1987 and Mackie 2003 slip into discussion of agenda manipulation. But it's been well known since Condorcet that even aggregation functions which satisfy condition (I), like his eponymous rule, are susceptible to agenda manipulation.

However, aggregation functions which violate (I) are susceptible to strategic misrepresentation of preferences, and uniquely susceptible to strategic misrepresentation of pairwise preferences between pairs consisting of one feasible and one infeasible option. Suppose we're aggregating by the Borda Rule, $X = \{a, b, \ldots, z\}$, and the agenda is $S = \{x, y, z\}$. Consider an individual whose sincere order is $x > y > z > a > b > \cdots > w$. Were she instead to submit the insincere order $x > a > b > \cdots > w > y > z$, she could reduce the positional score for y and z each by 23 points. This difference might be washed out in a large pool of mostly sincere individuals; but in a small pool of voters it could easily overwhelm the sincere preferences of others. Note, though, that the individual does not misrepresent her pairwise preferences over feasible options, or for that matter over the infeasible options. Rather she misrepresents her pairwise preferences over the set of pairs consisting of one option from $\{a, b, \ldots, w\}$ and either y or z. It is quite plain that it is precisely because the

[29] See: Riker 1961, 904; P. J. Hammond 1987, 122; Dasgupta and Maskin 2014, 137. For a lengthy discussion of manipulation, in particular for a careful distinction between strategic voting and manipulation of options, see: Riker 1982, 137–196.

Borda Rule violates condition (I) that it is susceptible to such manipulation. It is equally obvious that any aggregation function which satisfies condition (I) won't be susceptible to this sort of manipulation, because such a function won't be sensitive to how any element of S is ranked relative to any element of $X \setminus S$. (Here, $X \setminus S$ notates the set-difference between X and S, that is, the set of all elements of X not also in S.)

Where we have reason to worry about manipulation, especially about manipulation by this sort of strategic misrepresentation of preferences, we have corresponding reason to reject aggregation functions which violate condition (I). However, though this worry is legitimate for social aggregation, it is irrelevant for individual aggregation.

Social aggregation only allows for strategic misrepresentation of preferences because we don't have immediate access to individuals' preferences. Some way or other, they have to generate an external representation of their preferences on a ballot, and submit this to the institution in charge of aggregation. Further, individuals are able to misrepresent their preferences because they are full-fledged agents, and are able to reason about the balloting process. Neither of these conditions obtain in cases of individual aggregation. There is no separation of the individual's decision criteria from the aggregator. The aggregator is one and the same individual whose criteria are to be aggregated. The aggregator has immediate access to the weak orders generated by her decision criteria precisely because they are *her* decision criteria. Further, the decision criteria are not anything like autonomous agents capable of insincerity. As noted above, they are something like value considerations, the very essence of which is that they are always transparent in their evaluation of options.[30]

[30]For these reasons, I can safely ignore the family of results closely related to Arrow's Theorem about strategic voting and manipulability, like those of Gibbard 1973, Satterthwaite 1975, and Duggan and Schwartz 2000.

2.4.2 Argument 2: Condition (I) rules out lotteries

Riker 1982 argues that condition (I) rules out what he calls *lotteries*, which are aggregation functions that somehow incorporate an element of randomness in their mapping from profiles to aggregate orders(129–143). Random aggregation functions are objectionable because they are not responsive in the right way to changes in individual preferences. Thus, Riker's reasoning runs, we should reject aggregation functions which violate (I).

Riker, however, is unclear what he means by lotteries. There seem to be three possible interpretations. Recall that $\mathfrak{R}$ is the set of weak orders over X and $\mathfrak{P}$ is the set of possible profiles over X. First, Riker might have in mind a function that takes in a profile and returns a probability distribution over (some subset) of $\mathfrak{R}$. In fact, functions of this sort seem to be what Riker has in mind. But if the distribution returned is appropriately sensitive to differences in profile, such a function doesn't seem obviously objectionable. We might, for example, interpret it as communicating the probability, for each aggregate order in the distribution, that it was the order that would maximize social or individual utility. That's neither here nor there, though, because condition (I) is not a constraint on functions of this kind, and neither is it required to rule out functions like this. Such functions are preemptively ruled out by the formal structure of aggregation problems, because they are not aggregation functions. Aggregation functions are by definition functions from profiles to weak orders, not from profiles to probability distributions.

Second, he might have in mind a non-deterministic mapping from profiles to weak orders. Such a mapping would, for any profile input, return a weak order selected in a non-deterministic way, perhaps randomly according to a probability distribution somehow sensitive to the profile itself. But, though such a mapping has the same domain and range as an aggregation function, it is not a function, since each input is not mapped to a single output. Such a mapping might map any given profile to any number of distinct, possibly unrelated, aggregate orders. Once again, condition (I) is not required to rule out

such non-deterministic mappings; they are preemptively ruled out by the structure of aggregation problems.

There is a third possibility. Imagine that we proceeded through the space of possible profiles one by one, pairing each profile with a single, randomly-selected weak order. The resulting map would indeed be an aggregation function. This interpretation might be Riker's best option. Call these functions *Riker lotteries*. It's quite obvious why we should reject aggregation by Riker lotteries with respect to both social and individual aggregation problems. Riker lotteries fail to respond appropriately to differences in profiles. Condition (I) obviously rules out Riker lotteries.

But condition (I) is not required to rule out Riker lotteries. Indeed, there are conditions that rule out Riker lotteries, and which do not entail condition (I). This is obvious because the Borda Rule violates (I), but is not a Riker lottery. We can generate any number of other conditions that rule out Riker lotteries. In principle, all that's required is that we can predict how $f(\mathcal{P})$ and $f(\mathcal{P}')$ will differ on the basis of information about how $\mathcal{P}$ and $\mathcal{P}'$ differ. Or, to put the point somewhat differently, what's required is that given $\mathcal{P}$ we can somehow deterministically calculate $f(\mathcal{P})$. It is outside the scope of this essay to suggest and defend alternate conditions for ruling out Riker lotteries. All that matters is that condition (I) is not required to rule them out, and that there are conditions which will suffice that do not entail (I).

This argument, then, fails to motivate condition (I) as a constraint on either social aggregation functions or individual aggregation functions.

2.4.3 Argument 3: Functions which violate condition (I) can also violate the majority rule condition

At the core of the eighteenth century debate between Condorcet's camp and Borda's camp was the recognition that when there is a clear pairwise majority winner, the Borda Rule

might fail to select it. This is, it seems, the oldest criticism of the Borda Rule, and among the most common arguments used to motivate condition (I).[31] Let's put the point more precisely. Again, let f be an aggregation function, $\mathcal{P} = \langle R_1, R_2, \ldots, R_n \rangle$, and define $C_y^x = \left|\{i \mid xR_iy\}\right|$. Consider the following condition:

(M) Weak majority rule If $C_y^x > C_x^y$ then $xf(\mathcal{P})y$.

It is easy to see that the Borda Rule violates this condition. Let $S = X = \{x, y, z\}$. Suppose there are five voters, three of whom rank the options $x > y > z$ and two of whom rank the options $y > z > x$. x is plainly preferred to every option by a majority of voters. In particular, $C_y^x > C_x^y$. But the Borda Rule returns the aggregate order $y > x > z$. Thus, the Borda Rule violates (M).

As it happens, neither condition (I) nor condition (M) entails the other. In other words, the Borda Rule does not violate (M) merely because violates (I). We can easily describe an aggregation function that satisfies (M), but violates (I). As above, in §2.5.2, let $\mathbf{B}_x^+$ notate the Borda count for any x. Consider the function f defined such that, for all x, y, if $C_y^x > C_x^y$, then $xf(\mathcal{P})y$, and if $C_y^x = C_X^y$, then $xf(\mathcal{P})y$ if and only if $\mathbf{B}_x^+ > \mathbf{B}_y^+$. This function agrees with majority pairwise preference wherever it arises, and settles any ties that arise by the Borda Rule. It quite obviously satisfies (M), since it was defined to do just that. And it also quite obviously violates (I) for just the same reasons the plain Borda Rule does. In the other direction, every imposed aggregation function trivially satisfies (I), but violates (M).

However, I think there is a less pedantic and more charitable interpretation of this argument. Call any aggregation function which violates condition (I) a *non-(I)* function, and any aggregation function which violates condition (M) a *non-(M)* function. It is often assumed or argued that if any non-(I) function is potentially viable, then the Borda

[31] This criticism of the Borda Rule is actually a great deal older than either Condorcet's or Borda's work. Ramon Llull observed this as early as the thirteenth century. See my notes 15 and 17 above.

Rule and all other members of a closely related family of positional functions—call it the *Borda family*—are viable. Let's take this claim for granted. Then, if we can show that we have reason to reject any function from the Borda family, we can show that there is no viable aggregation function which violates (I). The Borda Rule is obviously a non-(M) function. Let's assume further that the Borda family consists exclusively in non-(M) functions. Given the assumptions made thus far, if we can show that violating (M) is a reason to reject an aggregation function, we will have shown that no viable aggregation function violates (I). Although this reason for (I) is fairly hedged and indirect, it still provides some motivation for the condition.

So should we reject aggregation functions which violate (M)? With respect to social aggregation, there is a strong case that we should. To reiterate an earlier point, in cases of social aggregation, the weak orders to be aggregated are the preferences of moral agents. Moral agents are the subjects of considerations of justice and fairness. Justice and fairness are generally understood to require that without good reason to do otherwise, we treat moral agents as equals. Equal treatment requires giving equal weight to the preferences of every individual. Suppose a majority of individuals prefer x to y, but the aggregate order prefers y to x. Under the banner of equal treatment, there seem to be only two plausible ways to justify this: It could be justified because the preferences of the individuals in the minority are more important—say because these individuals are more reasonable, or will be more deeply affected by the issue under decision, or are of a higher social rank. Or it could be justified because the preferences of the individuals in the minority for y over x are more intense than the preferences of the individuals in the majority for x over y.

The former option is a non-starter, because in the formal framework of aggregation problems, we don't have any information about the relative importance of anyone's preferences. The latter justification is also a non-starter. To be sure, preferences vary in

intensity. But in an aggregation problem, we have access only to each individual's ordinal ranking of options. This isn't sufficient to determine a unique cardinal utility function for each individual, let alone a joint utility scale on which interpersonal comparisons can be made. (More on this below.) So, in the formal framework of aggregation problems, individual preferences have to be treated as though they are discrete, on-off states, each to be counted once as a single unit of preference in the aggregate. One individual's preference for x over y is to be tallied as equal in intensity and importance to any other individual's preference for y over x. The preference of the majority has to prevail in the aggregate, on pain of unequal treatment of individuals when it does not.

What's to be made of the analog of this argument with respect to individual aggregation? We've already noted that an individual's decision criteria aren't moral agents, or agents of any kind, that there are no analogs of justice and fairness from which to argue for their equal treatment, and that there may even be cases in which they should be treated unequally. So the analogous argument simply doesn't get off the ground with respect to individual aggregation. If we don't begin with the defeasible presumption that all decision criteria are to be treated equally, we don't need justification to defeat that presumption. That certain functions which violate (I) sometimes also violate (M) is not, in and of itself, good reason to reject these aggregation functions.

There is, however, the lingering worry of comparisons of intensity of preference that arose as a subsidiary point in the discussion of this argument. I revisit this point in §§4.5 and 5 below.

2.4.4 Argument 4: Condition (I) eliminates the effect of judgments pertaining to infeasible options

One way to express the importance of condition (I) is that it prevents preferences involving infeasible options from affecting the aggregate order. For any aggregation function

that satisfies (I), if S is the set of feasible options, we need only consider individuals' preferences between options in S to determine the aggregate order over S. In other words, for functions satisfying (I), aggregate preferences over feasible options are independent of individuals' preferences over infeasible options. Hence the 'independence' component of the name. For aggregation functions which violate (I), the aggregate order over S depends on preferences over infeasible options.

Discussions of this point sometimes run aground on relatively uninteresting questions about feasibility. To forestall this, I'm going to shift terms. Call preferences involving options in $X \setminus S$ *off-agenda preferences*. Independence of the aggregate order over S from the influence of off-agenda preferences is clearly justified only if we have good reason to consider off-agenda preferences irrelevant. Indeed, the condition might have been been better called "independence from irrelevant preferences." Presumably, this justification of condition (I) is rooted in a plausible general principle of rationality that our decisions—whether individual or collective—should not be influenced by irrelevant information. Suppose that off-agenda preferences are always genuinely irrelevant. Then, for any aggregation function that violates (I), the aggregate order over S depends on irrelevant information. Thus, we have good reason to reject aggregation functions which violate (I).

There is a problem with this version of the argument. Off-agenda preferences don't always appear to be irrelevant. This is not a novel point, or even a contested point. Indeed, virtually every commentator who has written on Arrow's theorem concedes that, in principle, preferences pertaining to infeasible options might sometimes be relevant. This point is made quite clearly in Arrow 1951, Hildreth 1953, Goodman and Markowitz 1952, Sen 1970, Riker 1982, Sen 1982, and P. J. Hammond 1987, just to name a few. Perhaps a more illustrative way to put the point is this: Not all off-agenda options are equal with respect to relevance. Some are clearly more relevant than others. Relevance comes

in degrees.

Consider the following case. Suppose a city council is voting on a new community development project. Proposals for the development projects were submitted by citizens at an open forum meeting, and all proposals were open for discussion. A few of the proposed projects are construction of a new stadium, construction of a new library on the east side of town, construction of a new library on the west side of town, renovation of the historic shopping district, renovation of the waterfront district, expanding the city convention center, and construction of a spaceport. After preliminarily ranking all proposed options, the council determines that the spaceport is infeasible because it's absurd, the stadium is infeasible because it's too expensive, and the west-side library is infeasible because no suitable building site is available. Though the option is now infeasible, off-agenda preferences involving the west-side library seem especially relevant because of its similarity to another option in contention. They are clearly more relevant than off-agenda preferences involving the stadium. And off-agenda preferences involving either the west-side library or the stadium are clearly more relevant than off-agenda preferences involving the ludicrous spaceport.

To see the point somewhat more clearly, let's consider a pared down version of this case. Suppose that the initial set of options includes only the following projects: the east-side library, the west-side library, the convention center expansion, and the shopping district renovation. Let e, w, c, and d stand for each of these options respectively, so that $X = \{c, d, e, w\}$. Suppose that of the seven city councilors, four rank the options $w > c > e > d$, and three rank the options $e > w > d > c$. On this profile, the Condorcet Rule ranks the options $w > c > e > d$, and the Borda Rule ranks the options $w > e > c > d$. Both agree that w is the clear winner. Now suppose that the west-side library is ruled out for the reasons discussed above so that $S = \{c, d, e\}$. The Condorcet Rule returns the order $c > e > d$ whereas the Borda Rule returns the order $e > c > d$. The two aggregation

functions disagree, the former selects c, whereas the latter—sensitive to off-agenda preferences involving w—selects e. It is not obvious that the Borda Rule is in error. After all, everyone preferred some library or other to everything else, and three of four councillors preferred both libraries to anything at all. Though in the end some of these preferences turned out to be off-agenda, it's not obvious that they are irrelevant. Indeed, my intuition is that they are quite relevant. And—no matter that they'd have violated some consistency conditions in their own preferences by doing so—we can easily imagine that the four councillors who preferred the convention center expansion to the east-side library might have changed their tune if they knew there would be no library at all.

Advocates of condition (I) generally seem quite willing to concede all of this, while still arguing that we should reject aggregation functions which violate (I) because they are sensitive to off-agenda preferences. Their reasoning seems to run as follows: The relevance of off-agenda preferences does seem to come in degrees. Some off-agenda preferences are clearly more relevant than others, and some might even be downright worth considering. But there is no acceptable formal or practical rule for sorting off-agenda preferences by degrees of relevance. Formally, preferences involving options in $X \setminus S$ are indistinguishable with respect to their relevance. Practically, different individuals may disagree about which off-agenda preferences are relevant—P. J. Hammond 1987, in particular, worries about this problem—and to suggest that the relevance of off-agenda preferences can be settled by polling the collective is to invite a regress of aggregation problems. Without some way to filter out the downright irrelevant off-agenda preferences, all off-agenda preferences should be treated as irrelevant. Thus, we should reject aggregation functions which violate (I). At this point, we ought to wonder what licenses the conclusion that when we are unable to sort the relevant from the irrelevant off-agenda preferences we should treat them all as irrelevant. Presumably, this step in the reasoning is underpinned by another intuitively plausible general principle of rationality. The

principle is something likes this: It is a greater sin of rationality to allow patently irrelevant information to impact our decisions—whether individual or collective—than it is to ignore potentially relevant information.

This argument is, I think, quite compelling with respect to social aggregation. But there is a ready reply with respect to individual aggregation, which incorporates points from the preceding replies to Arguments 1 and 3 above (§§4.1. and 4.3, respectively.) First, to reiterate a point raised in reply to Argument 3, an individual's decision criteria are not autonomous agents. Second, to reiterate a point raised in reply to Argument 1, in cases of individual aggregation, the aggregator is one and the same as the individual whose preferences are to be aggregated. Because decision criteria are something like value considerations, rather than autonomous or quasi-autonomous agents, they can't disagree over which options are relevant. Indeed, they can't formulate any judgments of relevance at all. So this puts to rest Hammond-style worries about turf-wars over relevance.

More importantly, because the aggregator is aggregating her very own decision criteria, she is perfectly positioned as the arbiter of what is and isn't relevant. In other words, in cases of individual aggregation there is a general rule for sorting relevant off-agenda preferences from irrelevant off-agenda preferences. It is a practical, rather than a formal rule: an off-agenda preference is relevant precisely when and to the degree that the individual decision maker judges it to be. In many cases, it is quite reasonable to assume that all preferences involving all options that originally appeared in X will remain relevant. However, we need not assume that much. When faced with a decision in which $S \neq X$, an individual decision maker might very well determine that only some of her off-agenda preferences are relevant to the decision. Even so, any aggregation function sensitive to any off-agenda preferences will violate (I).

Thus, the argument that we should endorse condition (I) because it eliminates the effect of irrelevant off-agenda preferences fails with respect to individual aggregation. This

argument succeeded with respect to social aggregation, because we lacked a clear way to sort out which off-agenda preferences were relevant. We are not similarly handicapped with respect to individual aggregation.

2.4.5 Argument 5: Condition (I) prohibits interpersonal comparisons of utilities

From the outset, we can ignore versions of this argument that depend on the claim that condition (I) somehow prohibits interpersonal comparisons of utility, full stop.[32] There is an easily ignored fact of welfare economics, to which Scitovsky 1951 draws our attention: Even the most obvious collective welfare-based recommendations of economic policy necessarily involve some sort of interpersonal comparisons of welfare. Such comparisons might be crude, implicit, and qualitative, but they are never absent.[33] Hildreth 1953 makes much the same point, when he writes:

> ... as soon as we say that state x is socially preferred to state y for two states such that some individuals prefer x to y and others prefer y to x, we are thereby saying that the gains to those who prefer x are socially more important than the losses of those who prefer y. This implies that we have some basis for comparing the relevant gains and losses. Such a comparison is fundamentally an interpersonal comparison of utilities (90).

In other words, all aggregation functions—even those which satisfy condition (I)—involve some degree of implicit interpersonal comparisons of utilities.

The trouble is not merely that aggregation functions which violate condition (I) make interpersonal comparisons of utilities. Rather, it is that they mistakenly attempt—as Mackie 2003 so pithily puts it—"to squeeze cardinal blood from the ordinal turnip (145)." This line of reasoning is best illustrated with a specific example of an aggregation function which violates condition (I); as usual, I'll discuss the Borda Rule. Recall that at the

[32] Some commentators like Mackie 2003 and Hildreth 1953 seem to read an argument like this into Arrow 1951. I don't see any convincing evidence that Arrow puts forward anything quite so naïve. Nonetheless, this flatfooted version of the argument is widely echoed elsewhere in the literature on social choice theory.

[33] See Sen 1982, 264–5 for a valuable summary of Scitovsky 1951.

heart of the Borda Rule is what we might call a *positional scoring function* that, for each individual weak order, maps each option to a natural number. Here I've opted to specify the Borda Rule by way of a positional scoring function that maps each option to the number of other options to which it is strictly preferred, though other authors sometimes define the rule differently. These positional scores are then summed across individuals for each option, and the aggregate order is determined on basis of comparisons of these sums. For any two options, whichever has a greater aggregate positional score is strictly preferred in the aggregate.

Thus, the Borda Rule implicitly treats differences between adjacent positional scores as equivalent within individual weak-orders, and across individuals. It will help to precisely cash out these two points. For any option x, and any individual i, if x rises or falls in i's weak order, such that its positional score increases or decreases by exactly 1, this will change x's aggregate positional score by exactly 1, no matter what individual positional score x initially had with respect to i's order. And for any two individuals, i and j, and any option x, if x rises in i's weak order such that its positional score is increased by exactly 1, and falls in j's weak order such that its positional score is decreased by exactly 1, x's aggregate positional score will remain unchanged, no matter what individual positional scores x initially had with respect to i's and j's weak orders.

Proponents of condition (I) argue that there simply isn't any available justification for this implicit interpersonal comparison scheme. Two individuals with sharply distinct utility functions may have precisely the same ordinal preferences. From ordinal preferences alone, we cannot conclude that the utility difference between adjacent positional scores is identical either within or across individuals. To justify this comparison scheme, we'd need cardinal data on individual utility somehow measured on a common scale. The comparison scheme implicit in the Borda Rule isn't adequately grounded by the information available in an aggregation problem. Distinct aggregation functions which violate

condition (I) will all involve their own implicit interpersonal comparison schemes, which will also be too specific to be grounded by the information available in an aggregation problem. So we should accept condition (I) as a constraint on rational aggregation functions. Call this the *utility comparisons argument*.

This argument is at once the hardest to translate from the domain of social aggregation to the domain of individual aggregation, and also the toughest argument to reply to with respect to individual aggregation.

There is no obvious analog of individual utility with respect to individual decision criteria. Individuals are quite reasonably modeled with real-value functions over options (unique up to linear transformation) that communicate the individual's level of welfare if that option is selected. Decision criteria don't obviously have anything like levels of welfare. Thus it's not clear that an individual's decision criteria are similarly reasonably modeled with real-value functions over options (unique up to linear transformation) that communicate the level to which that decision criteria will be satisfied if that option is selected. One way to resist a similar line of argument with respect to individual aggregation would be to argue that there is no analog of utility for decision criteria. This preemptive line of reply isn't especially interesting, and I doubt it would be particularly compelling. While decision criteria don't obviously have levels of welfare, we can quite easily imagine a criterion being satisfied to various degrees. For the preemptive reply to work, we'd need to argue successfully that there there is some important disanalogy between an individual's level or welfare and a decision criterion's degree of satisfaction, such that the former but not the latter can be reasonably modeled with a real value function. I just don't see how any such argument could succeed.

So let's assume that there is some analog of utility with respect to an individual decision criterion. Call it *choiceworthiness*.[34] Choiceworthiness is to a decision criterion as

[34] I borrow this use of the term 'choiceworthiness' from MacAskill 2014.

utility is to an individual. For each decision criterion, there is an associated choiceworthiness function (unique up to linear transformation) which communicates the degree to which that criterion is, in some sense I won't bother specifying, satisfied by that option. Just as we can talk about utility of options in relation to a particular individual, we can talk about choiceworthiness of options in relation to a particular decision criterion.

The analog of the interpersonal comparisons argument then runs as follows: Individual aggregation functions which violate condition (I) implicitly make fairly specific inter-criteria comparisons of choiceworthiness. These comparisons would require robust cardinal data to justify. This data is unavailable in an aggregation problem where we have merely ordinal information. Thus we should reject aggregation functions which violate condition (I). Call this the *choiceworthiness comparisons argument.*

What do we make of these arguments? The utility comparisons argument is perhaps the most convincing motivation for condition (I) as it applies to social aggregation. I know of no adequate reply to it in the literature. Developing ideas from Armstrong 1951, the response in Goodman and Markowitz 1952 seems initially promising, but it fails for reasons discussed in Vickrey 1960 and Sen 1970.

However, I think that a Goodman and Markowitz-style reply succeeds with respect to the choiceworthiness comparisons argument where it failed with respect to the utility comparisons argument, again because of crucial differences between individuals and an individual's decision criteria.

Robbins 1932 offers the earliest thorough discussion of interpersonal comparisons of utility in the contemporary tradition of welfare economics. Sen 1982 notes that Robbins is widely misunderstood to have argued there that such comparisons are always groundless (265).[35] But—as he is at pains to make clear in Robbins 1938—his view of things is far less pessimistic. To be clear, Robbins does argue that what Sen later calls *descriptive* in-

[35]See: Sen's nt. 1, that page.

terpersonal comparisons will always be unjustified. Descriptive comparisons purport to compare actual utilities across individuals. Justifying descriptive comparisons requires utility measurements on some common scale for all individuals to be compared. Crucially, though, Robbins recognizes that we can make interpersonal comparisons of an altogether different type, what Sen later calls *prescriptive* comparisons. Prescriptive comparisons purport to state how we ought to compare utilities across individuals. Unlike descriptive comparisons, prescriptive comparisons are justified by appeal to ethical or political principles.

On this point, clarifying the confusion surrounding his earlier work, Robbins 1938 writes:

> I do not believe, and I never have believed, that in fact men are necessarily equal or should always be judged as such. But I do believe that, in most cases, political calculations which do not treat them *as if* they were equal are morally revolting. (635)

Appeal to ethical and political principles is precisely how we have justified the condition of anonymity which entails the non-dictatorship condition. And it is precisely how we can justify weighting each individual's weak-order equally in our aggregation functions. In the case of social aggregation, we—like Robbins—find the idea of aggregation functions which do not treat individuals as equals revolting to our ethical sensibilities, and generally incompatible with our most cherished ethical and political principles. Yet the individuals whose weak orders we collect in an aggregation problem might experience wildly different levels of utility. One individual might experience more utility if his least preferred option were realized, than another would experience if his most preferred option were realized. Despite this, we accept equal treatment of individuals as a constraint on aggregation functions without the least hesitation. The point here is that we are, in fact, extremely comfortable making prescriptive interpersonal comparisons of utility, no matter how much reluctance we feel making descriptive comparisons. We are

especially comfortable, and take ourselves to be especially well-justified when the prescriptive comparisons we make are egalitarian in nature.

Goodman and Markowitz's general reply to the utility comparisons argument against the Borda Rule is that the same principles that ground equal treatment of individuals also ground equal treatment of differences between adjacent ranks within and across individuals. Mackie 2003 elaborates on the same line of reply. Goodman and Markowitz 1952 and later Mackie 2003 build on ideas in from Armstrong 1951, so I begin with an overview of Armstrong's position on interpersonal comparisons.[36]

Armstrong suggested that we can find a common interpersonal unit of preference in the psychologists' notion of a *just noticeable difference* or, in Armstrong's preferred terms *just perceptible preference*.[37] Armstrong writes:

> . . . some approximation to a preference that is strictly marginal can be discovered by direct introspection; what is just perceptible is revealed to the individual directly and an individual is aware of preferences that are just perceptible as such. Furthermore, non-marginal preferences are clearly distinguishable as varying from weak to strong. If, then, we take the two person group, α, β, with α preferring A to B, and β preferring B to A, the utility-data for solving the group-welfare comparison are, in point of fact, given with the preference-data, since both α and β are aware of the strengths of their preferences. So long as the unit in which α and β measure their preferences is the same, the problem is solved, and the use of the same unit is ensured if both parties use the same number to describe the preference-strength of a 'just perceptible preference.' (268)

Armstrong quite clearly intends to lay the groundwork for a descriptive scheme of interpersonal comparisons.

But the attempt to ground a descriptive scheme of interpersonal comparisons in the notion of just noticeable differences runs into difficulties. Alas, his appeal to the notion of just noticeable differences does not in fact buy him the unit of comparison that he needs.

[36] Mackie, and Goodman and Markowitz, conspicuously fail to cite Armstrong. Mackie does cite Goodman and Markowitz, though rather unhelpfully. But, despite quite obviously lifting directly from Armstrong (see especially p. 259 of Goodman and Markowitz, as compared to pp.267–269 of Armstrong), Goodman and Markowitz make no mention of Armstrong.

[37] See: Armstrong 1951, 268.

Vickrey 1960 points out that Armstrong fails to take note of the fact that the degree to which individuals can discriminate options varies widely on basis of many, many factors (519-20). Further, even given two equally discriminating individuals, we have no reason to believe that they attach the same degree of felt subjective welfare to their just noticeable differences of satisfaction. Consequently, we have no reason to believe the common scale yielded by just noticeable differences of satisfaction is actually a scale of utility. The comparisons it yields are not comparisons of utility, but comparisons of measurements in just noticeable differences. This is not at all what we were after. We can illustrate the difficulty with an example. Suppose everyone measured distance by the length of his foot. One way we could compare measurements across individuals would be to treat all individual measurements as alike in scale. One individual's measurement of one foot is taken to be comparable to anyone else's measurement of one foot. This is hardly a way to generate joint scale. What results are at best pseudo-comparisons, too subjective to be put to any use, least of all for navigating the world

Goodman and Markowitz 1952, and later Mackie 2003 riff on Armstrong 1951 by building atop the notion of just noticeable differences a notion of an individual's *levels of discretion*. When an individual judges at least a just noticeable difference of satisfaction from one option to next, the options sit at different levels of discretion. We ought, according to these authors, treat the move from one level of discretion to another as comparable across individuals.[38] We can then treat each individual's weak-order as ranking her options by levels of discretion. We can then interpret an option's individual Borda Rank as indicating the level of discretion at which it sits for that individual. Thus, the comparisons implicit in the Borda Rule can be understood as based on levels of discretion, and thus these authors allege, the Borda Rule is not objectionable.

Setting aside the practical problems with this approach involving accurately assessing

[38]The question of precisely which principles of ethics or politics motivate this conclusion is left to the reader's imagination.

individuals' levels of discretion notwithstanding, it's clearly not the case that we should treat levels of discretion as having the same ethical significance across individuals. Sen 1970 builds on Vickrey's earlier arguments against Armstrong, Sen 1970 argues for this point quite convincingly (93-4). We might have two individuals, one of whom has very fine grained preferences, and thereby many levels of discretion, and another of whom has only coarse-grained preferences with just two levels of discretion. Clearly, the ethical significance of shifting one level of discretion is not the same for these individuals, what for the former individual may be an unnoticed triviality may have the utmost impact for the latter.

Its fair to conclude that prospects for this line of reply to the utility comparisons argument are bleak.

The analogous reply to the choiceworthiness comparisons argument does not meet with the same difficulties. In the case of individual aggregation, there is only one individual whose distinct decision criteria are to be aggregated. Thus there is only one individual whose just noticeable differences and levels of discretion are to be considered. To be clear, these may vary somewhat from one criterion to another. But, recall that in §2.4 we noted that in any case where the individual decision maker has a fully worked-out priority or weighting scheme for her decision criteria, she has thereby obviated the sort of conflict between criteria that was of interest in the first place. We can assume, then, that we're dealing with an individual who genuinely feels the pull of several different decision criteria. Moreover, though she may not feel the pull of each criterion in the same way, or even with quite the same intensity, she cannot measure or rank their deliberative significance in any straightforward way. This is, in a sense, just what it means for her to be a decision maker with multiple conflicting criteria.

But if the degree of relevance of these criteria is not clear to her, beyond the fact that each criterion is relevant, then she clearly ought to treat changes in levels of discretion

equally across criteria. That is, she ought to treat differences between adjacent ranks as of comparable deliberative significance. Here, unlike in the case of social aggregation, when she treats differences between adjacent ranks as comparable there is no underlying fact of the matter she can run afoul of. There are no two individuals with different subjective welfare experiences, whom we can be problematically misrepresenting when we treat them as comparable. The decision maker's criteria are relevant only because, and only to the degree that, they are relevant to her. If she cannot sort them by relevance, there is no further fact of the matter to uncover. Thus, she should treat differences between adjacent ranks as comparable across her decision criteria, because its the only principled attitude she can adopt toward them; she should treat them as comparable precisely because she has no reason treat them as incomparable. Thus, the line of reply advanced by Goodman and Markowitz, and Mackie, succeeds with respect to the choiceworthiness comparisons argument.

2.5 Rational violations of condition (I) in individual aggregation

Condition (I) is not, as we saw in §3, easily understood, and when it's made clear, it lacks the intuitive appeal so many commentators have taken as sufficient motivation for the condition. Beyond that, the usual arguments intended to motivate condition (I) as a constraint on rational social aggregation functions simply do not translate the the case of individual aggregation. It seems that we are left with no reason at all to accept condition (I) as a constraint on rational individual aggregation functions. Even so, I suspect that partisans of condition (I) might take a little more convincing. I turn now to arguing directly that certain individual aggregation functions which violate condition (I) are rational.

The point I argue for here is that it is rational for individual decision makers to be sensitive to off-agenda preferences in precisely the way that they would be if they were to

violate condition (I) by aggregating with a function like the Borda Rule. In other words, it is not merely that condition (I) fails to be a constraint on rational individual aggregation functions, but that some individual aggregation functions are rational precisely because they violate condition (I). Recall that off-agenda preferences are preferences involving items from the original set of options, X, which for some reason or other, are not included in the agenda, S.

My argument here is quite brief. First, note that in both kinds of aggregation, the inputs to an aggregation function are of a fundamentally different kind than the output, even if both are modeled with weak orders. The point is more obvious in social aggregation: individual orders and the aggregate social order are not the same kind of thing. The former guide individual choice behavior; the latter guides social choice behavior. But the point is felt more acutely with respect to individual aggregation. As we've noted, individual decision criteria are not themselves decision makers, and so their associated weak orders do not directly guide any kind of choice behavior. For this reason, we should not be troubled if what grounds the ranking determined by a decision criterion differs somehow from what grounds overall preferences.

One respect in which I think they differ is that ranking options according to a specific decision criterion sometimes integrally involves consideration of imagined options—in particular imagined best or worse cases—in a way that formation of overall preferences might not. When I suggest that the ranking integrally involves consideration of the imagined option, I mean that each option's place in the ranking is determined by how it compares to the imagined option(s). Suppose, for example, that Ashima is choosing between vacations, and among her decision criteria are cost, distance from home, and excitement. One way for her to rank the available options according to excitement is to imagine an ideally exciting vacation and then compare available options to it. Available options will rank higher to the degree that they resemble the ideally exciting vacation. Fairly clearly,

then, with respect to this particular decision criterion, in her overall decision it's far more relevant how far each available option is from the ideal, than how it compares pairwise to other available options. Whether this is the only way for her to generate this criteria-specific order is irrelevant. All that matters is that we can make sense of her ordering options according to a specific criterion in this way—that is, by comparing them individual to an ideal option, rather than pairwise to one another. If she does this, there's a sense in which the comparisons to the ideal option are integral to, or constitutive of the weak-order generated by the decision criterion. To ignore these options is to undermine the weak-order determined by the criterion.

There is an obvious objection at this point: If the weak-orders associated with particular decision criteria can be determined in this fashion—by comparison, one at a time, to an imagined ideal rather than by pairwise comparison with one another—then this completely undermines the binary account of choice that is standard in philosophy. This objection would be spot-on, if I had suggested that overall preferences might work like this. But I have not suggested that at all. I have merely suggested that rankings of options according to specific criteria might work thus. As I've just pointed out, even if we model them similarly, we should not confuse the two. An individual's overall preferences directly guides her choice behavior. Her individual decision criteria only guide her choice behavior insofar as they are somehow aggregated into a single overall order. Thus, this objection is misplaced.

Returning to the main thread of the argument, wherever the ranking associated with a criterion depends on such imagined options, these options are, necessarily infeasible, in the sense that they're not available to choose. After all, they're imaginary. Therefore, preferences involving these options are necessarily off-agenda preferences. Any aggregation function that satisfies condition (I) necessarily prevents these preferences from having any impact on the aggregate order. But if the order determined by that decision

criterion is grounded in comparisons to an imagined option, ignoring off-agenda preferences involving that option renders the ranking associated with that criterion more-or-less meaningless. If the decision maker aggregates by a function that satisfies condition (I), she ignores what grounds the weak order associated with the criterion in question. Perhaps I stretch the use of the word 'rational' when I suggest that it would be irrational in these cases to use aggregation functions which satisfy condition (I). But if a means of decision making erodes the very basis for rational choice, it seems that means should be called irrational.

In such cases, only aggregation functions which violate condition (I) are sensitive to off-agenda preferences in the right way to respect the role they play in grounding the weak orders associated with certain decision criteria. And it is precisely because they violate condition (I) that they are appropriately sensitive to these off-agenda preferences. Thus, far from being a constraint on rational individual aggregation functions, satisfying condition (I) is sometimes directly opposed to rational aggregation.

2.6 Prospects for modeling multiple objective decisions as aggregation problems

Above I argued that we should not take Arrow's condition (I) as a constraint on rational individual aggregation functions. As I noted above (and is reiterated time and again in the literature on social choice theory), we're supposed to be struck dumb by the sheer intuitive appeal of this condition. After all, Arrow himself takes the "essential argument" in favor of the condition to be its "direct appeal to intuition" and other partisans of the condition almost unanimously echo this sentiment.[39] All other substantive motivations for the condition are supposed to be subsidiary to its intuitive appeal. But in §3.2, I show that the condition clearly lacks the intuitive glow that was supposed to have drawn us in like moths. Abstract away from concrete examples, and the condition—like so many

[39] See: Arrow 1983, 51.

other abstruse formal principles—is neither immediately attractive nor immediately repulsive to the intuition. Then, in §4, I argued that the standard substantive arguments for condition (I) advanced in the social case fail to extend to the individual case.

At this point, I hope I've shown that we have no reason whatsoever to take condition (I) as a constraint on rational individual aggregation functions.

Then, in §5, I argued that individual decision makers should be sensitive to off-agenda preferences in precisely the way they would be if they were to violate condition (I). My argument here is best understood as a balance-of-reasons argument. Balance-of-reasons arguments sometimes leave us wanting something more, in particular when the bookkeeping of opposing reasons is a bit murky. But in this case, there is nothing more to want. The ledger is quite clear: There is no reason to take condition (I) as a constraint on individual aggregation functions, and at least one good reason to reject it.

This means that, though Arrrow's Theorem might make us pessimists about rationally resolving social aggregation problems, we can remain optimists about rationally resolving individual aggregation problems. So long as we don't think individual aggregation functions are subject to additional constraints of rationality sufficient to generate an Arrow-style impossibility result, then we should expect to find individual aggregation functions that satisfy all of the constraints rationality imposes.

This re-opens the door to a modeling approach for what are—after Keeney and Raiffa 1976—standardly called *multiple objective decisions*; we can model these decisions as individual aggregation problems. Multiple objective decisions are the sort of decisions we've had in mind, at least in the background, from the beginning of this essay. An individual faces a multiple objective decision when she evaluates her options according to multiple, possibly conflicting criteria, each of which she sees as motivationally relevant with respect to her choice. To model such a decision as an aggregation problem, we model the decision-maker's decision criteria as a tuple of weak orders over her options, from which

her aggregate—that is overall preference order—is to be determined by application of an aggregation function. Until now philosophers working on the problem of multiple objective decisions—for example, Steedman and Krause 1985, Kavka 1991, and Ellis 2006—have taken this approach to be a non-starter because of Arrow's Theorem. Part of the pay-off of this essay is to give this modeling approach a second chance.

This modeling approach faces pressing questions, both formal and foundational in nature. At the formal level I have to confront the question of whether there are additional constraints on rational individual aggregation functions, and if so, whether they are sufficient to generate an impossibility result. Then, given some set of constraints, I'll need to know whether they determine one unique aggregation function, or many distinct aggregation functions. If a plurality of aggregation functions satisfy the constraints, I'll have to face the further question of which of these is the correct aggregation function. At a more foundational level, I'll have to confront questions about how well this model actually captures the essential features of a multiple objective decision. For starters, I'll need to give a more precise account of just what decision criteria are, and whether they can be appropriately modeled by weak orders. At perhaps a deeper level I'll need to say more about why we should think the output of models like this is in any sense correct, or normatively binding.

I defer the task of developing this modeling approach in more detail to future work. For now, at least, it remains an enticing possibility.

Bibliography

Ainslie, George (1985). "Beyond microeconomics. Conflict among interests in a mulitple self as a determinant of value." In: *The Multiple Self*. Ed. by Jon Elster. Cambridge University Press (cit. on p. 24).

— (2001). *Breakdown of the Will*. Cambridge University Press (cit. on p. 24).

Anderson, Richard M. and Robert T. Clemen (2013). "Toward an Improved Methodology to Construct and Reconcile Decision Analytic Preference Judgments". In: *Decision Analysis* 10.2, pp. 121–134 (cit. on p. 61).

Armstrong, W. E. (1951). "Utility and the Theory of Welfare". In: *Oxford Economic Papers* 3, pp. 259–71 (cit. on pp. 110, 112, 113).

Arrow, Kenneth J. (1950). "A Difficulty in the Concept of Social Welfare". In: *Journal of Political Economy* 58 (cit. on pp. 77, 81, 82, 88, 89, 91).

— (1951). *Social Choice and Individual Values*. 1st ed. Cowles Commission for Research in Economics: Monograph 12. John Wiley and Sons (cit. on pp. 82, 88, 89, 91, 103, 107).

— (1963). *Social Choice and Individual Values*. 1st ed. John Wiley and Sons (cit. on pp. 82, 88, 89, 91).

— (1983). "The Principle of Rationality in Collective Decisions". In: *Social Choice and Justice*. Vol. 1. Collected Papers of Kenneth J. Arrow. Belknap Press, pp. 45–58 (cit. on pp. 88, 89, 118).

— (1987). "Oral History I: An Interview". In: *Arrow and the Ascent of Modern Economic Theory*. Ed. by George R. Feiwel. New York University Press (cit. on p. 89).

Baier, Kurt (1958). *The Moral Point of View*. Cornell University Press (cit. on p. 25).

Barry, Brian and Russell Hardin (1982). *Rational Man and Irrational Society*. Sage Publications (cit. on p. 89).

Baumann, Peter and Monika Betzler (2004). *Practical Conflicts: New Philosophical Essays*. Cambridge (cit. on p. 25).

Becker, Gordon M. and Charles G. McClintock (1967). "Value: Behavioral Decision Theory". In: *Annual Review of Psychology* 18, pp. 239–286 (cit. on p. 20).

Bettman, James R., Mary Frances Luce, and John W. Payne (2008). "Preference construction and preference stability: Putting the pillow to rest". In: *Journal of Consumer Psychology* 18, pp. 170–174 (cit. on p. 60).

Bordes, Georges and Nicolaus Tideman (1991). "Independence of Irrelevant Alternatives in the Theory of Voting". In: *Theory and Decision* 30.2, pp. 163–186 (cit. on pp. 89, 91, 94).

Brafman, Ronen I. and Moshe Tennenholtz (1996). "On the foundations of qualitative decision theory". In: *Proceedings of the 13th national conference on artificial intelligence*. American Academy of Artificial Intelligence (cit. on p. 8).

Brink, David O. (1996). "Moral Conflict and Its Structure". In: *Moral Dilemmas and Moral Theory*. Ed. by H. E. Mason. Oxford University Press (cit. on pp. 26, 28).

Butler, Joseph (2006). *The Works of Bishop Butler*. Ed. by David E. White. Boydell & Brewer (cit. on p. 77).

Chang, Ruth (1997a). *Incommensurability, Incomparability and Practical Reason*. Harvard University Press (cit. on p. 47).

— (1997b). "Introduction". In: *Incommensurability, Incomparability and Practical Reason*. Ed. by Ruth Chang. Harvard University Press, pp. 1–34 (cit. on p. 47).

Cherniak, Christopher (1986). *Minimal Rationality*. MIT Press (cit. on p. 24).

Clemen, Robert T. (1996). *Making Hard Decisions: An Introduction to Decision Analysis*. 2nd ed. Duxbury Press (cit. on p. 20).

Conlisk, John (1996). "Why bounded rationality?" In: *Journal of Economic Literature* 34, pp. 669–700 (cit. on p. 29).

Dancy, Jonathan (2004). *Ethics Without Principles*. Oxford University Press (cit. on p. 56).

Dasgupta, Partha and Eric Maskin (2014). "On the Robustness of Majority Rule". In: *The Arrow Impossiblity Theorem*. Ed. by Eric Maskin and Amartya K. Sen. Kenneth J. Arrow Lecture Series. Columbia University Press. Chap. 2, pp. 29–42 (cit. on pp. 88, 89, 96).

Davidson, Donald (2001a). *Essays on Actions and Events: Philosophical Essays Volume 1*. Clarendon Press (cit. on pp. 25, 26, 28).

— (2001b). "How is Weakness of the Will Possible?" In: *Essays on Actions and Events: Philosophical Essays Volume 1*. Clarendon Press (cit. on p. 25).

Davidson, Donald, Patrick Suppes, and Sidney Siegel (1957). *Decision Making: An Experimental Approach*. Stanford University Press (cit. on pp. 6, 8).

Debreu, G. (1960). "Topological methods in cardinal utility theory". In: *Mathematical Methods in the Social Sciences, 1959*. Ed. by K. J. Arrow, S. Karlin, and P. Suppes. Stanford University Press (cit. on p. 70).

Denicolò, Vincenzo (2000). "Independence of Irrelevant Alternatives and Consistency of Choice". In: *Economic Theory* 15.1, (cit. on pp. 89, 90).

Doyle, John and Richmond H. Thomason (1999). "Background to qualitative decision theory". In: *AI Magazine* 20.2, pp. 55–68 (cit. on p. 9).

Duggan, John and Thomas Schwartz (2000). "Strategic manipulability without resoluteness or shared beliefs: Gibbard-Satterthwaite generalized". In: *Social Choice and Welfare* 17.1, pp. 85–93 (cit. on p. 97).

Dummett, Michael (1984). *Voting Procedures*. Oxford (cit. on p. 89).

Ellis, Stephen (2006). "Multiple Objectives: A Neglected Problem in the Theory of Human Action". In: *Synthese* 153.2, pp. 313–338 (cit. on pp. 18, 28, 78, 95, 120).

Elster, Jon (1985a). "Introduction". In: *The Multiple Self*. Ed. by Jon Elster. Cambridge University Press (cit. on p. 24).

— (1985b). *The Multiple Self*. Cambridge University Press (cit. on p. 19).

Feiwel, George R., ed. (1987). *Arrow and the Foundations of the Theory of Economic Policy*. New York University Press.

Fine, B. and K. Fine (1974). "Social Choice and Individual Ranking I". In: *The Review of Economic Studies* 41.3, pp. 303–322 (cit. on p. 87).

Fischer, Gregory W. (1973). "Experimental Applications of Multi-attribute Utility Models". In: *Utility, Probability, and Human Decision Making*. Ed. by Dirk Wendt and Charles Vlek. Vol. 11. Theory and Decision Library. D. Reidel Publishing Company. Chap. 1, pp. 7–46 (cit. on pp. 20, 30, 31).

Fishburn, Peter C. (1970). "Arrow's Impossibility Theorem: Concise Proof and Infinite Voters". In: *Journal of Economic Theory* (cit. on pp. 70, 82).

Franklin, Benjamin (1772). "Letter to Joseph Priestly, September 19, 1772". In: *Benjamin Franlin: Representative Selections*. Ed. by Frank Mott and Chester Jorgenson. American Book Company, pp. 348–349 (cit. on p. 27).

Gaertner, Wulf (2009). *A Primer in Social Choice Theory: Revised Edition*. LSE Perspectives in Economic Analysis. Oxford University Press (cit. on pp. 79, 81, 82, 86).

Gärdenfors, Peter (1973). "Positionalist voting functions". In: *Theory and Decision* 4.1, pp. 1–24 (cit. on pp. 80, 87).

Gibbard, Allan (1973). "Manipulations of Voting Schemes: A General Result". In: *Econometrica* 41 (cit. on p. 97).

Gigerenzer, Gerd and Reinhard Selten, eds. (2002). *Bounded Rationality: the Adaptive Toolbox*. MIT Press (cit. on pp. 24, 29, 30).

Goodman, Leo A. and Harry Markowitz (1952). "Social Welfare Functions Based on Individual Rankings". In: *American Journal of Sociology* 58.3, pp. 257–262 (cit. on pp. 89, 103, 110, 112, 113).

Gorman, W. M. (1968a). "Conditions for Additive Separability". In: *Econometrica* 36.3, pp. 605–609 (cit. on p. 71).

— (1968b). "The Structure of Utility Functions". In: *The Review of Economic Studies* 35.4, pp. 367–390 (cit. on p. 71).

Gowans, Christopher W. (1987). *Moral Dilemmas*. Oxford University Press (cit. on p. 25).

Hammond, John S., Ralph L. Keeney, and Howard Raiffa (1999). *Smart Choices*. Broadway Books (cit. on pp. 20, 27, 33).

Hammond, Peter J. (1987). "Social Choice: the Science of the Impossible?" In: *Arrow and the Foundations of the Theory of Economic Policy*. Ed. by George R. Feiwel. New York University Press, pp. 116–134 (cit. on pp. 87, 96, 103, 105).

— (1988). "Consequentialist foundations for expected utility". In: *Theory and Decision* 25.1, pp. 25–78 (cit. on pp. 15, 76).

Hansson, Bengt (1973). "The independence condition in the theory of social choice". In: *Theory and Decision* 4.1, pp. 25–49 (cit. on pp. 81, 82, 88, 89).

Hare, Richard M. (1952). *The Language of Morals*. 14. Oxford Clarendon Press (cit. on p. 25).

Hildreth, C. (1953). "Alternative conditions for social orderings". In: *Econometrica* 21, pp. 81–94 (cit. on pp. 103, 107).

Horty, John F. (2012). *Reasons as Defaults*. Oxford University Press (cit. on pp. 27, 28, 56).

Howard, Ronald A. (1966). "Decision Analysis: Applied Decision Theory". In: *Proceedings of the Fourth International Conference on Operations Research* (cit. on p. 19).

Jeffrey, Richard C. (1965). *The Logic of Decision*. University of Chicago Press (cit. on pp. 9, 16).

Johansen, Leif (1977). *Lectures on Macroeconomic Planning: Part 1*. North-Holland (cit. on p. 29).

Johnson, Eric J., Mary Steffel, and Daniel G. Goldstein (2005a). "Making Better Decisions: From Measuring to Constructing Preferences". In: *Health Psychology* 24.4 (Suppl.) S17–S22 (cit. on p. 60).

Kaplow, Louis and Steven Shavell (1999). "The Conflict between Notions of Fairness and the Pareto Principle". English. In: *American Law and Economics Review* 1.1-2, pp. 63–77 (cit. on p. 83).

— (2000). "Notions of Fairness versus the Pareto Principle: On the Role of Logical Consistency". In: *The Yale Law Journal* 110.2, pp. 237–249 (cit. on p. 83).

— (2003). "Fairness versus Welfare: Notes on the Pareto Principle, Preferences, and Distributive Justice". In: *The Journal of Legal Studies* 32.1 (cit. on p. 83).

Kavka, Gregory S. (1991). "Is Individual Choice Less Problematic Than Collective Choice?" In: *Economics and Philosophy* 7.02, pp. 143– (cit. on pp. 19, 78, 95, 120).

Keeney, Ralph L. (1992). *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press (cit. on pp. 7, 33–36, 42, 65).

Keeney, Ralph L. (2006). "Using Preferences for Multi-Attributed Alternatives". In: *Journal of Mult-Criteria Decision Analysis* 14.169–174 (cit. on pp. 19, 31, 56).

Keeney, Ralph L. and Howard Raiffa (1976). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press (cit. on pp. 7, 16, 20, 31, 36, 41, 42, 61, 65, 68, 71, 119).

Kelly, J.S. (1987). *Social Choice Theory: An Introduction*. Springer-Verlag (cit. on pp. 81, 90).

Kemp, Murray and Yew-Kwang Ng (1987). "Arrow's Independence Condition and the Bergson-Samuelson Tradition". In: *Arrow and the Foundations of the Theory of Economic Policy*. Ed. by George R. Feiwel. New York University Press, pp. 223–242 (cit. on pp. 89, 91).

Klir, George J. and Tina A. Folger (1988). *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall (cit. on p. 16).

Krantz, D. H. et al. (1971). *Foundations of Measurement*. Vol. 1. Academic Press (cit. on p. 70).

Leontief, Wassily (1947a). "A Note on the Interrelation of Subsets of Independent Variables of a Continuous Funciton with Continuous First Derivatives". In: *Bulletin of the American Mathematical Society* 53, pp. 343–350 (cit. on p. 71).

— (1947b). "Introduction to a Theory of the Internal Structure of Functional Relationships". In: *Econometrica* 15.4, pp. 361–374 (cit. on p. 71).

Lichtenstein, Sarah and Paul Slovic (2006). "The Construction of Preference: An Overview". In: *The Construction of Preference*. Ed. by Sarah Lichtenstein and Paul Slovic. Cambridge University Press (cit. on p. 60).

Lin, Hanti (2013). "Foundations of Everyday Practical Reasoning". In: *Journal of Philosophical Logic* 42.6, pp. 831–862 (cit. on pp. 9, 10).

— (2014). "On the Regress Problem of Deciding How to Decide". In: *Synthese* 191.4, pp. 661–670 (cit. on pp. 5, 6, 9, 29).

Luce, R. Duncan (1956). "Semiorders and a Theory of Utility Discrimination". English. In: *Econometrica* 24.2, ISSN: 00129682 (cit. on p. 89).

Luce, R. Duncan and Howard Raiffa (1957). *Games and Decisions*. John Wiley and Sons (cit. on p. 89).

— (1985). *Games and Decisions*. Dover (cit. on p. 88).

Luce, R. Duncan and John W. Tukey (1964). "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement". In: *Journal of Mathematical Psychology* 1, pp. 1–27 (cit. on p. 70).

MacAskill, William (2014). "Normative Uncertainty". Thesis for the degree of Doctor of Philosophy (D. Phil.) St. Anne's College, Oxford (cit. on p. 109).

Mackie, Gerry (2003). *Defending Democracy*. Contemporary Political Theory. Cambridge University Press (cit. on pp. 88, 89, 94, 96, 107, 112, 113).

Manheim, Marvin L. and Fred L. Hall (1968). *Abstract representation of goals*. Tech. rep. MIT, Department of Civil Engineering (cit. on p. 31).

Marcus, Ruth Barcan (1980). "Moral Dilemmas and Consistency". In: *Journal of Philosophy* 77.3, pp. 121–136 (cit. on pp. 26, 28).

Maskin, Eric (2014). "The Arrow Impossiblity Theorem: Where Do We Go From Here?" In: *The Arrow Impossiblity Theorem*. Ed. by Eric Maskin and Amartya K. Sen. Kenneth J. Arrow Lecture Series. Columbia University Press. Chap. 3, pp. 43–55 (cit. on pp. 83, 89).

Maskin, Eric and Amartya K. Sen, eds. (2014). *The Arrow Impossiblity Theorem*. Kenneth J. Arrow Lecture Series. Columbia University Press (cit. on p. 83).

Mason, H. E. (1996). *Moral Dilemmas and Moral Theory*. Oxford University Press (cit. on p. 25).

May, Kenneth O. (1954). "Intransitivity, Utility, and the Aggregation of Preference Patterns". In: *Econometrica* 22.1, (cit. on p. 77).

McLean, Iain (1995). "Independence of irrelevant alternatives before Arrow". In: *Mathematical Social Sciences* 30.2, pp. 107–126 (cit. on p. 89).

Miller, James R. (1966a). "The assessment of worth: a systematic procedure and its experimental validation". PhD thesis. MIT (cit. on p. 31).

Minsky, Marvin L. (2006). *The Emotion Machine*. Simon & Schuster (cit. on p. 24).

Mongin, Philippe (2005). "Spurious unanimity and the Pareto principle". In: *CPNSS Working Papers*. Vol. 1. 5. Centre for Philosophy of Natural, Social Science, London School of Economics, and Political Science (cit. on p. 83).

Mosteller, Frederick and Philip Nogee (1951). "An Experimental Measurement of Utility". In: *Journal of Political Economy* 59.5, pp. 371–404 (cit. on pp. 8, 15, 60).

Nash, John (1950). "The Bargaining Problem". In: *Econometrica* 18.2, pp. 155–162 (cit. on p. 89).

Nowlis, Stephen Michael and Itamar Simonson (1997). "Measuring constructed preferences: towards a building code". In: *Journal of Marketing Research* 34, pp. 205–218 (cit. on p. 60).

Nussbaum, Martha (1985). "Aeschylus and Practical Conflict". In: *Ethics* 95.2, pp. 233–267 (cit. on p. 25).

Parnell, Gregory S. et al. (2013). *Handbook of Decision Analysis*. Wiley Handbooks in Operations Research and Management Science. Wiley (cit. on p. 20).

Pattanaik, Prasanta K. (2002). "Positional rules of collective decision-making". In: *Handbook of Social Choice and Welfare*. Ed. by Kenneth J. Arrow, Amartya K. Sen, and Kotaro

Suzumura. Vol. 1. Handbook of Social Choice and Welfare. Elsevier, pp. 361–394 (cit. on p. 87).

Plato (1997). *Plato: Complete Works*. Ed. by John M. Cooper and D. S. Hutchinson. Hackett Publishing Company (cit. on p. 25).

Pratt, John W., Howard Raiffa, and Robert O. Schlaifer (1964). "The foundations of decision under uncertainty: an elementary exposition". In: *Journal of the American Statistical Association* 59, pp. 353–375 (cit. on p. 19).

Pruzan, Peter Mark and J. T. Ross Jackson (1963). "On the Development of Utility Spaces for Multi-Goal Systems". In: *Saertryk af Erhevervsøkonomisk Tidsskrift* 4, pp. 257–274 (cit. on p. 70).

Radner, R. and J. Marschak (1954). "Note on Some Proposed Decision Criteria". In: *Decision Process*. Ed. by R. M. Thrall, C. H. Coombs, and R. L. Davies. John Wiley (cit. on p. 89).

Raiffa, Howard (1968). *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Ed. by Frederick Mosteller. Behavioral Science: Quantitative Methods. Addison-Wesley (cit. on pp. 6, 16, 19, 20, 75).

— (1969). *Preferences for Multi-Attributed Alternatives*. Tech. rep. RM-5868-DOT/RC. Santa Monica, CA: The RAND Corporation (cit. on pp. 7, 19, 24, 31, 41, 42, 65).

— (2002). "Decision Analysis: A Personal Account of How It Got Started and Evolved". In: *Operations Research* 50.1, pp. 179–185 (cit. on p. 19).

Raiffa, Howard and Robert O. Schlaifer (2001). *Applied Statistical Decision Theory*. Tech. rep. Division of Research, Harvard Business School (cit. on p. 19).

Ray, Paramesh (1973). "Independence of Irrelevant Alternatives". In: *Econometrica* 41.5, (cit. on pp. 89, 92).

Resnik, Michael D. (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press (cit. on pp. 9, 16, 29, 54).

Riker, William H. (1961). "Voting and the Summation of Preferences: An Interpretive Bibliographic Review of Selected Developments during the Last Decade". In: *The American Political Science Review* 55.4, pp. 900–911 (cit. on p. 96).

— (1982). *Liberalism against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. Waveland Press (cit. on pp. 96, 98, 103).

Robbins, Lionel (1932). *An Essay on the the Nature and Significance of Economic Science*. Macmillan and Company (cit. on p. 110).

— (1938). "Interpersonal Comparisons of Utility: A Comment". In: *The Economic Journal* 48.192, pp. 635–641 (cit. on pp. 110, 111).

Ross, William D. (2002). *The Right and the Good*. Clarendon Press (cit. on p. 25).

Russell, Stuart J. and Eric H. Wefald (1991). *Do the Right Thing*. MIT Press (cit. on p. 24).

Samuelson, Paul (1967). "Arrow's Mathematical Politics". In: *Human Values and Economic Policy: A Symposium*. New York University Press (cit. on p. 89).

Sanbonmatsu, David M. and Russell H. Fazio (1990). "The Role of Atitudes in Memory-Based Decision Making". In: *Journal of Personality and Social Psychology* 59.4, pp. 614–622 (cit. on p. 61).

Sartre, Jean-Paul (1946). *L'existentialisme est un humanisme*. Les Editions Nagel (cit. on p. 25).

Satterthwaite, Mark A. (1975). "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions". In: *Journal of Economic Theory* 10.2, pp. 187–217 (cit. on p. 97).

Savage, Leonard J. (1954). *The Foundations of Statistics*. 1st ed. John Wiley and Sons (cit. on pp. 8, 9, 13, 15, 16, 54, 76).

Schlaifer, Robert O. (1959). *Probabilty and Statistics for Business Decisions*. McGraw-Hill (cit. on p. 19).

Schroeder, Mark (2007). *Slaves of the Passions*. Oxford University Press (cit. on pp. 22, 56).

Schwarz, Norbert (2007). "Attitude construction: evaluation in context". In: *Social Cognition* 25, pp. 638–656 (cit. on p. 60).

Scitovsky, Tibor (1951). "The State of Welfare Economics". In: *The American Economic Review* 41.3, pp. 303–315 (cit. on p. 107).

Sen, Amartya K. (1970). *Collective Choice and Social Welfare*. Holden Day (cit. on pp. 15, 76, 81, 89, 92, 103, 110, 114).

— (1982). *Choice, Welfare and Measurement*. Basil Blackwell (cit. on pp. 81, 89, 92, 103, 107, 110).

— (1993). "Internal Consistency of Choice". In: *Econometrica* 61.3, (cit. on pp. 15, 76).

— (2014a). "Arrow and the Impossibililty Theorem". In: *The Arrow Impossiblity Theorem*. Ed. by Eric Maskin and Amartya K. Sen. Kenneth J. Arrow Lecture Series. Columbia University Press. Chap. 2, pp. 29–42 (cit. on pp. 81, 82).

— (2014b). "The Informational Basis of Social Choice". In: *The Arrow Impossiblity Theorem*. Ed. by Eric Maskin and Amartya K. Sen. Kenneth J. Arrow Lecture Series. Columbia University Press. Chap. 2, pp. 29–42 (cit. on p. 86).

Simon, Herbert (1982). *Models of Bounded Rationality*. MIT Press (cit. on pp. 24, 30).

Simonson, Itamar (1989). "Choice based on reasons: the case of attraction and compromise effects". In: *Journal of Consumer Research* 16, pp. 158–174 (cit. on p. 60).

— (1990). "The effect of purchase quantity and timing on variety-seeking behavior". In: *Journal of Marketing Research* 27.150–162 (cit. on p. 60).

Slovic, Paul (1995). "The Construction of Preference". In: *American Psychologist* 50.5, pp. 364–371 (cit. on p. 60).

Smith, Holly (1991). "Deciding how to decide: Is there a regress problem?" In: *Foundations of Decision Theory*. Ed. by M. Bacharach and S. Hurley. Basil Blackwell (cit. on p. 29).

Steedman, Ian and Ulrich Krause (1985). "Goethe's Faust, Arrow's Possibility Theorem and the individual decision-taker". In: *The Multiple Self*. Ed. by Jon Elster. Cambridge University Press (cit. on pp. 19, 77, 78, 95, 120).

Stevens, Stanley S. (1946). "On the Theory of Scales of Measurement". In: *Science* 103.2684, pp. 677–680 (cit. on pp. 72, 76).

Thomason, Richmond H. (2009). "Logic and Artificial Intelligence". In: *Stanford Encyclopedia of Philosophy* (cit. on p. 9).

— (Forthcoming). "The Formalization of Practical Reasoning: Problems and Prospects" (cit. on p. 9).

Triantaphyllou, Evangelos (2000). *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer (cit. on pp. 10, 24).

Tzeng, Gwo-Hshiung and Jih-Jen Huang (2011). *Multiple Attribute Decision Making: Methods and Applications*. CRC Press (cit. on pp. 16, 20, 31).

Vickrey, William (1960). "Utility, Strategy, and Social Decision Rules". In: *The Quarterly Journal of Economics* 74.4, pp. 507–535 (cit. on pp. 89, 110, 113).

von Neumann, John and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press (cit. on pp. 7, 8, 15, 47, 75).

Wald, Abraham (1947). *Sequential Analysis*. Wiley (cit. on p. 20).

— (1950). *Statistical Decision Theory*. McGraw-Hill (cit. on p. 20).

Warren, Caleb, A. Peter McGraw, and Leaf Van Boven (2011). "Values and preferences: defining preference construction". In: *WIREs Cognitive Science* 2.193–205 (cit. on pp. 60, 61).

Weatherson, Brian (2012). "Knowledge, Bets, and Interests". In: *Knowledge Ascription*. Ed. by Jessica Brown and Mikkel Gerken. Oxford University Press (cit. on pp. 6, 9).

Wendt, Dirk and Charles Vlek, eds. (1973). *Utility, Probability, and Human Decision Making*. Vol. 11. Theory and Decision Library. D. Reidel Publishing Company (cit. on p. 10).

Williams, Bernard (1987). "Ethical Consistency". In: *Moral Dilemmas*. Ed. by Christopher W. Gowans. Oxford University Press (cit. on p. 25).

Wilson, Robert (1972). "Social choice theory without the Pareto Principle". In: *Journal of Economic Theory* 5.3, pp. 478–486 (cit. on pp. 81, 88).

Winterfeldt, Detlof v. and Gregory W. Fischer (1973a). *Multi-Attribute Utility Theory: Models and Assessment Procedures*. Tech. rep. Engineering Psychology Laboratory, University of Michigan (cit. on pp. 10, 31, 65).

— (1973b). "Multi-Attribute Utility Theory: Models and Assessment Procedures". In: *Utility, Probability, and Human Decision Making*. Ed. by Dirk Wendt and Charles Vlek. Vol. 11. Theory and Decision Library. D. Reidel Publishing Company. Chap. 2, pp. 47–86 (cit. on pp. 7, 10).