THE HANDBOOK OF RATIONAL & SOCIAL CHOICE

Edited by
PAUL ANAND
PRASANTA K. PATTANAIK
& CLEMENS PUPPE

THE HANDBOOK OF

RATIONAL AND SOCIAL CHOICE

This page intentionally left blank

THE HANDBOOK OF

RATIONAL AND SOCIAL CHOICE

An Overview of New Foundations and Applications

Edited by

PAUL ANAND

PRASANTA K. PATTANAIK

and

CLEMENS PUPPE

OXFORD

UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Cape Town
Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul Karachi
Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw
with associated companies in Berlin Ibadan
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2009

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
The Handbook of rational and social choice / edited by Paul Anand,
Prasanta K. Pattanaik and Clemens Puppe.
p. cm.
Includes index.
ISBN 978-0-19-929042-0
1. Decision making. 2. Social choice. I. Anand Paul. II. Pattanaik, Prasanta K.
III. Puppe, Clemens, 1960-
HD30.23.H356 2009
Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire
ISBN 978–0–19–929042–0
1 3 5 7 9 10 8 6 4 2

Contents

PART II SOCIAL CHOICE AND WELFARE

List of Contributors

Mohammed Abdellaoui, Centre National de la Recherche Scientifique

Carlos Alós-Ferrer, University of Konstanz

Paul Anand, Open University and Oxford University

Chuck Blackorby, University of Warwick

Han Bleichrodt, Erasmus University

Walter Bossert, Université de Montréal

David Donaldson, University of Stirling

Keith Dowding, Australian National University

Jürgen Eichberger, University of Heidelberg

Marc Fleurbaey, CNRS, University Paris–Descartes and London School of Economics

Wulf Gaertner, University of Osnabrück and London School of Economics

Simon Grant, Rice University

Alan Hájek, Australian National University

Edi Karni, Johns Hopkins University

David Kelsey, University of Exeter

Christian List, London School of Economics and Political Science

Paola Manzini, Queen Mary, University of London

Marco Mariotti, Queen Mary, University of London

Edward F. McClennen, Syracuse University

John Miyamoto, University of Washington, Seattle

Klaus Nehring, University of California, Davis

Prasanta K. Pattanaik, University of California, Riverside

Fabienne Peter, University of Warwick

Clemens Puppe, University of Karlsruhe

Daniel Read, Durham University

Karl H. Schlag, Universitat Pompeu Fabra, Barcelona

Ulrich Schmidt, Hannover Leibniz University and Christian-Albrechts-Universität zu Kiel

Erik Schokkaert, Katholieke Universiteit, Leuven

Kotaro Suzumura, Hitotsubashi University

Aki Tsuchiya, University of Sheffield

Bertil Tungodden, Norwegian School of Economics and Business Administration

Martin van Hees, University of Groningen

Timothy Van Zandt, INSEAD

Yongsheng Xu, Georgia State University

This page intentionally left blank

INTRODUCTION

PAUL ANAND
PRASANTA K. PATTANAIK
CLEMENS PUPPE

1 Introduction

The idea that we make decisions in order to improve human welfare is hardly unique to economics, but it is of particular importance to researchers in the related fields of decision theory and social choice. One of the most interesting features of these closely related fields is how, over the past three decades or so, they have reevaluated dramatically their understanding of what is analytically possible and normatively desirable. From work on decision-making and rational choice there is considerable evidence that people do not always behave in a manner consistent with expected utility—and these "violations" have undoubtedly influenced the development of new representation theorems as well as more conceptual work on the proper characterization of rationality. In almost parallel developments, the field of social choice and welfare has been similarly transformed by concerns about the proper informational basis for making ethical assessments of welfare, leading to the development of theories and evidence that operationalize approaches which are essentially nonutilitarian in nature.

This Handbook, based on contributions from many of the researchers who have made significant contributions to these literatures, brings some of the main developments together and highlights the related technical and normative themes that have helped to transform the way rational choice is theorized, at least in more

formal disciplines. Without any attempt at being comprehensive, we offer, below, some thoughts about the significance and recurrent ideas in the chapters to follow, in the hope that they will help the reader's understanding of a challenging but exciting corpus of research. To highlight the various links between the chapters, we have grouped them together in this overview, not strictly following their later ordering in the book.

2 Expected Utility and Beyond

It is not possible to understand the recent developments properly without some account of subjective expected utility theory, an account that is provided by Simon Grant and Tim Van Zandt in the opening chapter. For much of the twentieth century, economists used expected utility as a basis for modeling the integration of preferences and beliefs whilst allowing for differential attitudes to risk, because of its axiomatic structure. There are other interpretations of the decision calculus that it embodies, and the version on which the authors focus is one that also highlights the fact that probability can be interpreted as a measure of subjective belief. Their overview discusses a range of conceptual and technical issues that arise in the formulation of expected utility theory, including the hidden assumption of consequentialism, the weak axiom of revealed preference, the calibration of utilities and beliefs, and the linearity (independence) assumptions, to mention a few. In a number of cases (for example, dynamic utility and state-dependent expected utility) these topics are discussed in more detail by subsequent contributors, but what the review does show is just how much structure and how many issues can be formulated or conceptualized by the expected utility approach. Even though, for a growing number of purposes, one might take the theory as a point of departure, expected utility is a theory that clearly merits and rewards serious consideration and is open to many kinds of applications.

Moving on, however, and driven as much by empirical evidence as by interest in trying to understand the implications of generalizing expected utility, decision theorists have constructed a variety of alternatives to the subjective expected utility, among the best known of which is the rank-dependent utility (RDU) model. This was one of the first non-expected utility theories to be formalized and is both simple and powerful, as Mohammed Abdellaoui's survey demonstrates. The basic idea is that risky options should be evaluated by the weighted sum of their utilities, where the weights are generalized, nonlinear functions of the relevant probabilities and depend on the ranking of the outcomes to which they apply. This is an important generalization of expected utility as it enables us to model, for example, choices where small probabilities need to be treated nonlinearly either because we want

to ignore them or, conversely, because we wish to accord them "disproportionate" weight.

To some, dropping the linearity assumption was normatively questionable, but as the chapter notes, and as other Handbook contributors discuss in more depth elsewhere, there are good normative reasons to believe that rational agents need not be restricted to weighing utilities only in direct proportion to their probabilities. In addition, and from a historical perspective, the development of RDU has been closely associated with its capacity to explain behavior observed in experimental settings, and this success is a central focus of the chapter. As Abdellaoui notes, many studies have found the weighting function of probabilities to be *S-shaped*, mirroring findings on risky choice which tends to be concave (risk-averse) for gains and convex (risk-preferring) for losses. These findings are perhaps the most important empirical results in decision theory, and they are reflected in a more recent development of RDU known as cumulative prospect theory.

In the real world, of course, not all uncertainties come neatly packaged in terms of probabilities, and this fact has given rise to questions about other ways in which we might conceptualize credence. There are many proposals for alternatives, ranging from potential surprise through to weight of evidence, but one approach that has attracted widespread interest is based on the distinction between risk and uncertainty. Its origins can be traced back to discussions in Keynes and Knight, though it is the work of Ellsberg on range-defined probabilities and violations of linearity which is a particular starting point for modern research.

In their survey of this area, Jürgen Eichberger and David Kelsey discuss the practical motivation for interest in work on ambiguity, the currently most significant alternative to probability, Ellsberg's initial analysis of the area and follow-up work by experimentalists. Their particular interest concerns different ways in which the concept might be modeled, and their discussion includes different prior beliefs, Choquet integrals and capacities, and the unresolved issue of how to distinguish between the measure of ambiguity and our attitudes to ambiguity. Apart from a wide range of conceptual and technical issues, they discuss some intriguing economic applications which show how ambiguity aversion is able to explain some of the paradoxes observed in financial markets that other traditional theories cannot.

3 Rationality, Axioms, and Consistency

For a long time, it was felt that the axioms of expected utility, introduced by Ramsey and explored more explicitly by de Finetti and von Neumann and Morgenstern,

were compelling accounts of constraints on how decision-makers should behave. This view was somewhat paradoxically supported by an argument which held that rationality was purely instrumental and only prescribed how people should behave *conditional* on their preferences, whatever form they took. In recent years, the main substantive axioms of expected utility have been scrutinized from the perspective of their normative appeal, and in general, these arguments have been found to be wanting. In fact, two independent pieces of research deal in surprisingly similar ways with the axioms of transitivity and independence (linearity) and come to strikingly similar conclusions.

In his survey of arguments relating to the view that rational agents must always have transitive preferences, Paul Anand examines the logical structure of claims of attempts to argue for the view, as well as some of the accounts that explain why agents would want to violate transitivity. Transitivity's defence is represented by arguments to do with the logical structure of preference, its substantive meaning, and the money-pump argument, and in each case he demonstrates that there are logical problems in the cases made, and that they would be difficult to repair. Moreover, the chapter goes on to argue that there are situations in which violations of transitivity are to be expected, and that these are often associated with counterfactual or contextual aspects of choice: what a person might otherwise have chosen may determine what the maximizing action is. Though Anand emphasizes here the weakness of the founding arguments that were constructed to make the case for transitive preference, he also stresses the constructive value for decision theory in allowing a class of theories that drop transitivity. Not only can such models be formalized, but they also often appear to be most appealing when context is a significant aspect of the decision-making process, a point, he argues, that links both theories such as those developed by Peter Fishburn and our assessment of examples and counterexamples that have been much discussed by philosophers.

As we have already seen, a second axiom which has played a pivotal role in rational choice theory is independence (linearity), and the chapter by Ned McClennen surveys the arguments that try to justify why rational agents should accept linearity as a constraint on their preferences or choices. Like the survey of arguments concerning intransitivity, the chapter essentially focuses on logical validity and similarly concludes that all the significant arguments ultimately fail to demonstrate why a rational agent should weigh probabilities of outcomes only in a linear manner. Five kinds of arguments are highlighted: self-evident appeal, independence due to disjunctive options, independence as dominance, openness to reinterpretation, and a form of money-pump argument. Not all these are equally appealing or amenable to assessment, but some of them are closely connected with important themes in decision theory. For example, although the valuation of a particular gamble cannot appeal to complementarity of outcomes, which are mutually exclusive by construction, the chances of payoffs within gambles are conjoint. As Ellsberg's ambiguity

and other problems illustrate, this conjointness can be an obvious and appealing source of complementarity within the gamble that one would then expect to shape valuation.

Very similar kinds of arguments have been deployed in the context of trying to understand what constraints rationality might impose on belief, and these are surveyed in a chapter by Alan Hájek. In this case, the idea (developed most notably by Ramsey) is that a book can be made against an agent if his or her beliefs do not satisfy certain consistency constraints. Potentially, although the dominance or money-pump reasoning found in discussions about preference is very similar, beliefs suggest something more objective, even in Bayesian decision theory, which might give rise to slightly different conclusions. These issues have been discussed mostly by logicians, and their discussions raise concerns additional to those articulated in the two chapters discussed above. How we move from the axioms of a mathematical theorem about the probability calculus to constraints on human behavior is argued not to be trivial, though the issue is clearly a crucial one for all the decision sciences. Interpretation and justification are explored extensively in this essay, which leads to a discussion of synchronic and diachronic Dutch Book arguments, echoing a similar distinction in the context of money-pump arguments about preference structures. What differs, compared to those latter arguments, is the conclusion that some Dutch Book arguments are more plausible than others, a view that raises the question why this might be so and underwrites Hájek's call for a more unified treatment of the field.

4 Rationality over Time

Although for many purposes, theorizing choice as a point-in-time exercise with fixed options, beliefs, and values is a useful approach, we know that time can (in situations ranging from personal decision-making about finances, through human capital formation, to the allocation of environmental resources and intergenerational transfers) play a key role, and therefore merits explicit analysis. Some of the issues involved are touched on in a number of chapters, but two are notable for focusing on the valuation of temporally distinct outcomes and the process of learning.

In their chapter on choice over time, Paola Manzini and Marco Mariotti unpack the standard approach to discounted valuation (the "exponential model" used extensively by economists since Samuelson's advocacy of it) in which a deferred payoff is accorded less weight by means of a discount factor applied multiplicatively and repeatedly. The approach makes good sense when employed to account for the opportunity costs of money, but there has long been a question

in economics about what to do when it comes to choices involving *utility* over time.

Suppose you were indifferent to $1 today and $2 tomorrow, then by the stationarity axiom, you should be indifferent to $1 in a year's time and $2 in a year and a day's time. Many people, the evidence suggests, would not be so indifferent and have reward–delay trade-offs that vary over time, violating stationarity and the models that require it, exponential discounting included. As the authors note, there is little reason to accept this axiom, even normatively, and their survey goes on to evaluate a range of other discounting models, including their own theory of vague discounting, in terms of accuracy as descriptive theories as well as from a normative perspective. They do not conclude in favor of an outright winning model, but argue for a more theoretically rigorous assessment of empirical evidence, suggesting that experimental evidence has probably been confounded with attitudes to risk which, when properly accounted for, lead to much lower and probably more intuitively reasonable estimates of personal discount rates.

The second paper in this pair, by Carlos Alós-Ferrer and Karl Schlag, deals with a different, though related, topic and centers on the question of how agents learn about the value of options. Not only does learning take place over time, but it is expensive, and there has been growing interest in understanding ways in which agents can act to minimize such costs. For a variety of reasons which include learning, mimicry is an important strategy in the biological world, and it is a concept that has found its way into an increasing number of interesting economic applications, from the following of fads through to population dynamics and the identification of conditions which allow imitation to be a stable and successful strategy for subpopulations.

In the literature on which they focus, a learning rule is thought of as the probability of choosing an action that is conditional on having performed an action and observed its outcome, *and* on observing the outcomes and actions of other agents in the population. A variety of such rules can be formulated, including proportional imitation whereby the agent switches between acts depending on the difference between her own outcome and the best observed outcome achieved by others. The analysis bears similarities to work in evolutionary game theory, and it is natural to ask similar kinds of questions about such rules. For example, does following a learning rule lead decision-makers to select their best actions in the long run? What happens when Prisoner's Dilemma games are repeated by populations where only a minority of players are defectors? What happens if we allow for the fact that agents who observe each other may also be acting in similar environments? This chapter addresses these and other related problems whilst highlighting, in a number of places, topics and areas in need of further work.

5 Experimental Evidence and Economic Applications

Rational choice theories inform and are shaped by empirical work, and there are two kinds of empirical engagement surveyed in this volume. In a chapter by Daniel Read, we are invited to consider a range of experimental evidence that compares observed choices and decisions against the axioms of expected utility theory. To see what is as stake, the chapter discusses an attempt to make experimental violations of the independence disappears by defining a further lottery over the gambles being chosen. Evidence like this was used, at least early on, to suggest that people really were expected-utility-maximizers and that non-expected utility maximizing behavior was artificial.

However, Read's analysis cautions against overly simplistic interpretations of evidence. In the first place, the fact that new choice problems can be constructed in which independence is appealing, does not address the fact that there remain other simpler and possibly more transparent choice problems where nonlinearity is more desirable to many people. Secondly, even if the assumptions of expected utility were the only possible axioms of consistency, this does not mean that inconsistent behavior might not under certain circumstances bring about outcomes that were more desirable to the decision-maker. Thirdly, and possibly most significantly, the very fact that the axiomatic structure of choice behavior does vary across contexts might be regarded as further evidence itself against theories which incorporate context-free accounts of rational choice. This last possibility provides a focus for discussion in the chapter, which then concludes with a consideration of how the axioms of expected utility might most usefully be interpreted as small-world theories in Savage's sense, rather than global constraints on all actions.

A complementary survey, by Han Bleichrodt and Ulrich Schmidt considers uses of non-expected utility theories in economic analyses. In work on the purchase of insurance, a number of standard results turn out not to be robust to relaxing assumptions of expected utility—something the authors discuss in the context of insurance type demanded (whether it is deductible or not) and the best arrangements for sharing risk between members of a group.

They consider also the literature on auction design, an area of game theory where utility plays a significant role, and again are able to point to a number of results—for example, relating to optimal bidding in different auctions—which do not hold in a world where agents are non-expected utility-maximizers. Finally, they consider applications to health (insurance and medical decision-making) and highlight a number of important insights that have emerged. For instance, policies designed to enhance consumer choice between insurance plans have been argued to be unlikely

to work in the presence of loss aversion, whilst non-expected utility provides a more satisfactory framework for testing such policy measures.

6 State-Dependent Utility, Diversity, and Context

Within the Savage framework, options were deemed to be materially complete descriptions of the world, and though this is helpful for some purposes, the move distances theory from revealed preference methodology as observers have no independent way of telling what is materially complete for the decision-maker. It also puts a lot of potentially interesting features of decision problems beyond analysis, by assigning them to the option descriptions—which are primitive. However, for at least some purposes this approach doesn't go far enough, and it finesses the analysis of issues that can be central to choice and decision-making, as the next pair of chapters demonstrate.

In an overview of work on state-dependent utility, which picks up theoretically where Read's discussion of Savage's small-world justifications for expected utility leaves off, Edi Karni highlights a class of problems where context-free restrictions appear not to hold. As he notes in his discussion of examples concerning disability, employment insurance, and consumption behavior itself, there are many economic choice problems where the decision-maker's preferences cannot be assumed to be exogenous with respect to the prevailing state of the world. State-dependent utility functions may be ruled out by the assumption of state-independent preferences, but there are other reasons why this latter assumption might be undesirable. It is necessary to observe the decision-maker's response to a shift in state probabilities if one is to prize apart state-dependent preferences and beliefs, and the chapter discusses a method for doing this. In this case, many of the standard results from expected utility do seem to carry over to worlds where state dependence is allowed.

Another way of thinking about context is to say that the options themselves make up the context. Often we choose schools or go to particular shops because they have a wide range of courses or options, and this very different aspect of choice is discussed in a chapter by Klaus Nehring and Clemens Puppe. Their simple concern is how we should measure the diversity of a set of options, or more generally the diversity of sets of arbitrary objects, and this is a concern that spans both the field of individual decision-making and the field of social choice. Thus far, these fields have worked with feasible sets as the unit from which analysis begins, but that leaves open the question whether formal theory can say anything beyond highlighting formal properties of choices from feasible sets. Work on diversity suggests a positive

response, and as the authors note, diversity theory is potentially a particularly valuable tool for assessing species diversity in global environmental policy choice problems.

The literature which the authors address sees the similarity of objects as being the key to diversity assessments. A seminal contribution by Weitzman defines the marginal diversity of an element with respect to a set in terms of its minimal dissimilarity, and then derives the diversity of a set by recursive applications of this idea. However, the approach is not problem-free and has given rise to recent work that proposes a multi-attribute model in which, for example, the diversity value of a set of species is given by the total weight of all different features possessed by the species in the set. The survey of this area ranges from discussions of applications through geometrical representation to the distinction between absolute and relative notions of diversity and questions about the extent to which these latter concepts are related. In doing so, the chapter discusses techniques and concepts that not only are of interest in rational choice but also contribute to the discussion of freedom and its measurement, as discussed in the second part of this volume.

7 Beyond Utilitarianism—Formal and Philosophical Theories

The chapters in the second part of the handbook deal with a variety of topics in the theory of social choice and welfare, many of which have emerged only in recent years. Traditionally, welfare economics has been consequentialist insofar as it assesses the rightness of social actions in terms of the goodness of their consequences. Furthermore, it has tended to assume that the only consequences relevant for assessing the goodness of social actions are the utilities of individuals. Utilitarianism is a special case of consequentialism and is widely used as a basis for cost–benefit analysis and health-care measurement (where the interpretations and operationalizations are somewhat different).

Developments in social choice theory over the past couple of decades, however, have examined a variety of novel theoretical issues, most of which are not utilitarian and some of which might usefully be regarded as non-consequentialist. Some of these topics, especially those to do with freedom, equity, responsibility, and capabilities, as well as empirical work on health, demography, and theory testing, are covered in the surveys that follow; but two chapters, by Pattanaik and by Suzumura and Xu, help frame these new areas of work by considering some of the conceptual motivations, as well as the technical issues, to which they give rise.

In his chapter, Prasanta Pattanaik provides a summary account of the version of utilitarianism that underpins traditional welfare economics before considering a number of problems, particularly questions about the difficulties associated with the formulation and inclusion in social decision-making of individual rights and freedoms, agent-relative claims, and the demands of procedural justice. Freedom has been a prominent theme in social choice, and the chapter provides an account of how this line of work develops through Sen's pioneering contribution on rights and the literature inspired by it, and links back to topics discussed by John Stuart Mill in the nineteenth century. The chapter highlights the fact that utilitarian welfare economics has difficulty dealing with a number of important issues, including multiple preferences, endogenous desires, and the urgency of preferences before concluding with a discussion of some of the problems associated with the sum-maximization aggregation rule when applied to utilities.

In a manner that complements this overview, Suzumura and Xu survey an area which they have done much to develop. Specifically, their work provides a formal framework for characterizing non-consequentialist social choices at the axiomatic level. A key advantage of this approach is that it helps us to understand whether and how non-consequentialist claims might be formulated in a coherent manner; it also enables us to assess the extent to which these conditions are compatible with other desirable aspects of the social choice procedures.

The third chapter in this grouping highlights possibly one of the youngest sub-fields covered by this Handbook: namely, the field of judgment aggregation. It is also, perhaps inevitably, a relatively theoretical area, though the survey by Christian List and Clemens Puppe highlights practical motivation also. The central question for this literature concerns what happens when we want to aggregate individual judgments on logically interrelated propositions. At the heart of this work is a striking and novel paradox—if we consider aggregation by majority voting on premises and conclusions, it is possible for the majority judgment on the premises to be incompatible with the majority conclusions. The relevance of this problem for social choice is evident by noting that preference relations can be interpreted as logically or otherwise interconnected judgments on binary propositions of the form "x is better than y". The classical preference aggregation problem thus appears as a special case.

The literature that List and Puppe discuss develops conditions that might apply to judgment aggregation and then examines the extent to which these are mutually consistent. The starting point is rather Arrowian, and indeed one of the earliest theorems in the field, by List and Pettit, finds that there is no judgment aggregation procedure which satisfies four important conditions—universal domain, collective rationality, anonymity, and systematicity. These conditions can be generalized and refined, but, possibly more significantly, the literature now contains a number of theorems that demarcate precisely the borderline between the cases in which satisfactory aggregation is possible and those in which it is not. The literature is, as

we noted, relatively young, but the range of techniques and interpretational issues at play suggests a new dimension to social choice theory by embedding the preference aggregation problem in a larger conceptual context.

8 Justice and Welfare

Work on justice and inequality is perhaps more widespread within economics than might be supposed, notwithstanding the fact that efficiency and equity are the two core criteria by which changes in economic welfare are judged. This owes much to Vilfredo Pareto's discussion of the issues, but it would be difficult to overestimate the more recent influence that the philosopher John Rawls (and his *Theory of Justice*) has had on more recent developments in the field. In a sense, all the chapters on social choice are concerned with aspects of justice, though a number of them particularly help to highlight fairness and how theories are continuing to evolve both in response to current social change as well as in reaction to theories that may be less potent today than when they were first proposed.

In her analysis of the Rawlsian approach to justice as developed and revised over several decades, Fabienne Peter particularly identifies Rawls's interest in reconciling justice with value pluralism through democratic institutions in which society is seen as a fair system of cooperation and individuals as politically free and equal persons. Rawls's theory is profound in its derivation of just institutions and principles. The overview of his theory given in this chapter emphasizes his political conception of democracy, which departs from the welfare traditions on a number of points.

Peter's careful and potentially corrective reading of Rawls draws our attention to a number of features that have been overlooked or misinterpreted in discussions of his work. Rawls is, for example, often criticized for trying to ground the egalitarian outcomes of deliberations behind his famous veil of ignorance in rationality alone, whereas in fact he also recognizes that persons have an ability to be reasonable, and that this capacity plays a vital role in understanding how fair cooperation can come about. Her discussion is wide-ranging, but it includes concerns about primary goods which, though attractive to many economists, suggest a more objective analysis of inequality than welfarism might have encouraged, held back as it was by a reluctance to make interpersonal comparisons for reasons which Arrow attributed to respect for individual autonomy.

Autonomy has for a long time been important in economics, though only recently has it received the formal attention assumed more discursively. And if autonomy is important, then there are questions about how personal responsibility can, or should, be reflected in the welfare evaluation of social orderings and

economic policies. Such questions are the focus of a chapter by Marc Fleurbaey which provides an overview of this relatively recent literature and makes a number of connections between practical concerns in welfare economics (particularly redistributive policies) and their ethical foundations.

Two central conceptions of responsibility emerge from this overview. The first is a liberal approach in which personal responsibility tempers the urge to reduce or eliminate inequalities, and the second is a more utilitarian approach that can recognize the extent to which individuals are responsible for inequalities, though only insofar as this is relevant to the maximization of total utility. Various properties associated with both approaches are discussed, before a second distinction is made between compensatory policies and reward principles, which Fleurbaey then combines to derive a fourfold classification of social choice procedures. These approaches are compared before returning to the issue of freedom, particularly as discussed in the capabilities approach. Although there are a number of distinct and essentially different ways to respond to the theoretical challenges posed by responsibility, in the end, Fleurbaey argues that if fair comparisons of responsibility are to be made and used to shape redistributive policies, we should not employ only the evidence of the choices people make without also considering the worth of the feasible menu from which those choices were made.

Taking a cognate but more explicitly distribution-oriented view, Bertil Tungodden considers theories concerning the welfare and ethical bases for concern about inequality. Why is it that so many people are concerned about what they perceive to be large and unwarranted socioeconomic inequalities? How can our intuitions about inequalities be codified? And what arguments are necessary to support particular views about inequality? These are perennial questions in this area, and they are ones that continue to attract substantial attention.

Tungodden's particular focus is on providing an account of prioritarianism, an alternative distributional framework to egalitarianism and one that is based, as Derek Parfit put it, on the principle that benefiting people matters more, the worse off they are. He begins, however, with an overview of the conceptual foundations of egalitarianism in a discussion that emphasizes the need to impose restrictions on egalitarian intuitions if one is to come to any useful conclusions when comparing distributions. The discussion helps to highlight some of the conflicts between versions of equalitarian and utilitarian doctrines, both of which have been closely linked to economic theory, and it brings out prioritarian aspects of a literature that aims to understand the theoretical implications and demands of equality promotion. The chapter is especially notable for the direct links it is able to make between discussions about absolute poverty lines and measures of inequality with foundational issues, including philosophical discussions about the concept of sufficiency, and it concludes with a discussion of still open debates about the consequences of prioritarianism that follow from its analysis of distribution under conditions of uncertainty.

Even in the ethical realm, experimental work can play a vital role, both as a means of parametrizing and evaluating theoretical hypotheses and also as a way of directing theoretical interest or even exploring issues for which there is little or no satisfactory theory. In a chapter that helps to identify this literature, Wulf Gaertner seeks to make sense of the area of empirical social choice which begins to tell us something about the ethical theories people use in resource allocation problems.

As Gaertner notes, equality is an idea that was discussed by Aristotle in ways that relate closely to current concerns though the modern literature largely takes off from work by Yaari and Bar-Hillel who note four reasons (different needs, tastes, beliefs and effort/contribution) for departing from equal division in just allocation problems. Some of the choices examined in these experiments involve realistic, if representative, moral dilemmas (for example, do you use limited public funds to better the situation of a handicapped child or alternatively improve the education of possibly more than one "gifted" child?) and Gaertner is able to report longitudinal results based on work with German students over a 13 year period as well as comparative work with students from Israel and central Europe which evidence sometimes dramatic variations across culture and time. Perhaps one of the most significant findings in this area is the fact that Pareto optimal improvements are widely supported but are objected to by a small number when the recipient of gain was someone else, and by a larger number when the beneficiary is better off than the assessor.

9 Health and Demography

As we have already suggested, applications can serve a number of valuable scientific roles, though inevitably newer theories take time to feed through to such work. Nonetheless, these areas are important both in their own right, because of their substantive value, and because they illustrate, more generally, how the decision sciences can shape, and be informed by, the needs of policymakers.

Health remains an area where both rational and social choice have been applied extensively and consistently, thereby providing public policy with principled ways of respecting patient autonomy, improving efficiency, and fostering procedural fairness. The quality-adjusted life-year (QALY) has been advanced, at different times, as an aid to all three of these desiderata, and therefore is of particular significance in the intellectual history of applied decision science. In their chapter, Aki Tsuchiya and John Miyamoto consider some of the questions that have arisen from the engagement of rational and social choice with health. How should health gain be measured at the individual level? And how should such measures be incorporated into the social welfare function? These are two key questions that dominate work in

this area. The QALY lies at the center of answers to these questions and has attracted a number of attempts to justify its use, two of which are focused on by the authors of this chapter—a welfarist approach which emphasizes the link to average utility functions of the population and a non-welfarist approach which emphasizes a more political-economic justification in terms of what health institutions have been set up to promote. The authors consider the extent to which utility sum maximization is an acceptable social choice rule in this domain and the manner in which issues of equity and inequality aversion have been incorporated into the social choice rules, thereby giving us an invaluable feel for how theory is being developed at the sharp end, so to speak, of policy and practice.

One of the most interesting and challenging application areas is the formalization of utility theory applied to problems of population ethics. The area has been developed particularly by Charles Blackorby, Walter Bossert, and David Donaldson, who survey work that had its early beginnings in the analyses of Derek Parfit, a philosopher and author of the "repugnant conclusion" problem which is still much discussed. Briefly, Parfit's idea was that one might object to social metrics of well-being which prescribe possibly enormous increases in population size, to increase total welfare, at the cost of quality of life. For many, there seems to be a critical minimal level below which the quality of a life fails to justify its existence, and the authors of this chapter discuss how such critical level utilitarian philosophies can be formalized. Whether a life worth living must be modeled as a critical level of utility is an open question, but if one accepts the case, their demonstration that critical level utilitarianism gives rise to a social ordering under certain conditions is surely a result of central importance.

10 Freedom and Capabilities

Looking across economic theory, philosophy, and the field of economic development, possibly the single most important, shared interest over the past two decades is found in work concerning the theoretical analysis and practical implications of freedom. When this new work started to emerge, one could caricature freedom, not too unfairly, as a concept that was widely valued but largely unanalyzed in economics, a captive of libertarian politics, and the subject of a philosophical debate about the distinction between positive and negative freedom that had run its course. Whilst this newer research may not have resolved all the questions that could be of interest, it does appear to have identified a much richer set of topics that are more compatible with a wider range of approaches to justice.

In a chapter that focuses on some of the key issues in formal theory and political philosophy, Keith Dowding and Martin van Hees survey a literature emanating

largely from a paper by Pattanaik and Xu (1990). That paper discussed three axioms of freedom in particular, and showed that the only measure satisfying all three conditions was the cardinality rule. This rule measures freedom in terms of the number of options available, but there has been widespread agreement from the start that this is unreasonable. For example, is the freedom to choose between a pair of beans really the same as the freedom implied by being able to choose between two rewarding jobs?

Dowding and van Hees suggest that the key to understanding the difficulty faced by attempts to measure freedom is the distinction between the extent of our freedom, on the one hand, and its value, on the other. The extent of an agent's freedom, they suggest, is best accounted for by the diversity of the options, an idea explored in more depth in the chapter by Nehring and Puppe, whilst the value of freedom has something to do with the real opportunities (in the ordinary sense) that a person has. The chapter goes on to consider a range of issues that derive from this distinction, blending conceptual and formal debates into their assessment, and concludes with a discussion of issues that warrant further research. Particularly notable is their call for the introduction of institutions into the theorizing of freedom and a discussion of how a person's own preferences, through the outcomes of strategies they may induce in others, can determine how free that person is.

The value of freedom, through Amartya Sen's work in economics and philosophy and international policymaking circles, has also become the center of an approach to social choice and welfare economics known as the "capabilities approach", one that is surveyed in a final chapter by Erik Schokkaert. Originating in concerns about the informational basis of utilitarian welfare economics, which emphasized the importance of recognizing additional ethical claims, like rights, the approach has evolved significantly over time but remains true to a simple formal framework that distinguishes between what people can do (capabilities), as opposed to what they do (functionings), and emphasizes the multidimensional nature of functionings and consumption, and the heterogeneity of individuals in transforming their resource entitlements into welfare outcomes.

Schokkaert's chapter provides an overview of some of the topics and disciplines that have been touched by the approach, highlighting theoretical issues concerning the identification of capability sets, their measurement in empirical settings, the selection of appropriate dimensions, and the still open issue of weighing different dimensions to provide a single index of capability or functioning. This chapter is particularly rich with challenges and questions to researchers wanting to take the area forward, but it is also constructive about ways in which theory can guide empirical work, and indeed has done so already. As the author notes, the approach has been influential in development policy by shaping the evolving construction of the Human Development Index, with the result that there is now a widespread consensus that poverty and quality of life are essentially multidimensional issues. It would be difficult to overestimate either how much the capabilities approach has

influenced thinking about the meaning of economic development, as Schokkaert allows, or, as the volume suggests, the extent to which this approach owes its policy insights to a thoughtful consideration of the foundations of social choice.

11 Summary

Whilst it would be difficult for any single volume to be comprehensive, these chapters collectively cover a number of key, normative developments in the related fields of decision theory and social choice, fields that in turn underpin much of the analysis in economics and beyond. Furthermore, in this overview we have highlighted only some of the issues covered in the chapters, with a slight bias towards some of the recurrent themes. In both decision theory and social choice, there have been parallel movements away from the basic theories towards more general or explicit accounts that have desirable normative features or provide a better fit with the empirical evidence.

Decision theory is very much concerned with the conceptualization of preference and belief and the selection of action based on these two inputs. Subjective expected utility theory provides a useful framework for understanding such issues, but there are issues to do with the conceptualization of uncertainty and the significance of context that require different approaches to those found in the work of Savage or von Neumann and Morgenstern. In addition, a set of questions which might loosely be thought of as being related to context provide grounds for wanting to generalize both transitivity and independence, assumptions that it turns out can indeed be weakened in a variety of contexts that include general equilibrium theories.

Social choice, on the other hand, has its origins in Arrow's formalization of Condorcet's voting paradox and for some time focused on preference aggregation. Sen's concerns about the inappropriate informational basis for social choice to which that framework gave rise have, in turn, also encouraged research which now emphasizes issues of equity and freedom in group decision-making. The role of rights alongside utilities suggests a need to integrate different kinds of claims, though the literature has moved on to consider what people are free to do. Freedom is concerned with the amount and diversity of choice that people have, and this is clearly difficult, though not impossible, to measure. Moreover, the capabilities implied by these claims are closely related to values like autonomy and responsibility, for which accounts also need to be given. And perhaps most significantly, it is perfectly possible to be concerned about the distribution of freedoms within a society and thereby to progress from the rudimentary idea that freedom is identified only by the absence of other-imposed constraints, important as this is.

Although much of the work is technically demanding, the picture is one of fields that continue to grow and develop in exciting ways. Issues to do with dynamics, multidimensionality, and the role of reasons all raise many, often connected questions that are only beginning to be addressed. Moreover, the papers highlight a fertile dialogue between economic theory and philosophy that has continued to flourish in recent years whilst involving valuable exchanges with other areas of social science. The rules of the game have surely not been set in stone, however, and what we now know should encourage us to continue thinking about how best to construct adequate theories of rational and social choice.

This page intentionally left blank

PART I

UTILITY THEORY, RATIONALITY, AND DECISION-MAKING

This page intentionally left blank

CHAPTER 1

EXPECTED UTILITY THEORY

SIMON GRANT
TIMOTHY VAN ZANDT

1.1 Introduction

This Handbook is a modern look at decision theory and social choice. It emphasizes recent developments that go beyond the classic theory of utility and expected utility. However, a sensible starting point is the classical theory, a benchmark that will frame departures considered in subsequent chapters. Furthermore, it remains the prominent workhorse for applied economics.

This chapter presents the main features of that classical theory. The topic is broad and has been treated so much elsewhere that we do not attempt to be technically comprehensive. Instead, we opt for a succinct and pedagogical treatment that will be accessible to nonspecialists yet still useful to the specialist for framing the subsequent chapters.

The theory reviewed here has had a long and complicated development. Expected utility dates back at least to Bernoulli (1738). As a resolution to the famous St. Petersburg paradox, he used a logarithmic utility index defined over wealth to compute a finite price for a gamble with an unbounded expected value. Even the

We are deeply indebted to Peter Wakker for providing detailed comments on two earlier drafts and helping shape this chapter. We also benefited from the feedback of Paul Anand and Clemens Puppe, from discussions with Enrico Diecidue, and from corrections by Philippe Delquié and Alminas Zaldokas.

modern foundations are a half-century old, with the development of an axiomatic choice-theoretic foundation for expected utility occurring roughly from the mid-1920s to the early 1960s.

We neither structure the chapter around such historical development nor attempt to explain it; a proper treatment would be too lengthy and would distract from the content of this chapter, whereas a succinct treatment would have too many unfair omissions. Instead, our goal is to clarify the interpretation and mathematical role of the various axioms that form the theory. We have the benefit of hindsight and of our freedom to deviate from the chronology of the theory's development.

Rather than delve immediately into choice under uncertainty, we start from scratch with basic preference theory (Sections 1.2–1.7). Not all of this volume relates to decision under uncertainty, but our other motive for reviewing this theory is to emphasize the assumptions that are already implicit when one represents choices over uncertain prospects by a complete and transitive preference relation.

The fundamental mathematical properties of expected utility representations are additivity and linearity. Our main expository innovation is to remain in the general choice setting a little longer (Sections 1.8–1.10) in order to explore the independence assumptions that lead to such structure *without reference to decision under uncertainty*. We can thereby (a) clarify the mathematics without the baggage of the uncertainty framework and (b) discuss the axioms in other contexts, such as multi-attribute decision problems and intertemporal choice, so that their significance in decision under uncertainty is then better understood.

We thus do not reach decision under uncertainty until the middle of the chapter. We first introduce the "states of the world" (Savage) framework for representing uncertainty in Section 1.11. This precedes our treatment of the lotteries representation of objective uncertainty (Section 1.13), in a reversal of the usual (and historical) order, because the states framework is useful and intuitive whether uncertainty is subjective or objective and because we refer to it when interpreting the axioms of the lotteries model. Overall, we proceed as follows. After presenting the states framework in Section 1.11, we take our characterization of preferences and utility as far as an independence axiom and state-dependent expected utility in Section 1.12. We then turn to expected utility for lotteries (objective uncertainty) in Section 1.13. In Sections 1.14 and 1.15, we return to the full subjective expected utility representation in the states model with state-independent utility.

As noted, we emphasize the link between the independence axioms in consumer theory, in expected utility for objective lotteries, and in expected utility under subjective uncertainty. Fishburn and Wakker (1995) provide a comprehensive study of how these different versions arose and evolved historically. The way in which we motivate the axioms—as following from dynamic consistency and consequentialism—is analogous to the treatment in Hammond (1988).

Our decision to study additive and linear utility before any applications to decision under uncertainty has the advantages we claimed above. It comes at two costs.

First, those familiar with expected utility theory may be disconcerted to see, in the first half of the chapter, familiar axioms and theorems with only passing references to the seminal work in expected utility theory from which they originated. However, such omissions are rectified in the second half of the chapter, and with this warning, the reader should suffer no ill consequences. The second cost is that we have to invent "context-free" names for several old axioms from expected utility theory.

1.2 Descriptive, Prescriptive, and Normative Theories

Decision theory has two goals: to *describe* how agents *do* make decisions (descriptive decision theory) and to *prescribe* how agents *should* make decisions (prescriptive decision theory). As in any theoretical modeling, decision theory balances accuracy and simplicity. A prescriptive decision theory that is too complicated to learn or implement is hardly useful for a decision-maker. A descriptive theory should be simple, because it is meant to be a framework that organizes and unifies a variety of situations and behavior, because it should be tractable enough to derive conclusions, and because we may need to estimate the parameters of the theory from a limited amount of data.

A third branch of decision theory, normative decision theory, tries to describe how a hypothetical, infinitely intelligent being would make decisions. This may sound speculative and impractical, but it provides important foundations for descriptive and prescriptive theories. A normative theory is inherently simpler than an overtly descriptive or prescriptive theory, because it need not concern itself with such complications as (a) errors or forgetting and (b) the heterogeneity of the intelligence and experience of decision-makers. There are only a few ways to be perfect, but many ways to be imperfect!

Simplicity is good, but how does a normative theory serve the goals of descriptive or prescriptive theory? Humans are goal-oriented and work hard to pursue their goals. Any descriptive theory should capture this first-order consideration, and a normative model is a powerful and parsimonious way to do so. To develop a prescriptive theory that helps mortal humans (with all their limitations) make decisions, it is useful to understand how unboundedly rational beings would make decisions.

This chapter develops classical rational choice theory, particularly expected utility theory, as a normative model. We leave it to subsequent chapters to evaluate how well it serves as a descriptive or prescriptive theory.

1.3 A Review of Choice, Preferences, and Utility

Before tackling decision-making under uncertainty, we study the theory of choice without uncertainty. Let's call our decision-maker "Anna".

In most of this chapter, we limit attention to *static choice*. This means we consider a single decision that Anna makes. We thus suppress the fact that Anna may anticipate having to make choices in the future and that current decisions are intertwined with such future decisions. However, we do not thereby suppress time: the objects of choice could be time paths of consumption.

We ignore much of the information processing that Anna must do in order to make a decision. The only input to Anna's problem is the set of possible alternatives, which we allow to vary. There is a fixed set X containing all potential alternatives, and Anna is presented with a set $A \subset X$ from which she must choose, called a *choice set*. For example, X is the set of all consumption bundles and A is the budget set, which depends on prices and the agent's wealth. Alternatively, X could be the set of potential presidential candidates and A the set of candidates on the ballot. We want a model of what Anna would choose from each set of feasible alternatives.

Assume that X is finite, and that every nonempty subset A of X is a potential feasible set. For each set $A \subset X$ of feasible alternatives, let $C(A)$ be the elements that Anna might choose from A. She must always choose something, which means that $C(A)$ is nonempty, but $C(A)$ may contain more than one item because of indifference. Call $C(\cdot)$ her *choice rule*. Our game is to find some conditions that lead to a simple representation of choice. The representation should allow one to derive qualitative conclusions in models without knowing specific choices and should have few parameters, all of which could be estimated empirically.

There are two complementary approaches. One is to start with a choice rule as a primitive. From this "empirical" object, one derives (revealed) preferences. The other approach is to start with preferences as the primitive. From these, one derives the choice rule. We follow the first (revealed preference) approach.

1.4 Hidden Assumption: Consequentialism

We will be as explicit as space permits about hidden assumptions. For example, we have already outlined the static nature of the model. Another hidden assumption is consequentialism: Anna cares about *consequences*—not how consequences are achieved.

In the decision model we have developed so far, "consequences" refers to the alternatives, and "consequentialism" means that Anna does not care how she ends up facing a particular choice set $A \subset X$ or how she chooses from A. For example, in consumer theory, one might suppose that Anna cares only about consumption and not about the prices that determine which consumption bundles are affordable. In decision-making under uncertainty, we may suppose that Anna cares only about probabilities of different outcomes and not about how these probabilities are generated (e.g. about whether uncertainty is resolved in one step or instead in a multistage lottery).

We can redefine consequences in order to circumvent any particular violation of consequentialism. If Anna cares about prices, then we define a consequence not only by how much she consumes of different goods but also by the prices she faces (we have thus redefined the set X). If Anna cares about the stages at which uncertainty unfolds, then a consequence can be defined to include such information.

However, if we proceed this way without constraint, then—in our abstract model—a consequence would be defined by not only the element of X that Anna ends up with but also by the set A from which she was able to select it. Hence, there would be no link or consistency conditions between the choices Anna might make from $A \subset X$ and those she might make from $A' \subset X$. We would have a vacuous theory. Instead, we must be able to define consequences in a sufficiently restrictive way that we can envision Anna being presented with different sets of consequences and caring only about those consequences, not about how those sets were generated.

From the empirical side of the theory, there is another hidden assumption. We are supposedly studying a single decision that Anna makes. However, for that decision she will end up facing a single choice set $A \subset X$. We will never be able to observe an inconsistency between $C(A)$ and $C(A')$ unless one of the following two conditions is satisfied.

1. When presented with different *hypothetical* choice sets, Anna can report the choices she would make.
2. Anna faces different instances of this problem at different times, and how she chooses from any choice set does not vary from one instance to the other.

Thus, we assume that one of these two conditions holds.

1.5 Weak Axiom of Revealed Preference

Let x and y belong to X. Then x is *revealed weakly preferred* to y if $x \in C(A)$ for some $A \subset X$ containing x and y (i.e. if y is available but x may be chosen). If also

$y \notin C(A)$, then we say that x is *revealed preferred* to y, or, if we want to emphasize that the preference is not weak, that x is *revealed strictly preferred* to y.

The *weak axiom of revealed preference* (WARP) states that if x is revealed weakly preferred to y, then y is not revealed preferred to x. This is defined explicitly in Axiom 1.

Axiom 1 (WARP). Let x and y belong to X, and let A and B be subsets of X containing x and y. If $x \in C(A)$ and $y \in C(B)$, then $x \in C(B)$.

WARP is a natural axiom of a normative theory, but it is at best an approximation for descriptive or prescriptive theories. Especially when choice sets are large and complex, achieving such consistency is difficult.

This one consistency condition gets us far. First, it implies that the choice rule can be summarized by or deduced from more limited information: the binary choices or *preferences*. To make this statement more precise, we first define, for $x, y \in X$,

$x \succcurlyeq y$ if $x \in C(\{x, y\})$ (x is *weakly preferred* to y);
$x \succ y$ if $x \succcurlyeq y$ but not $y \succcurlyeq x$ (x is (*strictly*) *preferred* to y);
$x \sim y$ if $x \succcurlyeq y$ and $y \succcurlyeq x$ (x is *indifferent* to y).

The binary relations $\succcurlyeq$, $\succ$, and $\sim$ are called the (weak) preference, strict preference, and indifference relations, respectively.

Definition 1. *Let $\succcurlyeq$ be the preference relation defined for a choice rule $C(\cdot)$ as in the previous paragraph. Then the choice rule satisfies* preference maximization *if, for every $A \subset X$ and $x \in A$,*

$$x \in C(A) \Leftrightarrow x \succcurlyeq y \quad \forall y \in A.$$

In words, choices from large sets are consistent with binary choices. If we know Anna's binary choices (preferences), then we can derive $C(\cdot)$. Preference maximization implies a considerable savings in the amount of information required in order to know $C(\cdot)$.

It is also useful for Anna's preferences to satisfy some consistency conditions themselves.

Definition 2. *A binary relation $\succcurlyeq$ is a* weak order *if it satisfies the following conditions.*

1. *(Completeness) For all $x, y \in X$, we have $x \succcurlyeq y$ or $y \succcurlyeq x$ (or both).*
2. *(Transitivity) For all $x, y, z \in X$, if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $x \succcurlyeq z$.*

Completeness of the preference relation follows from the assumption that C has nonempty values. Transitivity is the important consistency condition; it follows from WARP.

Proposition 1. *The choice rule $C(\cdot)$ satisfies WARP if and only if (a) it satisfies preference maximization and (b) the preference relation is complete and transitive.*

Proof: Samuelson (1938) defined WARP in the context of consumer theory (budget sets) and single-valued demand. Arrow (1959) defined WARP (as we did) for general choice rules—that is, for abstract choice sets and possibly multivalued choices. Proposition 1 is implied by theorems 2 and 3 in Arrow (1959). □

One advantage of complete and transitive preferences is that they can be represented by a utility function.

Definition 3. *Let $\succcurlyeq$ be a preference relation on X and let $U: X \to \mathbb{R}$ be a function. Then U is a* utility representation *of $\succcurlyeq$, and $\succcurlyeq$ is* represented by the utility function *U, if for all $x, y, \in X$ we have*

$$x \succcurlyeq y \Leftrightarrow U(x) \geq U(y).$$

Proposition 2. *A preference relation $\succcurlyeq$ on X is complete and transitive if and only if it has a utility representation.*

Proof: The utility representation is easily constructed recursively. See Birkhoff (1948, thm 1, p. 31) for such a proof for countable X. □

Thus, WARP also means that Anna makes choices *as if* she maximized a utility function, as stated in the next corollary.

Corollary 1. *The choice rule $C(\cdot)$ satisfies WARP if and only if there is a utility function $U: X \to \mathbb{R}$ such that*

$$C(A) = \arg\max_{x \in A} U(x)$$

for all $A \subset X$.

If $U: X \to \mathbb{R}$ is a utility representation of a preference relation $\succcurlyeq$ on X and if $f: \mathbb{R} \to \mathbb{R}$ is any strictly increasing function, then $f \circ U: X \to \mathbb{R}$ is also a representation of $\succcurlyeq$. (The composition $f \circ U$ is called a *monotone transformation* of U.) Therefore, the magnitudes of the utility values of a representation have no particular meaning—they only define the ordinal relationship $\succcurlyeq$.

1.6 From Preferences to Choice Rules

Suppose instead that we let Anna's preference relation $\succcurlyeq$ be the primitive and then derive choices from this relation. (The primitive is thus any binary relation on X, which attains meaning as a preference relation through the subsequent use made of

it in the model.) In this approach, preference maximization becomes an *axiom* that defines the choice rule: For $A \subset X$,

$$C(A) = \{x \in A \mid x \succcurlyeq y \ \forall y \in A\}.$$

Given this axiom, the standard way to proceed is by assuming that preferences are complete and transitive, and then concluding that (a) the choice rule $C(\cdot)$ satisfies WARP and (b) preferences have a utility representation.

We highlight the axiom of preference maximization—that choices from large sets be consistent with binary choices—because it is a critical consistency condition. If one observed a violation of WARP or a violation of transitivity, then it would be reasonable to doubt preference maximization. Retaining it and merely tweaking the axioms on preferences could be a poor way to develop a theory that encompasses such empirical violations.

1.7 Infinite Choice Sets and Continuity Assumptions

Infinite choice sets require two modifications to the theories described so far. One such modification is that the choice rule might not be defined on all subsets of X: it has a restricted domain. For example, if $X = \mathbb{R}_+$ represents monetary prizes, and if Anna would always choose the higher of two monetary prizes, then preference maximization would imply that Anna's choice from all of X or from $[0, 1)$ is not well-defined.

This should be viewed as a nonsubstantive technical restriction. In practice, it is fair to say that Anna would always make a decision. (Not announcing a choice also represents a decision, since there must be some default outcome that occurs.) The kind of choice sets that are suppressed from the domain of C typically are unrealistic ("you can have as much money as you want"), arise only because an inherently discrete problem has been modeled as a continuous approximation ("you can have any amount of money in $[0, 1)$"), or would lead to well-defined choices if certain aspects of the problem were fully modeled (e.g. if the time it took to get finer and finer divisions of a good were taken into account).

The second modification of the theory, which may be needed only if the choice set is uncountable, is that there may be an important role for some kind of continuity assumption on the preference relation.

Definition 4. *Let X be endowed with a topology. A preference relation $\succcurlyeq$ on X is* continuous *if $\{y \in X \mid y \succcurlyeq x\}$ and $\{y \in X \mid x \succcurlyeq y\}$ are closed for all $x \in X$.*

One representation theorem on infinite choice sets is the following.

Theorem 1. Suppose that X is a separable metric space and that $\succcurlyeq$ is continuous, complete, and transitive. Then $\succcurlyeq$ has a continuous utility representation.

Proof: See theorem II in Debreu (1954). The actual topological assumption is more general: that the topology on X have a countable base. For a metric space, this assumption is equivalent to separability. See Debreu (1964) for a discussion and alternate proofs of variants on this result. □

In Theorem 1, the continuity assumption is needed in order to obtain any utility representation at all, but it has the added benefit that the theorem shows the existence of a continuous representation. In the approach that starts at preferences and derives choices, continuity assumptions are used to ensure that choices are well-defined at least on compact sets. Continuity of the representation may be important in further applications.

However, it would be surprising if continuity of preferences were required for the mere existence of a (not necessarily continuous) utility representation. In fact, there are nontopological variants of Theorem 1 (and of many other theorems in this chapter) that use an "order-theoretic" or "algebraic" approach. We present one example.

Define a subset Z of $(X, \succcurlyeq)$ (or of any weakly ordered space) to be *order-dense* in $(X, \succcurlyeq)$ if, for all $x, y \in X$ such that $x \succ y$, there exists $z \in Z$ such that $x \succcurlyeq z \succcurlyeq y$. For any $Y \subset \mathbb{R}$, the linearly ordered space $(Y, \geq)$ has a *countable* order-dense subset. One can think of a utility representation as embedding the ordered space $(X, \succcurlyeq)$ into $(\mathbb{R}, \geq)$. Therefore, $(X, \succcurlyeq)$ must also have a countable order-dense subset. This necessary condition is also sufficient for the existence of a utility representation.

Theorem 2. If $\succcurlyeq$ is complete and transitive and contains a countable order-dense subset, then $\succcurlyeq$ has a utility representation.

Proof: This is lemma II in Debreu (1954). It is closely related to results for linear orders by Cantor (1895) and Birkhoff (1948, thm 2, p. 32). (See a complete proof of Birkhoff's theorem in Krantz *et al.* (1971, thm 2.2).) One proves Theorem 2 by studying the induced linear order over equivalence classes of $(X, \succcurlyeq)$ and thereby reducing it to the case of linear orders. This is made explicit in theorem 3.1 of Fishburn (1970).

We provide a warning to the reader. The term "order-dense" has several similar definitions in the mathematics and mathematical economics literature. For example, for a linearly ordered space $(Y, \geq)$, Birkhoff (1948) defines $Z \subset Y$ to be *order-dense* in $(Y, \geq)$ if, for all $x, y \in Y \setminus Z$ such that $x > y$, there is $z \in Z$ such that $x > z > y$. This definition is equivalent to ours for a linearly ordered set but not for a weakly ordered set. □

Theorems 1 and 2 are linked by the following fact: If X is a separable metric space and $\succcurlyeq$ is continuous, then $(X, \succcurlyeq)$ has a countable order-dense subset.

An example of preferences that have no countable order-dense subset and hence have no utility representation are lexicographic preferences on $\mathbb{R}^2_+$: $(x_1, x_2) \succ (y_1, y_2)$ if and only if either $x_1 > y_1$ or $x_1 = y_1$ and $y_1 > y_2$. Any order-dense subset of $(\mathbb{R}^2_+, \succcurlyeq)$ is not countable because it must contain an element for every value of the first coordinate.

Apart from this brief aside, we stick to the topological approach throughout this chapter. The topological assumptions tend to be both easier to state and more familiar to the reader than are the algebraic assumptions. The interested reader may consult Krantz *et al.* (1971) and Fishburn (1970) for further development of the algebraic approach.

1.8 Additive Utility

Maintained assumptions. The rest of this chapter is about the *additional* structure that preferences and their utility representation may have. Therefore, as we proceed, and even as the set of alternatives varies, we avoid repetition by *maintaining the assumptions that choices satisfy preference maximization, that the symbol $\succcurlyeq$ denotes the preferences relation, and that $\succcurlyeq$ is complete and transitive.*

1.8.1 Overview

We study two types of structure that one might impose on preferences or utility: additivity and linearity. Why do we tackle these technical topics in the setting of abstract choice theory when this chapter is supposed to focus on choice under uncertainty? In decision under uncertainty, additivity is the most important structure imposed on preferences over state-contingent outcomes, and linearity is the most important assumption imposed on preferences over lotteries. We would like to understand the mathematics of these structures in the absence of a particular interpretation so that (a) the mathematical structure is clearer, (b) when we reach decision under uncertainty, we can focus on their interpretation rather than the math, and (c) we can better understand the relationship between expected utility theory and, say, intertemporal choice.

1.8.2 Additive Separability

In this section, we assume that X is a product set: $X = X_1 \times \ldots \times X_J$. Each component $j = 1, \ldots, J$ is called a *factor*. Here are some possible cases.

1. There are J goods; $(x_1, \ldots, x_J) \in X$ is a consumption bundle; and $X_j = \mathbb{R}_+$ is the set of possible quantities of good j.
2. There is one good with multiple varieties, which are characterized by J attributes; a particular variety is given by a list $(x_1, \ldots, x_J) \in X$ of attributes; and X_j is the set of possible values of attribute j.
3. There are J periods; $(x_1, \ldots, x_J) \in X$ is a time path of consumption; and X_j is the set of possible consumption bundles in period j.
4. There are J states of nature; $(x_1, \ldots, x_J) \in X$ is a state-contingent outcome; and X_j is the set of possible outcomes in state j.
5. There are J people; $(x_1, \ldots, x_J) \in X$ is an allocation; and X_j is the set of possible consumption bundles of agent j. (The decision-maker is an outside observer.)

Given preferences $\succcurlyeq$ on X, we are interested in whether there is a representation of the form

$$U(x_1, \ldots, x_J) = \sum_{j=1}^{J} u_j(x_j)$$

for some functions $u_j \colon X_j \to \mathbb{R}$. Such a representation is called *additive* or *additively separable*.

Suppose such a representation exists. Recall that any monotone transformation of U represents the same preferences. However, it need not be additive. The family of additive representations can be fairly narrow. Under a variety of assumptions, including those of Theorem 3 to follow, an additive representation is unique up to an affine transformation. That is, V is another additive representation if and only if $V = a + bU$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}_{++}$. The magnitudes of the utilities of the different factors have some meaning because they are aggregated across factors to determine the overall ranking of alternatives.

For example, suppose $x^1,\ x^2,\ x^3,\ x^4 \in X$, with $x^1 \succ x^2$ and $x^3 \succ x^4$. Compare the "extra kick" from getting x^1 rather than x^2 with the extra kick of getting x^3 rather than x^4 and measure this comparison by the ratio

$$\frac{U(x^1) - U(x^2)}{U(x^3) - U(x^4)}. \tag{1}$$

This ratio is the same for any additive representation U, yet can have any positive value if we allow for non-additive representations.

For this reason, additive utilities are often referred to as *cardinal* utilities. However, only ratios such as (1) are uniform across additive representations. Thus, additive utilities are not interpersonal cardinal scales as used in utilitarianism.

1.8.3 Joint Independence

An immediate consequence of an additive representation is that preferences satisfy *joint independence*, meaning that "how one ranks what happens with some factors does not depend on what happens with the other factors".

We introduce some notation to formalize joint independence. For a given set $K \subset \{1, \ldots, J\}$ of factors, let $X_K \equiv \prod_{j \in K} X_j$. For a partition $\{K, L\}$ of $\{1, \ldots, J\}$, we may write an element of X as (a, c), where $a \in X_K$ denotes the values for the factors in K and $c \in X_L$ denotes the values for the factors in L. For the special case in which K is a singleton $\{j\}$, we may write (x_j, x_{-j}), where $x_j \in X_j$ and $x_{-j} \in X_{-j} \equiv \prod_{k \neq j} X_k$.

Axiom 2 (Joint independence). For all partitions $\{K, L\}$ of $\{1, \ldots, J\}$, for all $a, b \in X_K$, and for all $c, d \in X_L$,

$$(a, c) \succcurlyeq (b, c) \Leftrightarrow (a, d) \succcurlyeq (b, d).$$

Joint independence (a term used by Krantz *et al.* 1971, p. 339) has also been called *conjoint independence*, *strong separability* (Strotz 1959), *independence among factors* (Fishburn 1970, sec. 4.1; Debreu 1960), and *coordinate independence* (Wakker 1989). It was called the *sure-thing principle* by Savage (1954) for a model in which the factors are states.

We mentioned intertemporal choice as a setting in which one may want to work with additive representations. Note how strong the joint independence condition can be in that setting. It implies that how Anna prefers to allocate consumption between periods 2 and 3 does not depend on how much she consumed in period 1. However, after a shopping and eating binge on Tuesday, Anna might well prefer to take it easy on Wednesday and defer some consumption until Thursday. Joint independence also implies that Anna's ranking of today's meal options does not depend on what she ate yesterday and that whether she wants to talk to her boyfriend today does not depend on whether they had a fight yesterday.

Yet additive utility is a mainstay of intertemporal models in economics. The assumption becomes a better approximation when outcomes are measured at a more aggregate level, as is typical in such models. Suppose, then, that the model does not include such details as meal options and social relationships, addressing instead one-dimensional aggregate consumption. Then joint independence in a two-period model is satisfied as long as Anna's preferences are monotone (though we will see that, in this case, the assumption is not enough to guarantee an additive utility representation). With more than two time periods, joint independence implies that Anna's desired allocation between periods 2 and 3 does not depend on period-1 consumption; this assumption is more plausible when a period is one year rather than one day.

1.8.4 A Representation Theorem

Joint independence is the main substantive assumption needed to obtain an additive representation. Here is one such representation theorem.

Theorem 3. Assume $J \geq 3$ and that the following hold:

1. X_j is connected for all $j = 1, \ldots, J$.
2. $\succcurlyeq$ is continuous.
3. $\succcurlyeq$ satisfies joint independence.
4. Each factor is essential: for all j, there exist $x_j, x'_j \in X_j$ and $x_{-j} \in X_{-j}$ such that $(x_j, x_{-j}) \succ (x'_j, x_{-j})$.

Then $\succcurlyeq$ has an additive representation that is continuous.

Proof: This result is similar to Debreu (1960, thm 3), except that the latter also assumes that each X_j is separable. Krantz *et al.* (1971, thm 13, sec. 6.11) provides an algebraic approach to such a representation; the topological assumptions are replaced by algebraic conditions on the preferences, and the continuity of the additive representation is not derived. Their theorem 14 then shows that our topological assumptions imply their algebraic conditions. Wakker (1988, thm 4.1) complements this by showing that, under the topological assumptions, any additive representation must be continuous. □

1.8.5 Revealed Trade-offs

Joint independence should not be confused with the weaker assumption of *single-factor independence* (also called *weak separability*): that preferences over each *single* factor do not depend on what is held fixed for the other factors. The fact that such independence must hold for preferences over multiple factors is the reason for the term "joint" in "joint independence".

For example, any monotone preference relation on $\mathbb{R}^J_+$ satisfies single-factor independence. Yet, even with the additional assumptions of Theorem 3, not all monotone preferences have additive representations. To illustrate this, we need only find a continuous and monotone function that cannot be transformed monotonically into an additive function; $U(x_1, x_2) = \min\{2x_1 + x_2, x_1 + 2x_2\}$ will do.

With just two factors, single-factor and joint independence are the same. This is why Theorem 3 assumed that there were at least three essential factors. To handle the two-factor case, one needs an assumption that is more explicit about the trade-offs Anna may make between factors. We now review such an axiomatization and its extension to J factors.

Let $j \in \{1, \ldots, J\}$, let $a_j, b_j, c_j, d_j \in X_j$, and let $x_{-j}, y_{-j} \in X_{-j}$.

$$\text{Suppose} \quad (x_{-j}, a_j) \preccurlyeq (y_{-j}, b_j) \qquad \text{but} \qquad (x_{-j}, c_j) \succcurlyeq (y_{-j}, d_j).$$

This reveals (one might say) that, keeping fixed the comparison x_{-j} versus y_{-j} on factors other than j, the strength of preference for c_j over d_j on factor j weakly exceeds the strength of preference for a_j over b_j on factor j. For non-additive utility, a switch in such revealed strength of preference could easily occur. That is, there might also be $w_{-j}, z_{-j} \in X_{-j}$ such that

$$(w_{-j}, a_j) \succcurlyeq (z_{-j}, b_j) \qquad \text{but} \qquad (w_{-j}, c_j) \prec (z_{-j}, d_j).$$

If there is such a reversal, then coordinate j *reveals contradictory trade-offs*. The assumption that *no coordinate reveals contradictory trade-offs* is also called *triple cancellation* for the case of $J = 2$ (Krantz *et al.* 1971) and *generalized triple cancellation* for $J \geq 2$ (Wakker 1989).

One way to put the "no contradictory trade-offs" condition into words is roughly as follows. When it holds, Anna could say, "When I choose among consumption paths for the week, one of my consideration is what I will do on Friday night. I prefer to go to a great movie on Friday rather than read a book. I also prefer to go to a dance rather than play poker. But the first comparison is more likely to swing my decision than the second one"—without any qualification about what is happening the rest of the week.

Our next theorem derives an additive representation from this condition.

Theorem 4. Assume $J \geq 2$ and that the following hold:

1. X_j is a connected for all $j = 1, \ldots, J$;
2. $\succcurlyeq$ is continuous;
3. no coordinate reveals contradictory trade-offs.

Then $\succcurlyeq$ has an additive representation that is continuous.

Proof: See Krantz *et al.* (1971, sec. 6.2) for the two-factor case. See Wakker (1989, thm III.6.6) for the J-factor case. □

1.9 Linear Preferences

1.9.1 Linear Utility

Suppose that X is a convex subset of $\mathbb{R}^n$. Then a subclass of utility functions that are additive across dimensions of the space are those that are linear.

Let $u \in \mathbb{R}^n$ be the vector representation of the linear utility function. That is, $U(x) = \sum_{i=1}^{n} u_i x_i = u \cdot x$, where $u \cdot x$ denotes the inner product. A typical indifference curve is

$$\{x \in X \mid u \cdot x = \bar{u}\}.$$

That is, the indifference curves are parallel hyperplanes (more specifically, the intersection of such parallel hyperplanes with X), and there is a direction u—perpendicular to the hyperplanes—that points in the direction of increasing utility.

These two conditions are also sufficient for the existence of a linear utility representation. We merely take the normal vector of one of the hyperplanes, pointing in the direction of increasing utility, and this defines a linear utility representation.

1.9.2 Linearity Axiom

What axioms on the preference relation give us linear and parallel indifference curves? Let's start with just the linearity of a particular indifference curve. Roughly speaking, we need that, for any two points on an indifference curve, any line passing through those points also lies on the indifference curve—at least where it intersects X. We can state this condition as follows.

Axiom L1. For all $0 < \alpha < 1$ and all $x, y \in X$,

$$\alpha x + (1 - \alpha)y \sim x \Leftrightarrow x \sim y. \tag{2}$$

Because Axiom L1 is an "if and only if" condition, it is stronger than convexity of the indifference curves and really does imply that they are linear subspaces. Take x and y on the same indifference curve and let $z \in X$ be any point on the line through x and y. If z lies between x and y, then we must have $z \sim x$ because of the $\Leftarrow$ direction of condition (2). If instead z lies beyond y, then y is a convex combination of x and z, and we have $z \sim x$ because of the $\Rightarrow$ direction of condition (2).

The second condition we want is for the translation of an indifference curve to be another indifference curve, so that the indifference curves are parallel. We might state this as follows.

Axiom L2. For all $x, y \in X$ and $z \in \mathbb{R}^n$ such that $x + z \in X$ and $y + z \in X$,

$$x + z \sim y + z \Leftrightarrow x \sim y.$$

However, there is a more succinct way to combine Axioms L1 and L2.

Axiom L3. For all $0 < \alpha < 1$ and all $x, y, z \in X$,

$$\alpha x + (1 - \alpha)z \sim \alpha y + (1 - \alpha)z \Leftrightarrow x \sim y. \tag{3}$$

Axioms L1 and L2 are equivalent to Axiom L3. By letting $z = y$ in condition (3), one obtains condition (2). Figure 1.1 illustrates Axiom L3 for z that is not colinear with x and y. The axiom then implies that $x \sim y$ if and only if $x' \sim y'$ and $x'' \sim y''$. One can see why this axiom implies that the indifference curves are parallel.

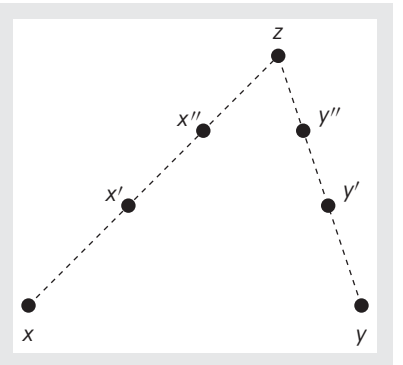

Fig. 1.1. An illustration of Axiom L3 or Axiom L: Axiom L3 states that $x \sim y$ if and only if $x' \sim y'$ and $x'' \sim y''$; Axiom L is the same but for $\succcurlyeq$.

However, Axiom L3 does not ensure that there is a common direction of increasing utility. To add this property, we can strengthen L3 by stating it in terms of $\succcurlyeq$ rather than $\sim$. We thus obtain the final statement of the linearity axiom.

Axiom L (Linearity) For all $0 < \alpha < 1$ and all $x, y, z \in X$,

$$\alpha x + (1 - \alpha)z \succcurlyeq \alpha y + (1 - \alpha)z \Leftrightarrow x \succcurlyeq y.$$

1.9.3 Continuity and the Representation Theorem

We have controlled for everything but the dimensionality of the indifference curves. It is fine for X to be one huge indifference curve; we can use the trivially linear utility function that maps all of X to 0. Otherwise, we need the dimensionality of the indifference curves to be one less than the dimensionality of X, so that each curve is the intersection of X and a hyperplane.

For example, lexicographic preferences on $\mathbb{R}^2$ satisfy the linearity axiom, and each indifference curve is indeed a linear subspace: a point. However, lexicographic preferences do not have a linear utility representation because the dimensionality of the indifference curves is too small.

A continuity assumption suffices to get the dimensionality correct. Consider, for example, an indifference curve through an alternative x that is neither the best nor the worst element of X. If preferences are continuous, then the set of points worse than x is open, and the set of points better than x is open; hence their union cannot be a connected set. This means that the indifference curve through x must separate

these two sets, which in turn requires that the indifference curve's dimension be one less than the dimension of X.

Given the linearity axiom, continuity is guaranteed by the following Archimedean axiom (which would be weaker than continuity in the absence of linearity).

Axiom 3 (Archimedean) For all $x, y, z \in X$, if $x \succ y \succ z$ then there exist α, β in $(0, 1)$ such that

$$\alpha x + (1 - \alpha)z \succ y \succ \beta x + (1 - \beta)z.$$

We now have our representation theorem.

Theorem 5. If $\succcurlyeq$ satisfies the linearity and Archimedean axioms, then $\succcurlyeq$ has a linear utility representation.

Proof: As we will see in Section 1.13, this proposition is similar to the expected utility theorem of von Neumann and Morgenstern (1944). However, this form is due to Jensen (1967, thm 8). It can be generalized to infinite-dimensional vector spaces. The simplest way is to develop a theory for an abstract generalization of convex sets, called "mixture spaces", as in Herstein and Milnor (1953). A comprehensive treatment can be found in Fishburn (1970, chs. 8 and 10). □

1.9.4 Linearity Implies Joint Independence

Suppose there are J factors, and each X_j is a convex subset of Euclidean space. Then $X = X_1 \times \ldots \times X_J$ is also a convex set. According to Theorem 5, if $\succcurlyeq$ satisfies the linearity and Archimedean axioms, then $\succcurlyeq$ has a linear utility representation, which can be written

$$U(x_1, \ldots, x_J) = \sum_{j=1}^{J} u_j \cdot x_j.$$

Since such a representation is additive across factors, $\succcurlyeq$ satisfies joint independence. Thus, with such a factor structure, a corollary to Theorem 5 is that the linearity and Archimedean axioms imply joint independence.

However, the Archimedean axiom is not needed for this implication. A direct proof of the stronger result, stated in the next proposition, clarifies the link between linearity and joint independence.

Proposition 3. *If $\succcurlyeq$ satisfies linearity, then $\succcurlyeq$ satisfies joint independence.*

Proof: Let $\{K, L\}$ be a partition of $\{1, \ldots, J\}$, let $a, b \in X_K$, and let $c, d \in X_L$. Assume $(a, c) \succcurlyeq (b, c)$. We show that $(a, d) \succcurlyeq (b, d)$.

Linearity and $(a, c) \succcurlyeq (b, c)$ imply

$$\frac{1}{2}(a, c) + \frac{1}{2}(a, d) \succcurlyeq \frac{1}{2}(b, c) + \frac{1}{2}(a, d). \tag{4}$$

By simple algebra, we can swap the terms a and b on the right-hand side of (4), so that this equation becomes

$$\frac{1}{2}(a, c) + \frac{1}{2}(a, d) \succcurlyeq \frac{1}{2}(a, c) + \frac{1}{2}(b, d). \tag{5}$$

Since (a, c) is a common term on both sides of the preference in (5), we obtain from a second application of linearity that $(a, d) \succcurlyeq (b, d)$. □

1.10 Cardinal Uniformity across Factors

1.10.1 Ordinal versus Cardinal Uniformity

Return to the setting of Section 1.8, in which there are J factors and $X = X_1 \times \ldots \times X_J$. There we found conditions under which Anna's preferences would have an additive representation

$$U(x_1, \ldots, x_J) = \sum_{j=1}^{J} u_j(x_j).$$

Suppose that each factor has the same set of possible alternatives: $X_j = \hat{X}$ for $j = 1, \ldots, J$. It is then meaningful to consider the special case in which preferences over each factor are the same—i.e. in which u_j and u_k represent the same ordinal preferences for $j \neq k$.

To state this condition in terms of preferences rather than utility, define the preference relation on factor 1 (for example) as follows. Pick $x_{-1} \in \prod_{j=2}^{J} X_j$ and then define $x_1 \succcurlyeq_1 x_1'$ by $(x_1, x_{-1}) \succcurlyeq (x_1', x_{-1})$; single-factor independence implies that this preference ordering on X_1 is the same no matter which x_{-1} we use to define it. Then the preferences $\succcurlyeq$ satisfy *ordinal uniformity across factors* if $\succcurlyeq_j$ is the same for all j.

However, ordinal uniformity adds little tractability beyond that of a general additive representation. It would be more interesting that the within-factor utility functions u_j were mere positive affine transformations of each other. Then we could

find a common utility function $u\colon \hat{X} \to \mathbb{R}$ and weights $\delta_1, \ldots, \delta_J > 0$ such that

$$U(x_1, \ldots, x_J) = \sum_{j=1}^{J} \delta_j u(x_j).$$

Such a representation has *cardinal uniformity across factors*.

For example, suppose that the factors are different time periods. Then such a representation is what we mean by *time-independent preferences with discounting*. We might normalize the factors by dividing by δ_1; then the new δ_1 equals 1, and we interpret δ_j as the discount factor from period j to period 1.

Such time independence of the utility representation simplifies economic models and allows one to derive stronger results. What is the essence of cardinal uniformity that allows for this? *We can separate the relative importance of the factors from the within-factor preferences.* The weights $\{\delta_1, \ldots, \delta_J\}$ provide a meaningful measure of the importance of the factors. For example, in a multi-agent intertemporal general equilibrium model, the assumption that all consumers have the same discount factors (i.e. the same impatience) is a powerful restriction even if their within-period utility functions u are heterogeneous. Without time-independent preferences, we would not even be able to formulate such as assumption.

Ordinal uniformity of preferences does not imply cardinal uniformity of the utility representation. Cardinal uniformity requires additional restrictions, which we explore in the rest of this section.

1.10.2 Ranking of Factors

As the preceding intertemporal consumption example illustrates, one of the key features of cardinal uniformity is that there is a measure $\{\delta_1, \ldots, \delta_J\}$ of the relative importance of the different factors. In terms of preferences, we can define an ordinal version of this condition as follows.

Axiom 4 (Joint ranking of factors) Suppose that preferences satisfy ordinal uniformity, and let $\succcurlyeq^*$ be the common-across-factors ordering on $\hat{X}$. Let $x^1, x^2, x^3, x^4 \in \hat{X}$ be such that $x^1 \succ^* x^2$ and $x^3 \succ^* x^4$. Let $K, L \subset \{1, \ldots, J\}$, and let, for example, x^1_K be the element of X_K that equals x^1 on each coordinate. Then

$$(x^1_K, x^2_{K^c}) \succ (x^1_L, x^2_{L^c}) \Leftrightarrow (x^3_K, x^4_{K^c}) \succ (x^3_L, x^4_{L^c}). \tag{6}$$

Joint ranking of factors is the same as Savage's P4 axiom, which we will see in Section 1.11. It yields an unambiguous weak order over the subsets of factors when we interpret $(x^1_K, x^2_{K^c}) \succ (x^1_L, x^2_{L^c})$ in equation (6) as meaning that the set of factors K is more important than the set of factors L.

Ordinal uniformity across factors has no implications for the trade-offs across factors, whereas joint ranking of factors does. For example, consider an intertemporal consumption example with a single good and monotone preferences. Ordinal uniformity holds merely because preferences are monotone. Joint ranking of factors implies additional restrictions, such as the following:

1. If Anna prefers the consumption path (100, 98, 98) over the consumption path (98, 100, 98), then she must also prefer (40, 30, 30) over (30, 40, 30). This captures the idea that Anna is impatient—not a normative assumption but one that is, at least, easy to interpret and empirically motivated.
2. The word "joint" (in "joint ranking of factors") applies because K and L can be arbitrary subsets of factors. Thus, suppose that Anna prefers the consumption path (100, 98, 98, 100, 98) over (98, 100, 100, 98, 98), which reflects fairly complicated trade-offs between the periods. Then she must also prefer (40, 30, 30, 40, 30) over (30, 40, 40, 30, 30). This goes well beyond impatience.

However, joint ranking of factors (with also joint independence and ordinal uniformity) is not yet sufficient for the existence of a cardinally uniform representation unless the choice set is rich enough for the axioms to impose enough restrictions. The problem is analogous to the way in which joint independence is sufficient for an additive representation when there are three factors but not when there are only two. We explore three ways to resolve this problem.

1.10.3 Enriching the Set of Factors

One approach is to enrich the set of factors so that it is infinite and perfectly divisible.

The first steps are to extend the definition of an additive representation to an infinite set of factors and then extend Theorem 3 to axiomatize the existence of such a representation. For this we refer the reader to Wakker and Zank (1999, thm 12). Joint independence remains the main behavioral assumption for additivity. Some additional technical assumptions are needed to handle the non-atomicity of the set of factors.

For example, suppose the set of factors is an interval $[0, 1]$ of time. A consumption path is a function $\tilde{x}\colon [0, 1] \to \hat{X}$. Joint independence and additional technical assumptions might yield an additive representation of the form

$$U(\tilde{x}) = \int_0^1 u(\tilde{x}(j),\, j)\, dj;$$

here, for each j, $u(\cdot, j)\colon \hat{X} \to \mathbb{R}$ is the within-factor utility for factor j. In this example, integration is with respect to the Lebesgue measure, so (a) any set of factors of Lebesgue measure 0 is negligible, in the sense that changing a consumption

plan on such a set does not change utility, and (b) the set of factors is non-atomic. Such features hold generally in the representation of Wakker and Zank (1999) and are important for what follows. Absent non-atomicity, the set of factors may not be rich enough for joint ranking of factors to imply cardinal uniformity. However, we repeat that these are essentially restrictions on the setting and not on the behavior of the decision-maker.

We now show that, in this example, joint ranking of factors is sufficient to go from an additive representation to one that is also cardinally uniform. To keep the illustration simple, we assume that $\hat{X}$ contains only the three elements a, b, c. The main step is the following lemma.

Lemma 1. *Assume ordinal uniformity and joint ranking of factors, and assume that* $a \succ^* b \succ^* c$. *Then the ratio*

$$\frac{u(b,\, j) - u(c,\, j)}{u(a,\, j) - u(c,\, j)} \tag{7}$$

is the same for a.e. j.

Before proving this lemma, observe that it allows us to obtain a cardinally uniform representation. Define $v(a) = 1$ and $v(c) = 0$, and let $v(b)$ be the common value of the ratio (7). Define also $\pi(j) = u(a,\, j) - u(c,\, j)$. Then

$$V(\tilde{x}) \equiv \int_0^1 \pi(j)\, v(\tilde{x}(j))\, dj$$

equals $U - \int_0^1 u(c,\, j)\, dj$. Thus, V is an additive representation with cardinal uniformity.

Proof of Lemma 1. We prove a contrapositive: Given ordinal uniformity and an additive representation, if the ratio (7) is not the same for almost every j, then joint ranking of factors does not hold.

Suppose (7) varies on a set of positive measure. Then there exist $\rho > 0$ and disjoint sets $K, L \subset [0, 1]$ such that

$$\frac{u(b,\, j) - u(c,\, j)}{u(a,\, j) - u(c,\, j)} > \rho \text{ for } j \in K \quad \text{and} \quad \frac{u(b,\, j) - u(c,\, j)}{u(a,\, j) - u(c,\, j)} < \rho \text{ for } j \in L. \tag{8}$$

Given non-atomicity, we can find $A \subset K$ and $B \subset L$ of strictly positive measure such that

$$\int_A (u(a,\, j) - u(c,\, j))\, dj = \int_B (u(a,\, j) - u(c,\, j))\, dj. \tag{9}$$

Let $\tilde{x}^1$ be equal to a on A and to c elsewhere; let $\tilde{x}^2$ be equal to a on B and c elsewhere. Observe that

$$U(\tilde{x}^1) - U(\tilde{x}^2) = \int_A (u(a,\, j) - u(c,\, j))\, dj - \int_B (u(a,\, j) - u(c,\, j))\, dj = 0. \tag{10}$$

That is, $\tilde{x}^1 \sim \tilde{x}^2$.

Joint ranking of factors implies that this same indifference should hold if $\tilde{y}^1$ and $\tilde{y}^2$ are defined accordingly except with a replaced by b. However, the two inequalities

$$\int_A (u(b, j) - u(c, j))\, dj > \rho \int_A (u(a, j) - u(c, j))\, dj,$$
$$\int_B (u(b, j) - u(c, j))\, dj < \rho \int_B (u(a, j) - u(c, j))\, dj$$

(which follow from the two inequalities in equation (8)), together with equality (9), imply that

$$\int_A (u(b, j) - u(c, j))\, dj > \int_B (u(b, j) - u(c, j))\, dj.$$

Therefore, as an analog to equation (10), we have $U(\tilde{y}^1) > U(\tilde{y}^2)$. Hence, joint ranking of factors does not hold. □

1.10.4 A Factor-Independent Version of No Contradictory Trade-offs

The second approach is to assume that $\hat{X}$ is connected and then use, as a starting point, the axiom that no coordinate reveals contradictory trade-offs. We capture state independence by modifying the axiom so that it holds even when we permute the indices. That is, if

$$(x_{-j}, a) \preccurlyeq (y_{-j}, b) \qquad \text{but} \qquad (x_{-j}, c) \succcurlyeq (y_{-j}, d),$$

then it cannot be the case that

$$(w_{-k}, a) \succcurlyeq (z_{-k}, b) \qquad \text{but} \qquad (w_{-k}, c) \prec (z_{-k}, d).$$

(We are being loose in the notation: the idea is that a, b, c, d show up in coordinate j in the first equation but in coordinate k in the second equation.) The resulting axiom is called *cardinal coordinate independence*, or *no contradictory trade-offs*, in Wakker (1989).

Proposition 4 (Wakker 1989). *Suppose that $\hat{X}$ is connected and that all factors are essential. Assume $\succcurlyeq$ is continuous and satisfies cardinal coordinate independence. Then $\succcurlyeq$ has a continuous additive representation that is cardinally uniform across factors.*

Cardinal coordinate independence subsumes joint independence, ordinal uniformity across factors, and joint ranking of factors. It captures the idea that the relative strength of trade-offs is uniform across factors.

1.10.5 The Linear Case

The third approach is to consider linear preferences. Suppose that $\hat{X}$ is a convex subset of Euclidean space and that we obtain an additive representation in which each u_j is linear. Recall that if u_1 and u_2 (for example) are two linear representations of the same ordinal preferences on $\hat{X}$, then u_2 is a positive affine transformation of u_1. Thus, linearity combined with the ordinal uniformity across factors gives us cardinal uniformity across factors!

We summarize this in a single representation theorem. If $\hat{X}$ is convex, then $X = \hat{X}^J$ is also convex and so we can impose the linearity and Archimedean axioms directly on $\succcurlyeq$. Linearity implies additivity; the axiom of joint independence is thus redundant. We merely need to add ordinal uniformity.

Proposition 5. *Suppose $\succcurlyeq$ satisfies the linearity, Archimedean, and ordinal uniformity across factors axioms. Then $\succcurlyeq$ has a utility representation of the form*

$$U(x_1, \ldots, x_J) = \sum_{j=1}^{J} \delta_j u(x_j),$$

where $u: \hat{X} \to \mathbb{R}$ is linear and continuous.

Proof: This is a consequence of Theorem 5 and the observation that ordinal uniformity implies cardinal uniformity for a linear representation. In Section 1.15, we will see that this is the expected utility theorem of Anscombe and Aumann (1963). □

1.11 States of Nature and Acts

We now reach (finally!) decisions under uncertainty. The first question is basic: How do we model uncertainty? We start with the states-of-the-world formulation and then move to the more reduced-form lotteries formulation.

1.11.1 Setup

In a static decision problem under uncertainty, Anna is uncertain about the outcomes of her actions. A formal model should make a careful distinction between those aspects of the world

1. that she controls, which we call an *action* and denote by $a \in A$;
2. that she is uncertain about and cannot influence, which we call a *state* and denote by $s \in S$;
3. that she cares directly about, which we call an *outcome* and denote by $z \in Z$.

These aspects are linked by a function F: $A \times S \to Z$ that summarizes how the outcome is determined by her actions (which she controls) and the state (which she cannot influence).

1.11.2 From Large to Small Worlds

Savage (1954, p. 9) describes a state of the world as "a description of the world, leaving no relevant aspect undescribed", but in chapter 5 he goes on to explain that an actual model of decisions (whether constructed by a decision theorist or contemplated by Anna) could not contain such detail. He then makes the following link between the abstract "large worlds", whose specification can omit no details, and the "small worlds" that appear in our models.

Let S^* be the set of "large worlds" or states as envisioned by Savage. A subset $E \subset S^*$ is called an *event*. To construct a model of a particular decision problem, we replace S^* by a partition $\{E_1, \ldots, E_n\}$ of S^* that is

- *fine* enough that, for every action, states in the same event lead to the same outcome,
- but otherwise as *coarse* as possible.

Each state s_j in the model corresponds to an event E_j in the partition.

Consider, for example, a portfolio choice problem in which Anna will invest a fixed amount of money in securities, hold the securities for one year, and then cash out the securities. The actions are the different portfolios that Anna could choose. Suppose that Anna cares only about the monetary payoff of her portfolio; such payoffs are the outcomes. Then a state could represent the values of all the securities at the time Anna cashes in her portfolio. Each "state" in the model is really an event. For instance, distinct large-world states within an event could differ in terms of the outcome of elections, the growth of the plants in Anna's garden, and other factors that do not affect portfolio returns.

1.11.3 From Actions to Acts

The decision-maker chooses an action $a \in A$. For example, Anna chooses a portfolio. Given action a, each state s leads to outcome $F(a, s)$. This defines a map $s \mapsto F(a, s)$ from S to Z. We call such a map an *act*. The act $s \mapsto F(a, s)$ is the one induced by the action a.

A first assumption is that, when choosing among actions, Anna compares the acts that they induce. This is plausible as long as we have defined outcomes carefully enough that they include any aspects of the action that Anna may care about (see Section 1.4). This leads to a reduced-form model in which Anna chooses among acts. Let's go a step further. Assume that Anna can contemplate hypothetical choices among any acts, even those that are not induced by any action.

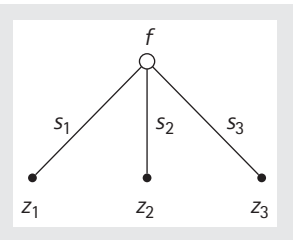

Fig. 1.2. Visualization of an act.

(Our distinction between actions and acts abuses vocabulary but is useful. When Savage coined the term "act", he was thinking of it as synonymous with "action". However, he jumped immediately to the reduced form and defined an act to be a mapping between S and Z. It is useful to retain the notion of the true action that Anna controls so that the model can more naturally describe actual decision problems.)

We can relate the resulting model to the ones developed earlier in the chapter as follows:

1. The set X of alternatives from which Anna chooses is the set of acts.
2. This set has a product structure in which the set of factors is the set S of states. Whereas we denoted an alternative by $x = (x_1, \ldots, x_J)$ in the factor model, here we use the equivalent functional notation $f\colon S \to Z$.
3. In each state, the set of outcomes is the same: Z. In the factor decomposition, this corresponds to the special case in which $X_j = \hat{X} = Z$ for each factor.

Although mathematically an act is just an element of a product set, it is appealing to visualize an act (when there are finitely many states) as in Figure 1.2, because it illustrates the unfolding of uncertainty. At the moment of making a choice, Anna stands in a position at or before the root node. Then one of the three states is realized—that is, one and only one of the branches from the root node occurs, leading to the consumption bundle (or, more generally, the outcome) $z = f(s)$.

1.12 Sure-Thing Principle and Additivity

Since X has a product structure, we can look for an additive representation of Anna's preferences $\succcurlyeq$.

1.12.1 Joint Independence

In this setting, the assumption of joint independence was called the *sure-thing principle* by Savage and has also been called the *independence axiom* (although the latter term has been used for related axioms in other settings). We pointed out that joint independence was a strong assumption when the factors are consumption of different goods or consumption in different time periods: one would never insinuate in those settings that joint independence was a normative axiom as opposed to a mere simplifying assumption. However, in this decision problem under uncertainty, the sure-thing principle is a plausible normative assumption.

With state-contingent consumption, only one of the states will ever be realized. It seems natural to most people, barring any mistakes Anna might make, that deciding how to spend the money Anna might earn if she wins a bet does not depend on what she might do if she loses the bet. We argue this more fully in the following section.

1.12.2 Dynamic Consistency

The notion of dynamic consistency can help us understand the sure-thing principle and judge whether it is a compelling normative axiom. Because our goal is not to develop a full theory of dynamic consistency for its own sake, and because such a framework is notationally intensive when set out formally, we keep the discussion informal.

In Section 1.3, we stated that (a) our model of static choice also applied to choosing a once-and-for-all plan in a temporal setting, but that (b) this chapter was not about making decisions at different decision epochs or the link between such decisions. Consider, however, such a possibility.

We are developing a normative model. Hence, though it would be difficult for her to do so, imagine that Anna can envision all the opportunities in which she will make decisions, that any uncertainty about available actions and outcomes of actions is reflected in her model of a set of states of the world, and that Anna understands how uncertainty is resolved over time as she learns new information. (All this would be modeled formally as a decision tree, which means an extensive game form in which there are moves by nature, but otherwise Anna is the only player.) A plan specifies the actions Anna would choose at each decision epoch (i.e. a plan is the same as a strategy in an extensive game form). Anna can then envision all possible plans and choose among them.

Suppose she then begins executing her preferred plan. At each decision epoch, she could reevaluate the plan—that is, consider all continuation plans on the subtree (subgame) that follows and choose among them. Dynamic consistency means that, as long as she has followed the *ex ante* preferred plan so far, Anna never wishes to deviate from the plan.

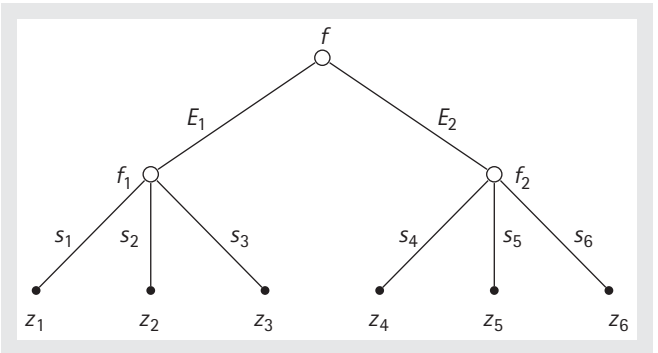

Fig. 1.3. An act that results from making decisions after observing $\{E_1, E_2\}$.

This paints an unrealistic picture of the way people actually make decisions. We do not set up a full model and choose among all possible plans. Even in an easily codified situation such as a chess game to be played during an afternoon, the set of possibilities is so large that it overwhelms our ability to envision and choose among them. Thus, we always have partially formulated plans; we constantly extend our partial model further into time; we review our available options; and we revise our decisions.

This is all due to the sheer complexity of devising a complete dynamic plan. However, it is fairly well accepted that dynamic consistency is a reasonable normative axiom—something Anna would satisfy if she were infinitely smart and had unlimited computational power.

Consider a simple situation in which there are six states, $\{s_1, s_2, s_3, s_4, s_5, s_6\}$, and in which Anna learns whether the state is in $E_1 \equiv \{s_1, s_2, s_3\}$ or in $E_2 \equiv \{s_4, s_5, s_6\}$ before choosing an action. She formulates a plan of what action to choose depending on whether she observes E_1 or E_2. Each such plan leads to an act.

For a given act f, we can represent the unfolding of information as in Figure 1.3. Observe that this is not a full decision tree; it depicts a single act and does not show the actions available at each decision epoch. Only nature's moves from the decision tree are shown. Behind the scenes we understand that different decisions lead to different acts, which means that Anna's choices determine which outcomes are at the leaves. The partial acts f_1 and f_2 specify the outcomes for the states in E_1 and E_2, respectively. We can write $f = (f_1, f_2)$.

If Anna is dynamically consistent, then if she observed E_1, she will want to stick to her plan. In other words, when choosing among acts defined on the states in E_1 after observing that event, she will still choose f_1 among the available acts. It is a fairly compelling assumption that, when she chooses among continuation plans having observed E_1, her preferences do not depend on what she thinks she might have done had E_2 been realized instead.

No theorem states that an infinitely smart person capable of such complex planning could not then be influenced in her decisions at each decision epoch by the plans she had made if other contingencies had been realized. "Fairly compelling assumption" simply means that these authors (at least) imagine that if we *were* so smart, then this is how we would make decisions.

This assumption, combined with dynamic consistency, implies that if $(f_1, f_2) \succcurlyeq (f_1', f_2)$, then also $(f_1, f_2') \succcurlyeq (f_1', f_2')$ for any other partial act f_2' defined on the states in E_2. More generally, it implies that preferences over acts satisfy the sure-thing principle.

1.12.3 Expected Utility Representation with State-Dependent Utility

Assume now that the sure-thing principle holds (along with other technical assumptions, such as those in Theorem 3), so that Anna's preferences over acts have an additive representation $\sum_{j=1}^{J} u_j(x_j)$, which we may write as

$$U(f) = \sum_{s \in S} u_s(f(s)).$$

Let $\pi\colon S \to \mathbb{R}$ be any probability measure on S such that $\pi(s) > 0$ for all $s \in S$. For each $s \in S$, define $v_s\colon Z \to \mathbb{R}$ by $v_s(z) = u_s(z)/\pi(s)$. Then

$$U(f) = \sum_{s \in S} \pi(s)\, v_s(f(s)). \tag{11}$$

Written this way, it it tempting to interpret U as an expected utility representation. Anna's utility in state s is $v_s(f(s))$ and, given the probability measure π on states, $U(f)$ is the expected value of Anna's utility. However, since the probability measure was chosen arbitrarily, the probabilities or beliefs used in this representation are indeterminate. Therefore, we cannot claim to have uncovered Anna's beliefs from her preferences.

The problem arises because the utility functions u_s in the additive representation have no link to each other. Utility is therefore said to be *state-dependent*.

1.12.4 Expected Utility Representation with State-Independent Utility

On the other hand, suppose that the representation satisfies cardinal uniformity across states and can thus be written as

$$U(f) = \sum_{s \in S} \delta_s u(f(s)),$$

where $\delta_s > 0$ for all s and $u\colon Z \to \mathbb{R}$. We then say that the utility is *state-independent*. We can normalize the weights $\{\delta_s \mid s \in S\}$ so that they sum to 1 by setting $\pi(s) = \delta_s / (\sum_{s' \in S} \delta_{s'})$; this is equivalent to multiplying the utility representation by $1/(\sum_{s' \in S} \delta_{s'})$. Then we have the representation

$$V(f) = \sum_{s \in S} \pi(s)\, u(f(s)). \tag{12}$$

If we take another additive representation, we know that it is a positive affine transformation of V. When we retain the restriction that the weights π sum to 1, the multiplicative and additive constants must end up in the function u. Thus, any additive representation of the preferences can be written in the form of (12) and will always end up with the same weighting function π. It then becomes more plausible to refer to these weights as probabilities or beliefs because they are uniquely identified.

1.12.5 Is Identification of Beliefs Important?

The answer is both "no" and "yes", as we now explain.

1.12.5.1 *Identification is a leap of faith*

Although the weights π in representation (12) are uniquely defined, their interpretation as beliefs remains a leap of faith. We cannot fully disentangle—from preferences over acts—whether Anna cares about what happens in a state because she considers the state likely or because the state affects her feelings. This is a general point, but we can give a concrete technological example in which preferences have the representation in equation (12) yet the weights $\pi(s)$ do not represent beliefs. Let the elements of Z be inputs in a single-output production process that is affected by state-contingent shocks such that output in state s, given inputs $z \in Z$, is $\theta_s u(z)$ (i.e. the production function in state s is $z \mapsto \theta_s u(z)$). An act is a state-contingent input plan $f\colon S \to Z$. Anna's beliefs are given by $\hat{\pi}$, and she is a risk-neutral expected utility maximizer with respect to output. Then her preferences over state-contingent input plans are given by

$$V(f) = \sum_{s \in S} \hat{\pi}(s)\, \theta_s\, u(f(z)).$$

The weights $\hat{\pi}(s)\, \theta_s$, written as $\pi(s)$ in equation (12), confound probabilities and productivity shocks. This is an argument that Karni (1985) gives in favor of studying state-dependent preferences and utility as opposed to inferring beliefs from a normalization of weights that is not derived from preferences or choice behavior.

1.12.5.2 *Identification of beliefs is not needed for Bayesian decision-making*

Are we concerned merely that Anna act as if she were probabilistically sophisticated and maximized expected utility, so that we can apply the machinery of Bayesian statistics to Anna's dynamic decision-making? Or rather, is our objective to uniquely identify her beliefs?

The latter might be useful if we wanted to measure beliefs from empirically observed choices in one decision problem in order to draw conclusions about how Anna would act with respect to another decision problem. Otherwise, the former is typically all we need, and state-dependent preferences are sufficient.

We can pick an additive representation of the form (11) with any weights π. Suppose that Anna faces a dynamic decision problem in which she can revise her choices at various decision nodes after learning some information (represented by a partition of the set of states). Given dynamic consistency, she will make the same decisions whether she makes a plan that she must adhere to or instead revises her decisions conditional on her information at each decision node. Furthermore, in the latter case her preferences over continuation plans will be given by expected utility maximization with the same state-dependent utilities and with weights (beliefs) that are revised by Bayesian updating. This may allow the analyst to solve her problem by backward induction (dynamic programming or recursion), thereby decomposing a complicated optimization problem into multiple simpler problems.

1.12.5.3 *Yet state independence is a powerful restriction*

The real power of state-independent utility comes from the structure and restrictions that this imposes on preferences, particularly in equilibrium models with multiple decision-makers. We already discussed this in the context of an intertemporal model with cardinally uniform utility. Let's revisit this point in the context of decision-making under uncertainty.

With state-independent utility, we can separate the relative probabilities of the states from the preferences over outcomes. For example, when the outcomes are money, we can separate beliefs from risk preferences. This is particularly powerful in a multi-person model, because we can then give substance to the assumption that all decision-makers have the same beliefs. Consider a general equilibrium model of trade in state-contingent transactions, such as insurance or financial securities. Suppose that all traders have state-independent utility with the same beliefs but heterogeneous utilities over money. If the traders' utilities are strictly concave (they are risk-averse) and if the total amount of the good that is available is state-independent (no aggregate uncertainty), then in any Pareto efficient allocation each trader's consumption is state-independent (each trader bears no risk).

1.12.5.4 *State independence is without loss of generality (more or less)*

It can be argued that state independence is without loss of generality: if it is violated, one can redefine outcomes to ensure that the description of an outcome includes everything Anna cares about—even things that are part of the description of the state. However, when this is done, some acts are clearly hypothetical.

Perhaps the two states are "Anna's son has a heart attack" and "Anna's son's heart is just fine". What Anna controls is whether her son has heart surgery. Clearly her preferences for heart surgery depend on whether or not her son has a heart attack. However, we can define an outcome so that it is specified both by whether her son has a heart attack and by whether he undergoes surgery. In order to maintain the assumption that the set of acts is the set of all functions from states to outcomes, Anna must be able to contemplate and express preferences among such hypothetical acts as the one in which her son has a heart attack and gets heart surgery in both states, including the state in which he does not have a heart attack!

Furthermore, when decision under uncertainty is applied to risk and risk sharing, the modeler assumes that preferences over *money* are state-independent. This is a strong assumption even if preferences were state-independent for some appropriately redefined set of outcomes.

1.13 Lotteries

1.13.1 From Subjective to Objective Uncertainty

We postpone until Section 1.14 a discussion of the axioms that capture state independence of preferences and yield a state-independent representation $U(f) = \sum_{s \in S} \pi(s)\, u(f(s))$. In the meantime, we consider how state independence combined with objective uncertainty allows for a reduced-form model in which choices among state-dependent outcomes (acts) is reduced to choices among probability measures on outcomes (lotteries). We then axiomatize expected utility for such a model.

One implication of state-independent expected utility is that preferences depend only on the probability measures over outcomes that are induced by the acts. That is, think of an act f as a random object whose distribution is the induced probability measure p on Z. Assume that S and Z are finite, so that this distribution is defined by $p(z) = \pi\{s \in S \mid f(s) = z\}$. We can then rewrite $U(f) = \sum_{s \in S} \pi(s)\, u(f(s))$ as $\sum_{z \in Z} p(z)\, u(z)$. In particular, Anna is indifferent between any two acts that have the same induced distribution over outcomes.

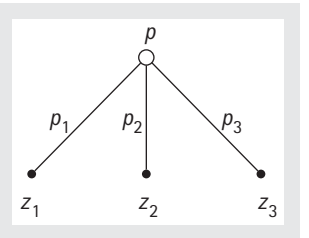

Fig. 1.4. A lottery.

Let us now take as our *starting point* that Anna's decision problem reduces to choosing among probability measures over outcomes—without a presumption of having identified an expected utility representation in the full model. We then state axioms within this reduced form that lead to an expected utility representation.

For this to be an empirical exercise (i.e. in order to be able to elicit preferences or test the theory), the probability measures over outcomes must be observable. This means that the probabilities are generated in an objective way, such as by flipping a coin or spinning a roulette wheel. Therefore, this model is typically referred to as one of objective uncertainty. The other reason to think of this as a model of objective uncertainty is that we will need data on how the decision-maker would rank all possible distributions over Z. This is plausible only if we can generate probabilities using randomization devices.

Thus, the set of alternatives in Anna's choice problem is the set of probability measures defined over the set Z of outcomes. To avoid the mathematics of measure theory and abstract probability theory, we continue to assume that Z is finite, letting n be the number of elements. We call each probability measure on Z a *lottery*.

Let P be the set of lotteries. Each lottery corresponds to a function $p\colon \to [0, 1]$ such that $\sum_{z\in Z} p(z) = 1$. Each $p \in P$ can equivalently be identified with the vector in $\mathbb{R}^n$ of probabilities of the n outcomes. The set P is called the *simplex* in $\mathbb{R}^n$; it is a compact convex set with $n - 1$ dimensions.

We can illustrate a lottery graphically as in Figure 1.4. The leaves correspond to the possible outcomes and the edges show the probability of each outcome. Figure 1.4 looks similar to the illustration of an act in Figure 1.2, but the two figures should not be confused. When Anna considers different acts, the states remain fixed in Figure 1.2 (as do their probabilities); what change are the outcomes. When Anna considers different lotteries, the outcomes remain fixed in Figure 1.4; what changes are the probabilities. This reduced-form model of lotteries has a flexibility with respect to possible probability measures over outcomes that would not be possible in the states model unless the set of states were uncountably infinite and beliefs were non-atomic.

By an *expected utility* representation of Anna's preferences $\succcurlyeq$ on P we mean one of the form

$$U(p) = \sum_{z \in Z} p(z)\, u(z),$$

where $u\colon Z \to \mathbb{R}$. Then $U(p)$ is the expected value of u given the probability measure p on Z. We call this a *Bernoulli representation* because Bernoulli (1738) posited such an expected utility as a resolution to the St. Petersburg paradox: that a decision-maker would prefer a finite amount of money to a gamble whose expected payoff was infinite. Bernoulli took the utility function $u\colon Z \to \mathbb{R}$ as a primitive and expected utility maximization as a hypothesis. His innovation was to allow for an arbitrary, even bounded, function $u\colon Z \to \mathbb{R}$ for lotteries over money rather than a linear function, thereby avoiding the straitjacket of expected value maximization—the state of the art in his day.

Expected utility did not receive much further attention until von Neumann and Morgenstern (1944) first axiomatized it (for use with mixed strategies in game theory). For this reason, the representation is also called a von Neumann–Morgenstern utility function. As we do here, von Neumann and Morgenstern took preferences over lotteries as a primitive and uncovered the expected utility representation from several axioms on those preferences.

1.13.2 Linearity of Preferences

Recall that P is a convex set, and recall from Section 1.9 that $\succcurlyeq$ admits a linear utility representation if it satisfies the linearity and Archimedean axioms. We proceed as follows.

1. We observe that a linear utility representation is the same as a Bernoulli representation.
2. We discuss the interpretation of the linearity and Archimedean axioms.

In this setting, linearity (Axiom L) is called the *independence axiom*.

So suppose we have a linear utility representation

$$U(p) = \sum_{z \in Z} u_z p_z$$

of Anna's preferences. We can write the vector $\{u_z \mid z \in Z\}$ of coefficients as a function $u\colon Z \to \mathbb{R}$ and use the functional form $p\colon Z \to [0, 1]$ of a lottery p. Then the linear utility representation can be written as

$$U(p) = \sum_{z \in Z} p(z)\, u(z), \tag{13}$$

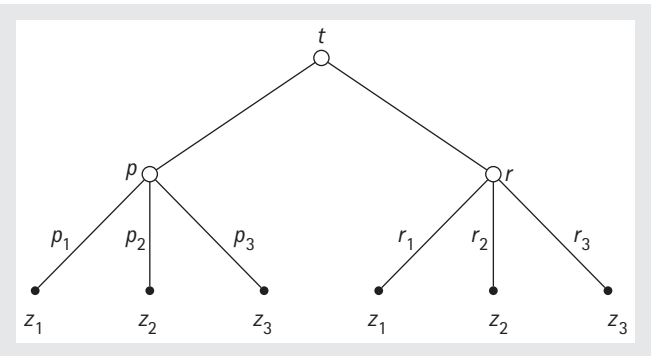

Fig. 1.5. A compound lottery.

that is, as a *Bernoulli representation*. Like any additive representation, this one is unique up to a positive affine transformation; such a transformation of U corresponds to an affine transformation of u. All this is summarized in our next theorem.

Theorem 6. If $\succcurlyeq$ satisfies the linearity (independence) and Archimedean axioms, then $\succcurlyeq$ has a Bernoulli representation.

Proof: This is an application of Theorem 5; as such, it is due to Jensen (1967, thm 8). Von Neumann and Morgenstern's representation theorem used a different set of axioms that implied but did not contain an explicit independence (linearity) axiom like our Axiom L. The role of the independence axiom, which we interpret further in what follows, was uncovered gradually by subsequent authors. See Fishburn and Wakker (1995) for a history of this development. □

1.13.3 Interpretation of the Axioms

The convex combinations that appear in the linearity and Archimedean axioms have a nice interpretation in our lotteries setting. Suppose the uncertainty by which outcomes are selected unfolds in two stages. In a first stage, there is a random draw to determine which lottery is faced in a second stage. With probability α, Anna faces lottery p in the second stage; with probability $1 - \alpha$ she faces lottery r. This is called a *compound lottery* and is illustrated in Figure 1.5.

Consider the overall lottery t that Anna faces *ex ante*, before any uncertainty unfolds. The probability of outcome z_1 (for example) is $t_1 = \alpha p_1 + (1 - \alpha)r_1$. As a vector, the lottery t is the convex combination $\alpha p + (1 - \alpha)r$ of p and r. Thus, we can interpret convex combinations of lotteries as compound lotteries.

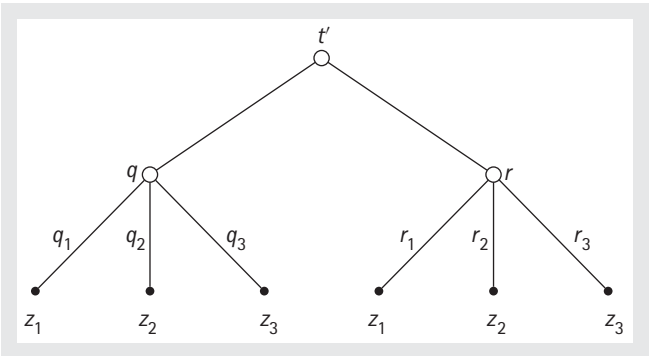

Fig. 1.6. Another compound lottery.

Consider this compound lottery and the one in Figure 1.6, recalling the discussion of dynamic consistency and the sure-thing principle from Section 1.12. Suppose Anna chooses t over t' and then, after learning that she faces lottery p in the second stage, is allowed to change her mind and choose lottery q instead. Dynamic consistency implies that she would not want to do so. Furthermore, analogous to our normative justification of the sure-thing principle, it is also natural that her choice between p and q at this stage would depend neither on which lottery she would otherwise have faced along the right branch of the first stage nor on the probability with which the left branch was reached. Together, these two observations imply that she would choose lottery t over t' if and only if she would choose lottery p over q. Mathematically, in terms of the preference ordering $\succcurlyeq$, this is Axiom L. It is called the *independence axiom* or *substitution axiom* in this setting, because the choices between t and t' are then independent of which lottery we substitute for r in Figure 1.6.

Thus, the justification for the independence (linearity) axiom in this lotteries model is the same as for the sure-thing principle (joint independence axiom) in the states model, but the two axioms are mathematically distinct because the two models define the objects of choice differently (lotteries vs. acts).

The Archimedean axiom has the following meaning. Suppose that Anna prefers lottery p over lottery q. Now consider the compound lottery t in Figure 1.6. Lottery r might be truly horrible. However, if the Archimedean axiom is satisfied, then, as long as the right branch of t occurs with sufficiently low probability, Anna still prefers lottery t over lottery q. This is illustrated by the risk of death that we all willingly choose throughout our lives. Death is certainly something "truly horrible"; however, every time we cross the street, we choose a lottery with small probability of death over the lottery we would face by remaining on the other side of the street.

1.13.4 Calibration of Utilities

The objective probabilities are used in this representation to calibrate the decision-maker's strength of preference over the outcomes. To illustrate how this is done, suppose Anna is considering various alternatives that lead to varying objectively measurable probabilities of the following outcomes:

e — Anna stays in her current employment;
m — Anna gets an MBA but then does not find a better job;
M — Anna gets an MBA and then finds a much better job.

We let $Z = \{e, m, M\}$ be the set of outcomes, and, since this is a reduced form, we view her choice among her actions as boiling down to the choice among the probabilities over Z that the actions induce. Furthermore, we suppose that she can contemplate choices among all probability measures on Z, and not merely those induced by one of her actions. We assume $M \succ e \succ m$, where (for example) $M \succ e$ means that she prefers getting M for sure to getting e for sure.

Anna's preference for e relative to m and M can be quantified as follows. We first set $u(M) = 1$ and $u(m) = 0$. We then let $u(e)$ be the unique probability for which she is indifferent between getting e for sure and the lottery that yields M with probability $u(e)$ and m with probability $1 - u(e)$—that is, for which $e \sim u(e)M + (1 - u(e))m$. The closer e is to M than to m in her strength of preference, the greater this probability $u(e)$ would have to be and, in our representation, the greater is the utility $u(e)$ of e.

The Archimedean axiom implies that such a probability $u(e)$ exists. The independence axiom then implies that the utility function $u\colon Z \to \mathbb{R}$ thus defined yields a Bernoulli representation of Anna's preferences. The actual proof of the representation theorem is an extension of this constructive proof to more general Z.

1.14 Subjective Expected Utility without Objective Probabilities

1.14.1 Overview

Let us return to the states and acts setting of Sections 1.11 and the state-dependent expected utility representation from Section 1.12. Recall the challenge—posed but not resolved—of finding a *state-independent* representation, so that

the probabilities would be uniquely identified and could be interpreted as beliefs revealed by the preferences over acts. This is called *subjective expected utility* (SEU).

One of the first derivations of subjective expected utility (involving the joint derivation of subjective probabilities to represent beliefs about the likelihood of events as well as the utility index over outcomes) appeared in a 1926 paper by Frank Ramsey, the English mathematician and philosopher. This article was published posthumously in Ramsey (1931) at about the same time as an independent but related derivation appeared in Italian by the statistician de Finetti (1931). The definitive axiomatization in a purely subjective uncertainty setting appeared in Leonard Savage's 1954 book *The Foundation of Statistics*.

In Section 1.12, we showed that the sure-thing principle implied additivity of the utility. We went on to say that SEU requires that the additive utility be cardinally uniform across states, but we stopped before showing how to obtain such a conclusion. Recall, further back, Section 1.10, where we tackled cardinal uniformity in the abstract factors setting. Axiomatizing cardinal uniformity was tricky, but we outlined three solutions. Each of those solutions corresponds to an approach taken in the literature on subjective expected utility.

1. Savage (1954) used an infinite and non-atomic state space as in Section 1.10.3. We develop this further in Section 1.14.2.
2. Wakker (1989) assumed a connected (hence infinite) set of outcomes and assumed cardinal coordinate independence, as we did in Section 1.10.4. Cardinal coordinate independence involves specific statements about how the decision-maker treats trade-offs across different states and assumes that such trade-offs are state-independent.
3. Anscombe and Aumann (1963) mixed subjective and objective uncertainty to obtain a linear representation, as in Section 1.10.5. We develop this in Section 1.15.

1.14.2 Savage

We give a heuristic presentation of the representation in Savage (1954). (In what follows, P*n* refers to Savage's numbering of his axioms.) Savage began by assuming that preferences are transitive and complete (P1: weak order) and satisfy joint independence (P2: sure-thing principle); this yields an additive or state-dependent representation. The substantive axioms that capture state independence are ordinal uniformity (P3: ordinal state independence) and joint ranking of factors (P4: qualitative probability).

As a normative axiom, P3 is really a statement about the ability of the modeler to define the set of outcomes so that they encompass everything that Anna cares about. Then, given any realization of the state, Anna's preferences over outcomes should be the same.

Because Savage works with an infinite state space in which any particular state is negligible, his version of P3 is a little different from ours, and he needs an additional related assumption. These are minor technical differences.

1. Savage's P3 states that Anna's preferences are the same conditional *on any nonnegligible event*, rather than on any state. With finitely many states, the two axioms are equivalent.
2. Savage adds an axiom (P7) that the preferences respect statewise dominance: given Anna's state-independent ordering $\succcurlyeq^*$ on Z, if f and g are such that $f(s) \succcurlyeq^* g(s)$ for all $s \in S$, then $f \succcurlyeq g$. With finitely many states, this condition is implied by the sure-thing principle and ordinal state independence.

Let us consider in more detail Savage's P4, which is our joint ranking of factors. We begin by restating this axiom using the terminology and notation of the preferences-over-acts setting.

Axiom 5 (Qualitative probability). Suppose that preferences satisfy ordinal state independence, and let $\succcurlyeq^*$ be the common-across-states ordering on Z. Let $A, B \subset S$ be two events. Suppose that $z^1 \succ^* z^2$ and $z^3 \succ^* z^4$. Let, for example, $(I_A z^1, I_{A^c} z^2)$ be the act that equals z^1 on event A and z^2 on its complement. Then

$$(I_A z^1, I_{A^c} z^2) \succ (I_B z^1, I_{B^c} z^2) \Leftrightarrow (I_A z^3, I_{A^c} z^4) \succ (I_B z^3, I_{B^c} z^4).$$

This axiom takes state independence one step further: it captures the idea that the decision-maker cares about the states only because they determine the likelihood of the various outcomes determined by acts. If preferences are state-independent, then the only reason why Anna would prefer $(I_A z^1, I_{A^c} z^2)$ to $(I_B z^1, I_{B^c} z^2)$ is because she considers event A to be more likely than event B. In such case, she must also prefer $(I_A z^3, I_{A^c} z^4)$ to $(I_B z^3, I_{B^c} z^4)$.

As explained in Section 1.10.3, ordinal state independence and qualitative probability impose enough restrictions to yield state-independent utility only if the choice set is rich enough—with one approach being to have a non-atomic set of factors or states. This is the substance of Savage's axiom P6 (continuity). The richer state space allows one to calibrate beliefs separately from payoffs over the outcomes.

1.15 Subjective Expected Utility with Objective Probabilities

1.15.1 Horse-Race/Roulette-Wheel Lotteries

Anscombe and Aumann (1963) avoided resorting to an infinite state space or axioms beyond joint independence and ordinal uniformity by combining (a) a lotteries framework with objective uncertainty and (b) a states framework with subjective uncertainty.

In their model, an act assigns to each state a lottery with objective probabilities. These two-stage acts are also called horse-race/roulette-wheel lotteries, but we continue to refer to them merely as acts and to the second-stage objective uncertainty as lotteries. Fix a finite set S of states and a finite set Z of outcomes. We let P be the set of lotteries on Z. An act is a function $f\colon S \to P$. Let H be the set of acts.

1.15.2 Linearity: Sure-Thing Principle and Independence Axiom

First notice that H, which is the product set P^S, is also a convex set and that the convex combination of two acts can be interpreted as imposing compound lotteries in the second (objective) stage of the unfolding of uncertainty. In other words, for any pair of acts f, g in H and any α in $[0, 1]$, $\alpha f + (1 - \alpha)g$ corresponds to the act h in H for which $h(s) = \alpha f(s) + (1 - \alpha)g(s)$, where $\alpha f(s) + (1 - \alpha)g(s)$ is the convex combination of lotteries $f(s)$ and $g(s)$.

In Section 1.9, we showed that $\succcurlyeq$ has a linear utility representation if $\succcurlyeq$ satisfies the linearity and Archimedean axioms. Let us consider the interpretation of such a utility representation and the interpretation of these axioms.

The dimensions of H are $S \times Z$, and a linear utility function on H can be written as

$$\sum_{s \in S} \sum_{z \in Z} u_{sz} p_{sz} = \sum_{s \in S} \sum_{z \in Z} u_{sz} \times f(s)(z). \tag{14}$$

On the left side, we have represented the element of H as a vector $p \in \mathbb{R}^{S \times Z}$; the probability of outcome z in state s is p_{sz}. On the right side, we have represented the element of H as an act $f\colon S \to P$; the probability of outcome z in state s is $f(s)(z)$. (That is, $f(s)$ is the probability measure or lottery in state s and $f(s)(z)$ is the probability assigned to z by that measure.) We used a "$\times$" on the right-hand side for simple multiplication to make clear that $f(s)(z)$ is a single scalar term. The order of summation in equation (14) is irrelevant.

For any probability measure π on S we can also write the linear utility function in the form

$$\sum_{s\in S}\pi(s)\sum_{z\in Z}f(s)(z)\times u_s(z). \tag{15}$$

We derived equation (15) from (14) by:

- dividing each coefficient u_{sz} by $\pi(s)$ and writing the result as $u_s(z)$; then
- reversing the order of multiplication so that $\sum_{z\in Z} f(s)(z)\times u_s(z)$ is recognized as the expected utility in state s—given that $f(s)$ is the probability measure on Z and $u_s\colon Z\to\mathbb{R}$ is the utility function on Z in state s.

Thus, (15) can be interpreted as the subjective expected value (over states S with subjective probability π) of the objective expected utility (over outcomes Z given objective probabilities $f(s)$ in state s). We call equation (15) a *state-dependent Anscombe–Aumann representation*. We thus have, as a corollary to Theorem 5 and this discussion, the following representation theorem.

Theorem 7. Assume that $\succcurlyeq$ satisfies the linearity and Archimedean axioms. Then $\succcurlyeq$ has a state-dependent Anscombe–Aumann representation.

The linearity axiom on $\succcurlyeq$ thus encompasses two independence conditions:

1. the *sure-thing principle* as applied to *subjective* uncertainty across different states (i.e. linearity implies joint independence over factors, as shown in Section 1.9.4);
2. the *independence axiom* as applied to *objective* uncertainty within each state (i.e. linearity of $\succcurlyeq$ implies linearity of the within-state preferences).

The normative arguments that justify these two axioms or principles, which we have already discussed extensively, also justify the linearity axiom in this Anscombe–Aumann framework.

1.15.3 State Independence

The probability measure π is still not uniquely identified because we have state-dependent utility. However, recall from Section 1.10.5 that the additional assumption of ordinal state independence is enough to obtain state-independent linear utility and thus to pin down the subjective beliefs. The overall representation becomes

$$U(f)=\sum_{s\in S}\pi(s)\sum_{z\in Z}f(s)(z)\times u(z). \tag{16}$$

We call equation (16) an *Anscombe–Aumann representation*.

We summarize this as follows.

Theorem 8. Assume that $\succcurlyeq$ satisfies the linearity, Archimedean, and ordinal state independence axioms. Then $\succcurlyeq$ has an Anscombe–Aumann representation.

Proof: This follows from Theorem 5 and the preceding discussion. It is also essentially Anscombe and Aumann (1963, thm 1), though their axiomatization is a bit different. □

1.15.4 Calibration of Beliefs

The simplicity of Theorem 8 and the fact that it is a mere application of linear utility masks the way in which beliefs and utilities are disentangled. We illustrate how such calibration happens using ideas that lurk in the proof of the theorem.

For example, consider a less reduced-form version of the story in Section 1.13.4. Anna chooses between two actions:

leave— leaving her current employment to undertake an MBA;
stay — staying put.

The relevant outcomes are the three enumerated in Section 1.13.4: (e) no MBA and staying in her current employment; (m) bearing the cost of an MBA without then finding a better job; and (M) bearing the cost of an MBA and then finding a better job.

The last element in the decision problem is the event E, the set of states in which Anna obtains the better job if she gets an MBA. We take this to be a state or elementary event in the small-worlds model; hence the set of states is $\{s_1, s_2\}$, where s_1 corresponds to event E and s_2 corresponds to event E^c. Therefore the two acts associated with the actions *leave* and *stay* are

$$\textit{leave}\,(s_1) = M, \quad \textit{leave}\,(s_2) = m;$$
$$\textit{stay}\,(s_1) = e, \quad \textit{stay}\,(s_2) = e.$$

Whether we have *leave* $\succcurlyeq$ *stay* or *stay* $\succcurlyeq$ *leave* seems to depend on two separate considerations: how good *Anna feels* the chances of obtaining a better job would have to be in order to make it worthwhile to leave her current employment; and how good *in her opinion* the chances of obtaining a better job actually are. What Anna does when she considers the first of these is quantify her *personal preference* for e relative to m and M. What she does when she considers the second is quantify her *personal judgment* concerning the relative strengths of the factors that favor and oppose certain events.

In order to calibrate these two considerations, we must assume that she can meaningfully compare *any* horse-race/lottery acts, not merely the acts *leave* and *stay* available to her in this problem. Thus, she must be able to express preferences

over hypothetical acts such as the act

$$g(s_1) = m, \quad g(s_2) = M,$$

(in the state s_2 where she would *not* find a good job if she got an MBA, she gets an MBA and finds a good job!) and the act that yields, in both states, a lottery with equal probability of the three outcomes.

We can first quantify the strength of Anna's personal preference for e relative to m and M by considering the constant acts (lotteries that are not state-dependent). That is, we abstract from the subjective uncertainty about the states and consider her preferences over objectively generated lotteries. This is the representation and calibration we covered in Section 1.13. We thereby let $u(m) = 0$ and $u(M) = 1$ and define $u(e)$ to be such that Anna is indifferent between e and the lottery $u(e)M + (1 - u(e))m$.

To quantify Anna's judgment concerning the likelihood of state s_1, we let $\pi(s_1)$ be the unique probability for which Anna is indifferent between the act *leave* and the act that leads, in every state, to the lottery with probability $\pi(s_1)$ on M and probability $1 - \pi(s_1)$ on m. The idea is that, given state-independent preferences, the state is simply a randomization device from Anna's point of view: if Anna is indifferent between these two acts, it is because the objective probability $\pi(s_1)$ is the same as Anna's subjective likelihood of state s_1.

1.16 Conclusion

Throughout this chapter we have emphasized the link between independence axioms in standard consumer theory, in expected utility theory for decision under objective uncertainty, and in expected utility theory for decision under subjective uncertainty. We contend that the independence axioms have considerable normative appeal in decision under uncertainty.

However, experimental and empirical evidence shows that behavior deviates systematically from these theories, implying that (not surprisingly) such normative theories make for only approximate descriptive models. Furthermore, many authors have disagreed with our claim that the independence axioms are normatively compelling.

There is now a vast literature that has developed generalizations, extensions, and alternatives to expected utility. We will not provide a guide to this literature; doing so would be beyond the scope of our chapter, whereas later chapters in this Handbook treat it extensively. However, as a transition to those chapters and as a further illustration of the content of the independence axioms, we outline some of the experimental violations.

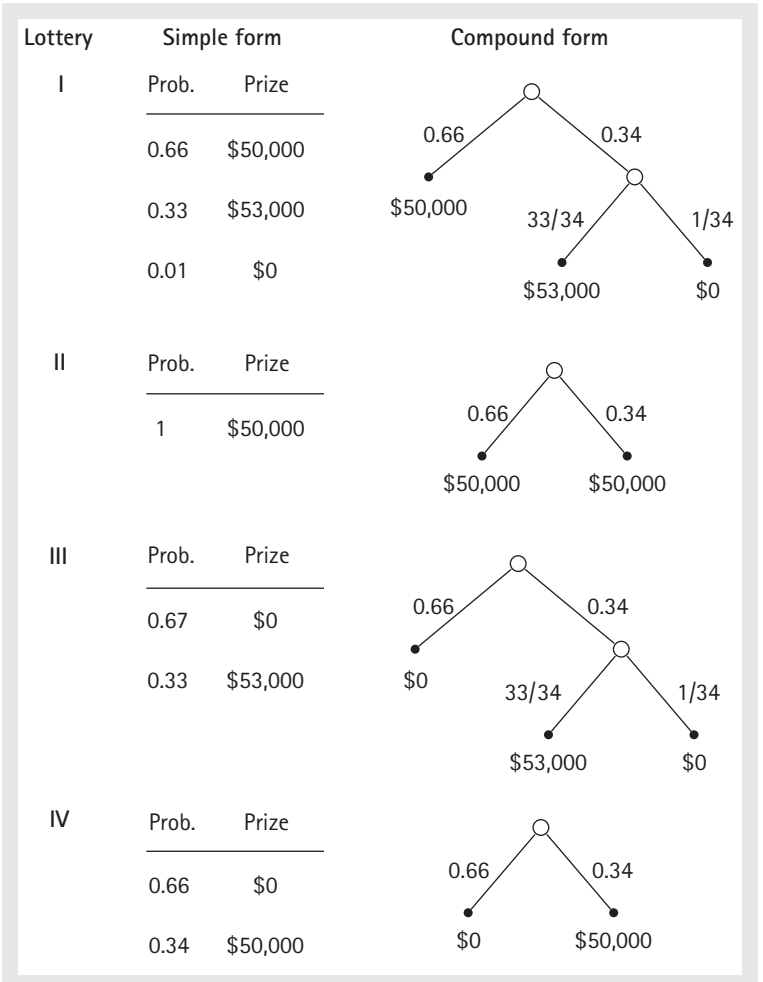

Fig. 1.7. Common consequence paradox (Allais paradox). The simple lotteries on the left are the reduced lotteries of the compound lotteries on the right. Preferences II $\succ$ I and III $\succ$ IV violate the independence axiom but are common for subjects in decision experiments.

One of the earliest and best-known tests of expected utility is the *common consequence paradox*, first proposed by Maurice Allais (1953). It is illustrated in Figure 1.7. Allais conjectured (and found) that most people would prefer lottery II to lottery I but would prefer lottery III to lottery IV (when presented as the simple lotteries on the left). By writing the simple lotteries as the compound lotteries on the right, we can see that such choices violate the independence axiom (Axiom L).

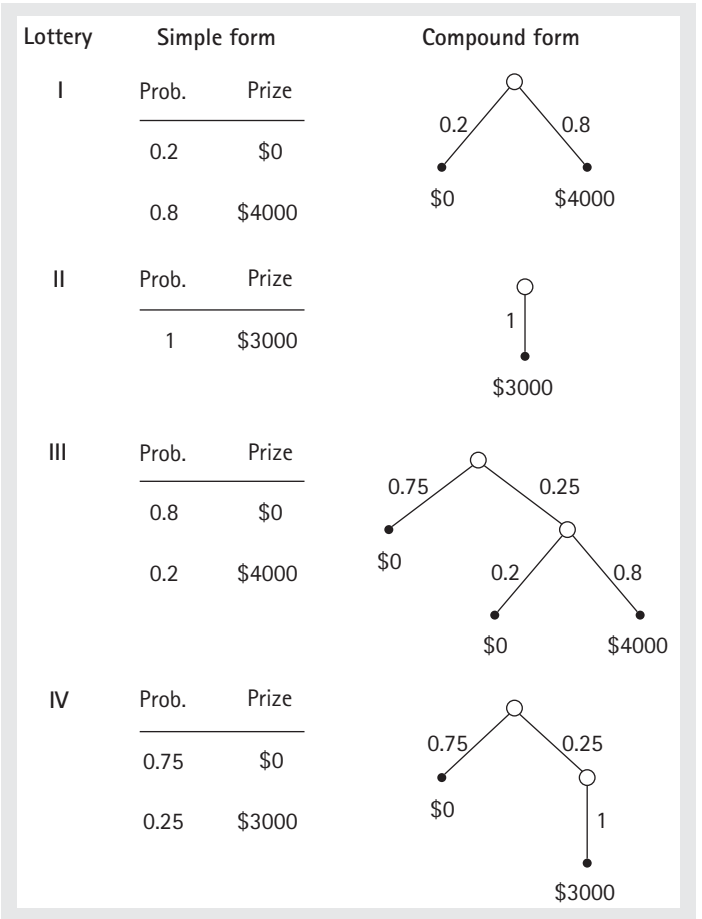

Fig. 1.8. Common ratio paradox. The simple lotteries on the left are the reduced lotteries of the compound lotteries on the right. Preferences II $\succ$ I and III $\succ$ IV violate the independence axiom but are common for subjects in decision experiments.

A closely related and frequently observed systematic violation of expected utility theory is the *common ratio paradox* (see Kahneman and Tversky 1979). This is illustrated in Figure 1.8. Again, the choices of II over I and III over IV are common but violate the independence axiom.

There has been debate about whether these violations are due to bounded rationality or whether the normative model needs adjustment, but there is certainly room for better descriptive models than the classic theory reviewed in this chapter (even if, for many applications, the classic theory has proved to be a suitably accurate approximation).

Systematic violations of expected utility theory—observed in choice problems such as these paradoxes—suggest the following: when altering a lottery by reducing the probability of receiving a given outcome, the portion of the probability we must place on a better outcome (with the remaining portion on a worse outcome) in order to keep the individual indifferent is not *independent* of the lottery with which we began. Yet the independence axiom implies that it is independent. Indeed, for any three outcomes H, M, L, with $H \succ M \succ L$, the trade-off for an expected utility maximizer is simply the constant slope of the indifference curves in the simplex of lotteries:

$$\frac{u(M) - u(L)}{u(H) - u(M)}.$$

When assessing the accumulated experimental evidence, Machina (1982) proposed that one could account for these observed systematic violations of expected utility by assuming that this trade-off is *increasing* the "higher up" (in preference terms) in the simplex is the lottery with which one starts. Geometrically, this corresponds to a "fanning out" of the indifference curves in the simplex. Many other patterns have been observed that depend on the size and sign of the payoffs.

In response, several versions of so-called non- or generalized expected utility have axiomatized and analyzed nonlinear representations of preferences over lotteries. These include, among others, rank-dependent expected utility of Quiggin (1982) and Yaari (1987), cumulative prospect theory of Tversky and Kahneman (1992) and Wakker and Tversky (1993), betweenness of Dekel (1986) and Chew (1989), and additive bilinear (regret) theories of Loomes and Sugden (1982) and Fishburn (1984).

Another famous experiment, whose results are inconsistent with *subjective* expected utility theory, is the Ellsberg paradox. Daniel Ellsberg (1961) proposed a number of thought-experiments to suggest that, in situations with ambiguity about the nature of the underlying stochastic process, preferences over subjectively uncertain acts would not allow for beliefs over the likelihood of events to be represented by a well-defined probability distribution.

One such choice problem involves an urn from which a ball will be drawn. Anna knows there are ninety balls in total, of which thirty are red. However, the only information she has about the remaining sixty balls is that some are black and some are white—she is not told the actual proportions. Consider two choice problems.

1. A choice between (f) an act that pays \$100 if the ball drawn is red and nothing if it is black or white, and (g) an act that pays \$100 if the ball drawn is black and nothing if it is red or white.
2. A choice between (f') an act that pays \$100 if the ball drawn is red or white and nothing if it is black, and (g') an act that pays \$100 if the ball drawn is black or white and nothing if it is red.

Ellsberg conjectured that many individuals would be averse to ambiguity in the sense that they would prefer to bet on "known" rather than "unknown" odds. In this example, they would strictly prefer the bet on red to the bet on black in the first problem ($f \succ g$)—indicating a subjective belief that black is *less* likely than red—but then prefer the bet on black or white to the bet on red or white in the second problem ($g' \succ f'$)—indicating a subjective belief that black is *more* likely than red. But such a preference pattern is inconsistent with beliefs being represented by a well-defined probability measure.

In response, models have been developed in which beliefs are represented by multiple measures and/or non-additive capacities (which is a generalization of a probability measure). Examples are Gilboa and Schmeidler (1989) and Schmeidler (1989).

We have mentioned only a small sample of critiques of classic expected utility theory and of the extensions to that theory. This theme is developed further in other chapters of this Handbook.

References

Allais, Maurice (1953). La psychologie de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21, 503–46.

Anscombe, F. J., and Aumann, R. J. (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics*, 34, 199–205.

Arrow, Kenneth J. (1959). Rational Choice Functions and Orderings. *Economica*, 26, 121–7.

Bernoulli, Daniel (1738). *Specimen theoriae novae de mensura sortis. Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 5, 175–92.

Birkhoff, Garrett (1948). *Lattice Theory*. New York: American Mathematical Society.

Cantor, Georg (1895). Beiträge zur Begründung der trans nieten Mengenlehre I. *Mathematische Annalen*, 46, 481–512.

Chew, Soo Hong (1989). Axiomatic Utility Theories with the Betweenness Property. *Annals of Operations Research*, 19, 273–98.

de Finetti, Bruno (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17, 298–329.

Debreu, Gerard (1954). Representation of a Preference Ordering by a Numerical Function. In R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), *Decision Processes*, 159–65. New York: Wiley.

—— (1960). Topological Methods in Cardinal Utility Theory. In K. J. Arrow, S. Karlin, and P. Suppes (eds.), *Mathematical Methods in the Social Sciences, 1959*, 16–26. Stanford, CA: Stanford University Press.

—— (1964). Continuity Properties of Paretian Utility. *International Economic Review*, 5, 285–93.

DEKEL, EDDIE (1986). An Axiomatic Characterization of Preferences under Uncertainty: Weakening the Independence Axiom. *Journal of Economic Theory*, 40, 304–18.

ELLSBERG, DANIEL (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643–69.

FISHBURN, PETER (1970). *Utility Theory for Decision Making*. New York: Wiley.

—— (1984). Dominance in SSB Utility Theory. *Journal of Economic Theory*, 34, 130–48.

—— and WAKKER, PETER (1995). The Invention of the Independence Condition for Preferences. *Management Science*, 41, 1130–44.

GILBOA, ITZHAK, and SCHMEIDLER, DAVID (1989). Maxmin Expected Utility with a Non-Unique Prior. *Journal of Mathematical Economics*, 18, 141–53.

HAMMOND, PETER J. (1988). Consequentialist Foundations for Expected Utility. *Theory and Decision*, 25, 25–78.

HERSTEIN, I. N., and MILNOR, JOHN (1953). An Axiomatic Approach to Measurable Utility. *Econometrica*, 21, 291–7.

JENSEN, NIELS-ERIK (1967). An Introduction to Bernoullian Utility Theory, I, II. *Swedish Journal of Economics*, 69, 163–83, 229–47.

KAHNEMAN, DANIEL, and TVERSKY, AMOS (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263–91.

KARNI, EDI (1985). *Decision Making under Uncertainty: The Case of State-Dependent Preferences*. Cambridge, MA: Harvard University Press.

KRANTZ, DAVID H., LUCE, R. DUNCAN, SUPPES, PATRICK, and TVERSKY, AMOS (1971). *Foundations of Measurement*, i: *Additive and Polynomial Representations*. New York: Academic Press.

LOOMES, GRAHAM, and SUGDEN, ROBERT (1982). Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. *Economic Journal*, 92, 805–24.

MACHINA, MARK J. (1982). 'Expected Utility' Analysis without the Independence Axiom. *Econometrica*, 50, 277–323.

QUIGGIN, JOHN (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization*, 3, 323–43.

RAMSEY, FRANK P. (1931). Truth and Probability. In *Foundations of Mathematics and Other Logical Essays*. London: K. Paul, Trench, Trubner, Co.

SAMUELSON, PAUL A. (1938). A Note on the Pure Theory of Consumer's Behaviour. *Economica*, 5, 61–71.

SAVAGE, LEONARD J. (1954). *The Foundations of Statistics*. New York: Wiley.

SCHMEIDLER, DAVID (1989). Subjective Probability and Expected Utility without Additivity. *Econometrica*, 57, 571–87.

STROTZ, ROBERT H. (1959). The Utility Tree—A Correction and Further Appraisal. *Econometrica*, 27, 482–8.

TVERSKY, AMOS, and KAHNEMAN, DANIEL (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.

VON NEUMANN, JOHN, and MORGENSTERN, OSKAR (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

WAKKER, PETER (1988). The Algebraic versus the Topological Approach to Additive Representations. *Journal of Mathematical Psychology*, 32, 421–35.

WAKKER, PETER (1989). *Additive Representations of Preferences: A New Foundation of Decision Analysis*. Dordrecht: Kluwer Academic Publishers.

—— and TVERSKY, AMOS (1993). An Axiomatization of Cumulative Prospect Theory. *Journal of Risk and Uncertainty*, 7, 147–76.

—— and ZANK, HORST (1999). State-Dependent Expected Utility for Savage's State Space. *Mathematics of Operations Research*, 24, 8–34.

YAARI, MENAHEM E. (1987). The Dual Theory of Choice under Risk. *Econometrica*, 55, 95–115.

CHAPTER 2

RANK-DEPENDENT UTILITY

MOHAMMED ABDELLAOUI

2.1 Introduction

Rank-dependent utility (RDU) is among the most popular families of models for decision under risk and uncertainty that deviate from the standard theory of expected utility. RDU was initially introduced by Quiggin (1982) for decisions with known probabilities (risk), and by Schmeidler (1989) for decisions with unknown probabilities (uncertainty). Subsequently, RDU has been incorporated in the famous original prospect theory (Kahneman and Tversky 1979) giving birth to cumulative prospect theory (Tversky and Kahneman 1992), descriptively the most sophisticated version of RDU.

Under classical expected utility, risk attitude results from the combination of mathematical expectation with the prevailing assumption of decreasing marginal utility, leading to risk aversion. The commonly observed violations of expected utility are handled in RDU through the introduction of non-additive decision weights reflecting what may be called chance attitude (Tversky and Wakker 1995). More specifically, RDU allows for coexistence of gambling and insurance, and explanations of the certainty and common ratio effects. Capturing chance attitude also allows individual preferences to depend not only on the degree of uncertainty, but also on the source of uncertainty (Tversky and Wakker 1995, p. 1255).

I thank Nathalie Etchart-Vincent and Peter P. Wakker for helpful comments and suggestions.

As pointed out by Diecidue and Wakker (2001), RDU models are mathematically sound. For instance, they do not exhibit behavioral anomalies such as implausible violations of stochastic dominance (Fishburn 1978). This is corroborated by the existence of axiomatizations that allow RDU preference representations of individual choice (Quiggin 1982; Gilboa 1987; Schmeidler 1989; Abdellaoui and Wakker 2005). RDU also satisfies another important requirement regarding empirical performance. It has been found in a long list of empirical works that RDU can accommodate several violations of expected utility (e.g. Harless and Camerer 1994; Tversky and Fox 1995; Birnbaum and McIntosh 1996; Gonzalez and Wu 1999; Bleichrodt and Pinto 2000; Abdellaoui, Barrios, and Wakker 2007). Many researchers also agree that RDU is intuitively plausible. Diecidue and Wakker (2001) provide compelling and intuitive arguments in this direction.

The purpose of this chapter is to bring into focus the main violations of expected utility that opened the way to RDU, the intuitions and preference conditions behind rank dependence, and, finally, a few recent empirical results regarding these models.

2.2 Background: Expected Utility and its Violations

Mathematical expectation was considered by early probabilists as a good rule to be used for the evaluation of individual decisions under risk (i.e. with known probabilities), particularly for gambling purposes. If a prospect (lottery ticket) is defined as a list of outcomes with corresponding probabilities, then one should prefer the prospect with the highest expected value. This rule was, however, challenged by a chance game devised by Nicholas Bernoulli in 1713, known as the St. Petersburg paradox. To solve his cousin's paradox, Daniel Bernoulli (1738) proposed the evaluation of monetary lotteries using a nonlinear function of monetary payoffs called utility. Two centuries later, von Neumann and Morgenstern (1944) gave an axiomatic basis to the expected utility rule with exogenously given probabilities. This allowed for the formal incorporation of risk and uncertainty into economic theory. Subsequently, combining the works of Ramsey (1931) and von Neumann and Morgenstern (vNM), Savage (1954) proposed a more sophisticated axiomatization of expected utility in which "states of the world", the carriers of uncertainty, replace exogenously given probabilities. Savage's approach is based on the assumption that decision-makers' beliefs about states of the world can be inferred from their preferences by means of subjective probabilities.

2.2.1 Expected Utility with Known Probabilities

Expected utility (Eu) theory with known probabilities has been axiomatized in several ways (e.g. vNM 1944; Herstein and Milnor 1953). Below, we will follow Fishburn (1970) and his approach based on probability measures to explain the axioms of expected utility.

Let $\mathcal{X}$ be a set of outcomes and $\mathbb{P}$ the set of simple probability measures, i.e. n-outcome *prospects* on $\mathcal{X}$, with $n < \infty$. By $\succcurlyeq$ we denote the preference relation "weakly preferred to" on $\mathbb{P}$, with "indifference" $\sim$ and "strict preference" $\succ$ defined as usual. The preference relation $\succcurlyeq$ is a *weak order* if it is complete and transitive. It satisfies first-order stochastic dominance on $\mathbb{P}$ if for all $P, Q \in \mathbb{P}$, $P \succ Q$ whenever $P \neq Q$ and for all $x \in X$, $P(\{y \in \mathcal{X}: y \succcurlyeq x\})$ is at least equal to $Q(\{y \in \mathcal{X}: y \succcurlyeq x\})$. For $\alpha \in [0, 1]$, the convex combination $\alpha P + (1 - \alpha)Q$ of prospects P and Q is a prospect (i.e. a probability measure). It can be interpreted as a compound (two-stage) prospect giving P with probability α and Q with probability $1 - \alpha$. The preference relation $\succcurlyeq$ is *Jensen-continuous* if for all prospects $P, Q, R \in \mathbb{P}$, if $P \succ Q$, then there exist $\lambda, \mu \in [0, 1]$ such that $\lambda P + (1 - \lambda)R \succ Q$ and $P \succ \mu R + (1 - \mu)Q$.

The key axiom of expected utility theory with known probabilities is called *vNM-independence*. It is usually formulated as follows:

vNM-independence. *For all* $P, Q, R \in \mathbb{P}$ *and* $\alpha \in [0, 1]$: $P \succcurlyeq Q \Leftrightarrow \alpha P + (1 - \alpha)R \succcurlyeq \alpha Q + (1 - \alpha)R$.

This axiom says that if a decision-maker has to choose between prospects $\alpha P + (1 - \alpha)R$ and $\alpha Q + (1 - \alpha)R$, her choice does not depend on the "common consequence" R. A Jensen-continuous weak order satisfying vNM-independence on the set $\mathbb{P}$ is necessary and sufficient for the existence of a utility function $u: \mathcal{X} \to \mathbb{R}$ such that

$$\forall P, Q \in \mathbb{P},\ P \succcurlyeq Q \Leftrightarrow E(u, P) \geq E(u, Q), \tag{1}$$

where $E(u, R) = \sum_{x \in X} r(x)u(x)$ for any prospect R. The utility function u is unique up to a positive affine transformation (i.e. unique up to level and unit).

2.2.2 Expected Utility with Unknown Probabilities

According to Savage, the ingredients of a decision problem under uncertainty are the *states of the world*, the carriers of uncertainty; the outcomes, the carriers of value; and the *acts*, the objects of choice. The set of states (of the world), denoted $\mathcal{S}$, is such that one and only one of them obtains (i.e. they are mutually exclusive and exhaustive). An *event* is a subset of $\mathcal{S}$. An *act* is a function from $\mathcal{S}$ to $\mathcal{X}$, the set of outcomes. The set of acts is denoted by $\mathcal{A}$. An act is *simple* if $f(\mathcal{S})$ is finite. When an act f is chosen, $f(s)$ is the consequence that will result if state s obtains. For outcome x, event A, and acts f and g: fAg (xAg) denotes the act resulting

from g if all outcomes $g(s)$ on A are replaced by the corresponding outcomes $f(s)$ (by consequence x). The set of acts $\mathcal{A}$ is provided with a complete and transitive preference relation $\succcurlyeq$ (Savage's axiom P1). Strict preference and indifference are defined as usual. An act f is *constant* if for all states s, $f(s) = x$ for some $x \in \mathcal{S}$. The preference relation on acts is extended to the set of consequences by the means of constant acts. Triviality of the preference relation is avoided by assuming that there exist outcomes x and y such that $x \succ y$ (Savage's axiom P5). An event A is said to be *null* if the decision-maker is indifferent between any pair of acts differing only on A.

In the vNM setup, the straightforwardness of "preference continuity" uses the natural richness of the interval of probabilities. In the Savagean setup, the absence of exogenously given probabilities requires defining preference continuity using a rich collection of events—hence Savage's axiom P6 called *small-event continuity*. It states that for any non-indifferent acts ($f \succ g$), and any outcome (x), the state space can be (finitely) partitioned into events ($\{A_1, \ldots, A_n\}$) small enough so that changing either act to equal this outcome over one of these events keeps the initial indifference unchanged ($xA_i f \succ g$ and $f \succ xA_j g$ for all $i, j \in \{1, \ldots, n\}$). This structural axiom generates an infinite state space $\mathcal{S}$. In the presence of a non-trivial weak order satisfying small-event continuity, Savage needs three additional key axioms: the sure-thing principle, eventwise monotonicity, and likelihood consistency.

Sure-thing principle: *For all events A and acts f, g, h and h', $fAh \succcurlyeq gAh \Leftrightarrow fAh' \succcurlyeq gAh'$.*

The sure-thing principle (axiom P2) states that if two acts f and g have a common part over $(-A)$, then the ranking of these acts will not depend on what this common part is. This axiom implies a key property of subjective expected utility: namely, separability of preferences across mutually exclusive events.

Eventwise monotonicity: *For all non-null events A, and outcomes x, y and acts f, $xAf \succcurlyeq yAf \Leftrightarrow x \succcurlyeq y$.*

Eventwise monotonicity (or axiom P3) states that for any act, replacing any outcome y on a non-null event by a preferred/equivalent outcome x results in a preferred/equivalent act.

Likelihood consistency: *For all events A, B and outcomes $x \succ y$ and $x' \succ y'$, $xAy \succcurlyeq xBy \Leftrightarrow x'Ay' \succcurlyeq x'By'$.*

Likelihood consistency (axiom P4) states that the revealed likelihood binary relation $\succcurlyeq^*$ (read "weakly more likely than") defined over events by

$$A \succcurlyeq^* B \quad \text{if for some } x \succ y,\ xAy \succcurlyeq xBy \tag{2}$$

is independent of the specific outcomes x, y used. It is noteworthy that the likelihood relation $\succcurlyeq^*$, representing beliefs, is not a primitive but is inferred from the preference relation over acts.

Savage (1954) shows that axioms P1 to P6 are sufficient for the existence of a unique subjective probability measure P^* on $2^{\mathcal{S}}$, preserving likelihood rankings (i.e. $A \succcurlyeq^* B$ if and only if $P^*(A) \geq P^*(B)$), and satisfying convex-rangeness (i.e. $A \subset \mathcal{S}, \alpha \in [0, 1] \Rightarrow (P^*(B) = \alpha P^*(A)$ for some $B \subset A$). The existence of P^* allows assigning a simple prospect to each simple act in $\mathcal{A}$. More specifically, an act f such that $f(\mathcal{S}) = \{x_1, \ldots, x_n\}$ induces the prospect $P_f = (x_1 : P^*(f^{-1}(x_1)), \ldots, x_n, : P^*(f^{-1}(x_n)))$. Moreover, if acts generate the same prospect, then they should be indifferent ($P_f = P_g \Rightarrow f \sim g$).

The preference relation over simple acts is extended to the set of induced prospects through the equivalence $f \succcurlyeq g \Leftrightarrow P_f \succcurlyeq P_g$. Furthermore, it can be shown that under axioms P1 to P6, vNM axioms are satisfied over the (convex) set of induced prospects. Consequently, there exists a vNM utility function u on $\mathcal{X}$, unique up to level and unit, such that the decision-maker ranks simple acts f on the basis of $E(P_f, u)$.

2.2.3 Violations of Expected Utility

Experimental investigations dating from the early 1950s have revealed a variety of violations of expected utility. The most studied violations concern the independence axiom and its analog for unknown probabilities, the sure-thing principle. Two "paradoxes" emerge as the most popular in the experimental literature: Allais (1953) and Ellsberg (1961). Moreover, numerous experimental studies have shown that risk aversion, the most typical assumption underlying economic analyses, is systematically violated.

2.2.3.1 *The Allais Paradox*

Allais (1953) provides the earliest example of a simple choice situation in which subjects consistently violate the vNM-independence axiom. Table 2.1 presents the two choice situations used in Allais' example: choice between prospects A and B in the first situation, and between A' and B' in the second situation.

The most frequent choice pattern is AB'. To show that these preferences violate the independence axiom, let C and D be two prospects such that C gives \$5M with probability 10/11 and nothing otherwise, and D gives nothing with certainty. Consequently, we have $A = 0.11A + 0.89A$, $B = 0.11C + 0.89A$, $A' = 0.11A + 0.89D$, and $B' = 0.11C + 0.89D$. According to the independence axiom, the preference between $A(A')$ and $B(B')$ should depend on A vs. C preference. Clearly, the

Table 2.1. Allais paradox

	Probabilities		
	$p = 0.01$	$p = 0.10$	$p = 0.89$
A	\$1m	\$1m	\$1m
B	0	\$5m	\$1m
A'	\$1m	\$1m	0
B'	0	\$5m	0

independence axiom requires either the choice pattern AA' or the choice pattern BB'. Following Allais, the certainty of becoming a millionaire encourages people to choose A, while the similarity of the odds of winning in A' and B' encourages them to opt for prospect B'.

2.2.3.2 *The Ellsberg Paradox*

Table 2.2 describes the two choice situations proposed in Ellsberg's example. The subject must choose an alternative (act) in each choice situation. Uncertainty is generated by means of the random draw of a ball from an urn containing thirty red (R) balls as well as sixty balls that are either black (B) or yellow (Y).

Savage's sure-thing principle requires that a strict preference for $f(g)$ should be accompanied by a strict preference for $f'(g')$. Nevertheless, Ellsberg claimed that many reasonable people will exhibit the choice pattern fg'. He suggested that preferring f to g is motivated by ambiguity aversion: the decision-maker has more precise knowledge of the probability of the "winning event" in act f than in act g. Similarly, in the second choice situation, the choice of act g' can be explained by the absence of precise knowledge regarding the probability of event Y. In terms of likelihood relation $\succcurlyeq^*$, it can easily be shown that, under expected utility, the choice pattern fg' implies two contradictory likelihood statements: namely, $R \succ^* B$ and $B \cup Y \succ^* R \cup Y$.

Table 2.2. Ellsberg paradox

	30 balls	60 balls	
	Red	Black	Yellow
f	\$1000	0	0
g	0	\$1000	0
f'	\$1000	0	\$1000
g'	0	\$1000	\$1000

Table 2.3. The fourfold pattern of risk attitudes

	Gain	Loss
Low probability	C($100, 0.05) = $14 risk seeking	C(−$100, 0.05) = −$8 risk aversion
High probability	C($100, 0.95) = $78 risk aversion	C(−$100, 0.95) = −$84 risk seeking

2.2.3.3 *The Fourfold Pattern of Risk Attitudes*

While expected utility does not impose any prior attitude towards risk, risk aversion represents the most typical assumption underlying economic analysis. Numerous studies have shown, however, that this assumption is systematically violated in a way that expected utility cannot explain. Table 2.3 reports aggregated experimental results by Tversky and Kahneman (1992) through median certainty equivalents $C(x, p)$, where (x, p) is the prospect offering $\$x$ with probability p, and nothing otherwise.

Tversky and Kahneman (1992) found evidence in favor of risk seeking (aversion) for low probability gains (losses), and risk aversion (seeking) for high probability gains (losses). Similar experimental results were reported in Cohen, Jaffray, and Said (1987) and Kachelmeier and Shehata (1992), among others. Under expected utility, these results cannot be explained by the shape of the utility function because they occur over a wide range of outcomes (see also Tversky and Wakker 1995).

2.3 Generalizing Expected Utility through Rank Dependence

2.3.1 Generalizations of Expected Utility

Researchers in decision theory have responded to the accumulation of experimental evidence against expected utility by developing new theories of choice with known and unknown probabilities. Many of them demanded, however, that theories generalizing expected utility should satisfy empirical, theoretical, and normative goals (e.g. Machina 1989).

The empirical goal stipulates that the new theory should fit the data better than expected utility. The theoretical goal imposes that the theory should be useful to conduct analysis of standard economic decisions. Following the normative goal, the new theory should have a "minimal" rationality content. While the empirical

and theoretical goals are clear and intuitively reasonable, the normative goal needs more explanation. Transitivity of choice may help to clarify the idea of a minimal rationality content. Indeed, despite the experimental evidence against transitivity (e.g. Tversky 1969), economists emphasize the self-destructive nature of violations of transitivity. A similar line of reasoning can be applied to stochastic dominance under risk and eventwise monotonicity under uncertainty. For instance, Machina (1989, p. 1623) explains that "whereas experimental psychologists can be satisfied as long as their models of individual behavior perform properly in the laboratory, economists are responsible for the logical implications of their behavioral models when embedded in social settings".

Most generalizations of expected utility have been elaborated for choice with known probabilities. Two important families of non-expected utility theories dominate the literature: utility theories satisfying the "betweenness" property and RDU theories. Betweenness implies that there is no preference for or aversion to a randomization between indifferent prospects. This assumption is weaker than the independence axiom, and it has the advantage of retaining much of its normative appeal. In fact, betweenness exhibits interesting characteristics in dynamic choice problems (Green 1987). Furthermore, it is a sufficient condition for the existence of a preference for portfolio diversification (Camerer 1989; Dekel 1989). In other terms, as far as the theoretical and normative goals are concerned, betweenness seems to be close to expected utility. The problem is that, on a descriptive ground, this axiom does not perform better than independence (e.g. Harless and Camerer 1994; Abdellaoui and Munier 1998). Weighted utility theory (Chew and MacCrimmon 1979), implicit weighted utility (Chew 1985), SSB utility theory (Fishburn 1988), and the theory of disappointment aversion (Gul 1991) are the most famous non-expected utility theories satisfying the betweenness property. Counterparts to SSB utility theory for choice under uncertainty are regret theory (Loomes and Sugden 1982) and SSA utility theory (Fishburn 1988).

The second family of non-expected utility models, called rank-dependent utility, has the advantage of accounting for experimental findings by psychologists and decision theorists showing that, in risky choices, subjects have a clear tendency to overweight small probabilities and to underweight moderate and high probabilities (e.g. Kahneman and Tversky 1979; Cohen, Jaffray, and Said 1987). It can be shown that such subjective treatment of probabilities is remarkably consistent with the Allais paradox and the fourfold pattern of attitude towards risk. More recent experimental investigations on individual decision-making with unknown probabilities have revealed similar findings. Individuals exhibit a clear tendency to subjectively overweight unlikely events and underweight likely events (e.g. Tversky and Fox 1995; Wu and Gonzalez 1999; Abdellaoui, Vossmann, and Weber 2005).

For decision under risk, the early experimental findings by Preston and Baratta (1948) showed that, for small changes in wealth (i.e. assuming a linear utility for money), subjects tend to overweight small probabilities (less than 0.2) and to

underweight large ones (above 0.2). Subsequently, descriptive models incorporating the transformation of single-outcome probabilities through a strictly increasing function w satisfying $w(\mathbf{0}) = 0$ and $w(1) = 1$ have been proposed. Handa (1977) suggested the evaluation of prospect $(p_1: x_1, \ldots, p_n: x_n)$ through $\sum_{i=1}^{n} p^*_i u(x_i)$, where $u(x_i) = x_i$, and $p^*_i = w(p_i)$, $i = 1, \ldots, n$. Then, Karmarkar (1978) proposed a more general formula where u is not necessarily the identity function and decision weights are normalized to sum to 1. Kahneman and Tversky (1979) suggested a more sophisticated subjective probability weighting approach including evaluation $\sum_{i=1}^{n} p^*_i u(x_i)$ for a subclass of prospects. However, these models share a drawback leading to violations of first stochastic dominance (Fishburn 1978; Kahneman and Tversky 1979). This observation led Quiggin (1981) to the basic idea of RDU: the attention given to outcome should depend not only on the corresponding probability but also on the favorability of this outcome as compared to other possible outcomes (Diecidue and Wakker 2001). Subsequently, the idea of rank dependence was extended to the case of unknown probabilities by Gilboa (1987) and Schmeidler (1989).

2.3.2 Rank-Dependent Utility for Known Probabilities

2.3.2.1 *Rank-Dependent Evaluation of Prospects*

To introduce the idea of rank dependence, consider the prospect $P = (p_1: x_1, \ldots, p_n: x_n)$ and assume that $x_1 \succcurlyeq \ldots \succcurlyeq x_n$. The RDU value of prospect P is given by

$$\sum_{i=1}^{n} \pi_i u(x_i), \tag{3}$$

where u is the utility function as in EU, and the *decision weight* π_i depends on the ranking of outcome x_i, $i = 1, \ldots, n$. The decision weights $\pi_i s$ are defined by

$$\pi_i = w(p_1 + \ldots + p_i) - w(p_1 + \ldots + p_{i-1}) \tag{4}$$

where w denotes the *probability weighting function*, i.e. a strictly increasing function from $[0, 1]$ to $[0, 1]$, satisfying $w(0) = 0$ and $w(1) = 1$. [1]

The shape of the probability weighting function w introduces optimism and pessimism in the subjective evaluation of prospects (see Diecidue and Wakker 2001). To clarify this point, consider, for example, the prospect $(\frac{1}{3}: \$100; \frac{1}{3}: \$10; \frac{1}{3}: 0)$. Following Eq. 4, the resulting decision weights are given by $\pi_1 = w(\frac{1}{3})$, $\pi_2 = w(\frac{2}{3}) - w(\frac{1}{3})$, and $\pi_3 = 1 - w(\frac{2}{3})$. If we assume that the probability weighting function w is convex, this implies that the weight attached to the worst outcome is higher than the

[1] We assume the convention: $\pi_1 = w(p_1)$.

weight attached to the best outcome ($\pi_3 > 1/3 > \pi_1$). This probability weighting corresponds to a pessimistic "probabilistic risk" attitude, which aggravates risk aversion in the presence of a concave utility function.

RDU is able to explain the most well-known violations of expected utility such as the Allais paradox. Indeed, using a RDU evaluation of prospects in Table 2.1 with $u(0) = 0$, the preference pattern AB yields:

$$\begin{cases} w(1)u(\$1\text{m}) > w(0.1)u(\$5\text{m}) + [w(0.99) - w(0.1)] \\ w(0.1)u(\$5\text{m}) > w(0.11)u(\$1\text{m}) \end{cases}$$

which together imply $w(1) - w(0.99) > w(0.11) - w(0.10)$. This last inequality reflects subjects' tendency to assign a less important subjective impact to the replacement of probability 0.10 by probability 0.11 than to the replacement of probability 0.99 by probability 1. By contrast, expected utility requires that such probability replacements should have the same subjective impact.

2.3.2.2 *A Key Preference Condition for Rank Dependence*

Most axiomatic approaches of RDU have assumed richness of the outcome space. For instance, it is a continuum in Quiggin (1982), Chew (1989), Segal (1989, 1990), Wakker (1994), and Chateauneuf (1999), and a solvable space in Nakamura (1995). Strangely enough, only three papers used richness in the probability dimension to characterize RDU for risk: Nakamura (1995), Abdellaoui (2002), and Zank (2004). Abdellaoui (2002, thm. 9, p. 726) shows that under usual conditions of a Jensen-continuous weak order satisfying first stochastic dominance, a preference condition called *probability trade-off consistency* is necessary and sufficient for RDU. Abdellaoui and Wakker (2005) propose a new version of this condition based on consistency of revealed orderings of decision weights.

Let $P = (p_1: x_1, \ldots, p_n: x_n)$ and $Q = (q_1: y_1, \ldots, q_m: y_m)$ denote the prospects yielding x_i with probability p_i and y_j with probability q_j respectively, where it is understood that $x_1 \succcurlyeq \ldots \succcurlyeq x_n$ and $y_1 \succcurlyeq \ldots \succcurlyeq y_m$. Assume that, under RDU, the corresponding decision weights are $\pi_1^P, \ldots, \pi_n^P$ for prospect P, and $\pi_1^Q, \ldots, \pi_m^Q$ for prospect Q. For given i and j, prospects P and Q can be rewritten as follows:

$$\begin{cases} P = p_i x_i + (1 - p_i)P^* \\ Q = q_j y_j + (1 - q_j)Q^* \end{cases}$$

where prospects P^* and Q^* are obtained by means of a suitable modification of P and Q (conditioning them on the nonoccurrence of x_i and y_i) respectively. Consider two outcomes z and t such that $t \succ z$ and assume that replacing *both* outcomes x_i and y_j by either z or t keeps unchanged the initial rank orderings in prospects P and Q.

To illustrate the idea of revealed orderings of decision weights, assume the following preferences:

$$\begin{cases} p_i z + (1 - p_i) P^* \sim q_j z + (1 - q_j) Q^* \\ p_i t + (1 - p_i) P^* \sim q_j t + (1 - q_j) Q^* \end{cases} \tag{5}$$

Because outcome z is replaced by a strictly preferred outcome t without changing the rank ordering of outcomes in the left prospect as well as in the right prospect, the corresponding decision weights ($\pi_i^P s$ and $\pi_j^Q s$) should remain unchanged. Under RDU, the left consequence replacement entails an improvement $\pi_i^P [u(t) - u(z)]$, and the right consequence replacement generates an improvement $\pi_j^Q [u(t) - u(z)]$. If the change of consequences does not result in a change of preference (i.e. indifference holds), then $\pi_i^P = \pi_j^Q$, meaning that the decision weight of probability p_i, in a rank of probability $p_1 + \ldots + p_{i-1}$, is equal to that of probability q_j in a rank of probability $q_1 + \ldots + q_{j-1}$. If the second indifference $\sim$ in (5) is replaced by strict preference $\succ$, then $\pi_i^P > \pi_j^Q$.

Intuitively, revealed rankings of decision weights should not be influenced by the consequences used to elicit them. In other words, we should not be able to find a pair of consequences z' and t' (keeping unchanged the initial rank ordering of outcomes in prospects P and Q) such that the resulting ranking of decision weights π_i^P and π_j^Q contradicts that obtained using consequences z and t. The corresponding consistency axiom is comprised in the following condition.

Consistency condition: *Consistency of decision weights revealed orderings holds if* $\pi_i^P > \pi_j^Q$ *and* $\pi_i^P = \pi_j^Q$ *for no probabilities* p_i *and* q_j, $i = 1, \ldots, n$, *and* $j = 1, \ldots, m$.

The above consistency condition can replace the probability trade-off condition of Abdellaoui (2002) to characterize RDU for known probabilities (see thm. 5.7 in Abdellaoui and Wakker 2005).

2.3.3 Rank-Dependent Utility for Unknown Probabilities

The analysis presented in this subsection deals with decision situations in which probabilities of uncertain events are not exogenously available. In this case, subjective degrees of belief interfere in individual choice. As for the Savagean setup presented in Section 2.2.2, we restrict our attention to simple acts. The rank-dependent evaluation of an act is also called *Choquet expected utility* (CEU). Under CEU, decision weights are obtained from a capacity, i.e. a set function W mapping events to $[0, 1]$, with $W(\varnothing) = 0$, $W(S) = 1$, and $W(A) \leq W(B)$ whenever $A \subset B$.

2.3.3.1 *Rank-Dependent Evaluation of Acts*

The rank-dependent evaluation of an act $f = (E_1 : x_1, \ldots, E_n : x_n)$, with $x_1 \succcurlyeq \ldots \succcurlyeq x_n$, is given by

$$\sum_{i=1}^{n} \pi_i u(x_i), \tag{6}$$

where u is the utility function as in SEU, and the *decision weight* π_i depends on the ranking of outcome x_i, $i = 1, \ldots, n$. The decision weights $\pi_i s$ are defined by

$$\pi_i = W(E_1 \cup \ldots \cup E_i) - W(E_1 \cup \ldots \cup E_{i-1}), \tag{7}$$

where W denotes a capacity on the state space $\mathcal{S}$.

Assume for simplicity that no indifference holds between two outcomes of the act f. The decision weight of event E_i is the marginal W contribution of this event to the *dominating* event $R_i = E_1 \cup \ldots \cup E_{i-1}$ of receiving a better outcome. Formally, this means that equation (7) can be rewritten as follows:

$$\pi_i = W(E_i \cup R_i) - W(R_i) \tag{8}$$

This equation reflects the dependence of the decision weight π_i on E_i and on the dominating event R_i. Because $R_j \supset R_k$ if and only if $k \geq j$, R_i reflects the rank of event E_i (see Abdellaoui and Wakker 2005).

For more general settings, given an act f, event R is the rank of an event A if $A \cap R = \varnothing$ and $f(t) \succcurlyeq f(t') \succcurlyeq f(t'')$ for all $t \in R$, $t' \in A$, and $t'' \in (A \cup R)^c$. To avoid problems posed by the rank ordering of states with equivalent consequences, we choose a rank ordering of states that is compatible with f and such that the states in R are ranked higher than the states in A, and the latter are ranked higher than those in $(A \cup R)^c$. A^R is a *ranked event* if R is the rank of A and $A \cap R$ is empty. For a ranked event A^R, the decision weight $\pi(A^R)$ is $W(A \cup R) - W(R)$.

A pessimistic attitude towards uncertainty means that the attention paid to an event (reflected by its decision weight) gets higher as the event gets rank-ordered worse. Formally, this corresponds to a *convex* W: the marginal W contribution of an event E to a disjoint event R, $W(E \cup R) - W(R)$, is increasing in R (with respect to set inclusion) for all E. In other words, worsening the ranking position (R_i increases with respect to set inclusion) increases the decision weight of E_i. Optimism is similarly characterized by concavity of W.

2.3.3.2 *A Key Preference Condition for Rank Dependence*

Abdellaoui and Wakker (2005) propose a key preference condition for RDU with unknown probabilities. This condition is similar to the one presented above for risk. It excludes inconsistent revealed orderings of decision weights.

Table 2.4. Modifying the Ellsberg paradox (m denotes $1000)

(a)							(b)						
B	$Y^{B'}$	R		R	$Y^{B'}$	B	B'	$Y^{B'}$	R'		R'	$Y^{B'}$	B'
m	**0**	0	$\succ$	m	**0**	0	m	**0**	0	$\sim$	m	**0**	0
m	**m**	0	$\succ$	m	**m**	0	m	**m**	0	$\succ$	m	**m**	0

For a given act f and a consequence x, the notation $x_{A^R} f$ designates the act resulting from f if all outcomes for event A are replaced by x, as did $x_A f$, but specifies that R is the rank of A in $x_A f$. That is, all outcomes under R are weakly preferred to x, and those under $(R \cup A)^c$ are weakly less preferred than x.

The ranked event A^R is *revealed more likely* than the ranked event $B^{R'}$, $A^R \succ^* B^{R'}$, if

$$\begin{cases} z_{A^R} f \ \sim z_{B^{R'}} g \\ t_{A^R} f \succ t_{B^{R'}} g \end{cases} \tag{9}$$

hold for some $t \succ z$. If the strict preference $\succ$ in (9) is replaced by indifference $\sim$, then A^R is revealed equally likely as the ranked event $B^{R'}(A^{\uparrow} R \sim^{\uparrow *} B^{\uparrow}(R^{\uparrow'}))$. It can easily be shown that, under CEU,

$$\begin{cases} A^R \succ^* B^{R'} \Rightarrow \pi(A^R) > \pi(B^{R'}) \\ A^R \sim^* B^{R'} \Rightarrow \pi(A^R) = \pi(B^{R'}). \end{cases}$$

A modified version of the Ellsberg three-color paradox can illustrate the idea of revealed (rank-dependent) likelihood rankings. Events are rank-ordered from best (left) to worst (right) outcomes in Tables 2.4a and b. The modified Ellsberg example displayed in Table 2.4b is obtained through the moving of a small event from R to B, leading to a smaller R' and a larger B'. Taking into account the similarity of preferences of Table 2.4b with preferences in (9), we infer that the ranked event $Y^{B'}$ is revealed more likely than the event $Y^{R'}$ $(Y^{B'} \succ^* Y^{R'})$. Consistency of revealed orderings of decision weights needs the following rank-dependent likelihood consistency.

Consistency condition: *Consistency of revealed orderings of decision weights holds if* $A^R \succ^* B^{R'}$ *and* $A^R \sim^* B^{R'}$ *hold for no ranked events* A^R *and* $B^{R'}$.

In the presence of a few standard conditions, the above consistency condition is necessary and sufficient for a CEU representation of preferences (Abdellaoui and Wakker 2005, thm. 5.7).

2.4 Empirical Elicitation of RDU Models

2.4.1 Cumulative Prospect Theory

Most attempts to elicit RDU experimentally have focused on its more descriptive and more general version called cumulative prospect theory (CPT), initially proposed by Tversky and Kahneman (1992). In this model, it is postulated that the carriers of utility are gains and losses, instead of the final asset position as typically assumed in EU and RDU. Therefore, CPT assumes that the utility function u satisfies $u(0) = 0$. Furthermore, this theory invokes two probability weighting functions: w^+ for gains and w^- for losses, leading to "sign dependence" (of the weighting function). The CPT value of a prospect $P = (p_1: x_1, \ldots, p_n: x_n)$ with $x_1 \geq \ldots \geq x_r \geq 0 \geq x_{r+1} \geq \ldots \geq x_n$ is given by

$$\sum_{i=1}^{r} \pi_i^+ u(x_i) + \sum_{i=r+1}^{n} \pi_i^- u(x_i), \tag{10}$$

where the decision weights for gains are defined by $\pi_i^+ = w^+(p_1 + \ldots + p_i) - w^+(p_1 + \ldots + p_{i-1})$ and the decision weights for losses by $\pi_i^- = w^-(p_i + \ldots + p_n) - w^-(p_{i+1} + \ldots + p_n)$. It can be easily shown that these decision weights do not sum to 1.

For choice under uncertainty, the weighting functions for gains and losses are denoted by W^+ and W^-. The value of an act $f = (A_1: x_1, \ldots, A_n: x_n)$ with $x_1 \geq \ldots \geq x_r \geq 0 \geq x_{r+1} \geq \ldots \geq x_n$ is given by a formula similar to (10), except that

$$\pi_i^+ = W^+(A_1 \cup \ldots \cup A_i) - W^+(A_1 \cup \ldots \cup A_{i-1})$$

and

$$\pi_i^- = W^-(A_i \cup \ldots \cup A_n) - W^-(A_{i+1} \cup \ldots \cup A_n).$$

RDU (CEU) corresponds to the special case where the weighting for losses is the dual of the weighting function for gains: $w^-(p) = 1 - w^+(1 - p)$ ($W^-(A) = 1 - W^+(S - A)$). For prospects (acts) with nonnegative outcomes, RDU (CEU) and CPT coincide.

2.4.2 Utility for Gains and Losses

Following Kahneman and Tversky (1979) and Tversky and Kahneman (1992), the shape of the utility function $u(.)$ reflects psychophysics of diminishing sensitivity: marginal impact (of money) diminishes with distance from the reference point 0. This means that the utility function is concave for gains, and convex for losses. Consequently, concavity of $u(.)$ on the gains domain contributes to risk aversion

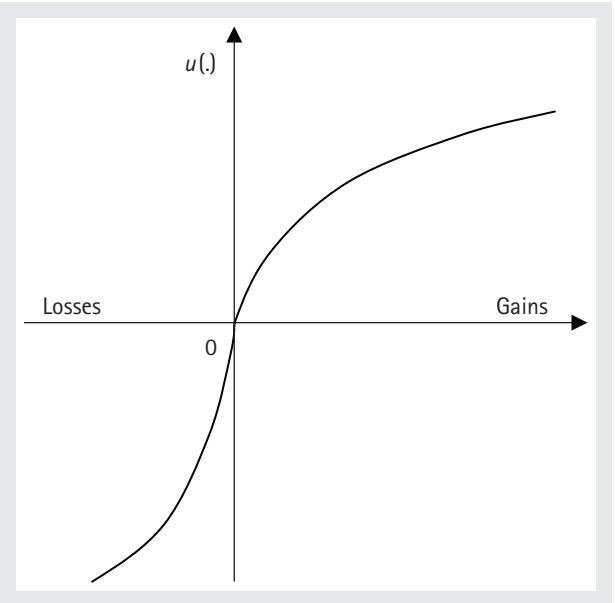

Fig. 2.1. Utility function for monetary gains and losses.

for gains, and convexity on the loss domain contributes to risk seeking for losses (e.g. Fox and See 2003). The utility function is also assumed steeper for losses than for gains, reflecting *loss aversion* (see Figure 2.1).

The traditional methods that are commonly used to measure utility under expected utility (e.g. certainty equivalent method, probability equivalent method) are no longer valid under CPT because they cannot account for probability weighting and loss aversion. Tversky and Kahneman (1992) opened the way to measure utility for gains and losses simultaneously by assuming specific parametric forms for utility and probability weighting. While recognizing the merits of such a parametric study and some others (e.g. Camerer and Ho 1994), one must also agree that the findings may have been confounded by the particular parametric families chosen. Wakker and Deneffe (1996) provided the first parameter-free utility elicitation method, the trade-off method, which is robust to probability weighting. Subsequently, Abdellaoui, Bleichrodt, and Paraschiv (2007) provided a nonparametric method to completely measure utility under CPT.

Most empirical studies of the shape of the utility function have confirmed that, under CPT/RDU, utility is concave for gains and convex for losses. Fennema and van Assen (1999), Abdellaoui (2000), Schunk and Betsch (2006), and Abdellaoui, Barrios, and Wakker (2007) found that utility for gains was concave at the aggregate level and for a majority of subjects. The available evidence is stronger, however, for gains than for losses. Fennema and van Assen (1999), Abdellaoui (2000), Etchart-Vincent (2004), Abdellaoui, Vossmann, and Weber (2005), and Schunk and Betsch

Table 2.5. Examples of power utility parameter estimates

Experimental study	$u(x)$	Median estimates	
		Gains	Losses
Tversky and Kahneman (1992)	$(*)\begin{cases} (x)^{\alpha}, & x \geq 0 \\ -\lambda(-x)^{\beta}, & x < 0 \end{cases}$	$\alpha = 0.88$	$\beta = 0.88$
Wu and Gonzalez (1996)		$\alpha = 0.50$	
Gonzalez and Wu (1999)		$\alpha = 0.49$	
Abdellaoui (2000)		$\alpha = 0.89$	$\beta = 0.92$

(*) λ stands for loss aversion coefficient.

(2006) found slightly convex utility for losses at the aggregate level. Baucells and Heukamp (2006) and Abdellaoui, Bleichrodt, and Paraschiv (2007) found strong support for an S-shaped utility function (i.e. concave for gains and convex for losses) at both the aggregate and individual levels. Table 2.5 reports median estimates of the utility function assuming a power parametric form. The values shown in Table 2.5 are consistent with an S-shaped utility for money, as postulated in prospect theory.

2.4.3 Weighting for Gains and for Losses

According to prospect theory, the shape of the probability weighting function reflects diminishing sensitivity. The weighting function has two reference points: impossibility and certainty. For decision under risk, diminishing sensitivity implies an inverse S-shaped probability weighting function that is concave for small probabilities and convex for middle and high probabilities, as depicted in Figure 2.2.

Overweighting of small probabilities contributes to risk aversion in the loss domain, and risk seeking in the gain domain. It thus counterbalances the effect of the shape of the utility function on attitude towards risk. Underweighting of intermediate and large probabilities reinforces risk aversion for gains and risk seeking for losses. Figure 2.2 illustrates an S-shaped probability weighting function (reflecting diminishing sensitivity). It shows that the lower probability interval $[0, q]$ has more impact than the middle interval $[p, p+q]$, which is bounded away from the lower and upper endpoints. Similarly, the upper interval $[1-q, 1]$ has more impact than the middle interval $[p, p+q]$. These two notions are known as lower subadditivity (LSA), and upper subadditivity (USA). For decision under uncertainty, diminishing sensitivity results in the overweighting of unlikely events and underweighting of likely events. LSA and USA of $W(.)$ are thus defined as follows: $W(A) \geq W(A \cup B) - W(B)$ and $1 - W(S - A) \geq W(A \cup B) - W(B)$, provided that A and B are disjoint, and $W(A \cup B)$ and $W(B)$ are bounded away from 1 and 0, respectively.

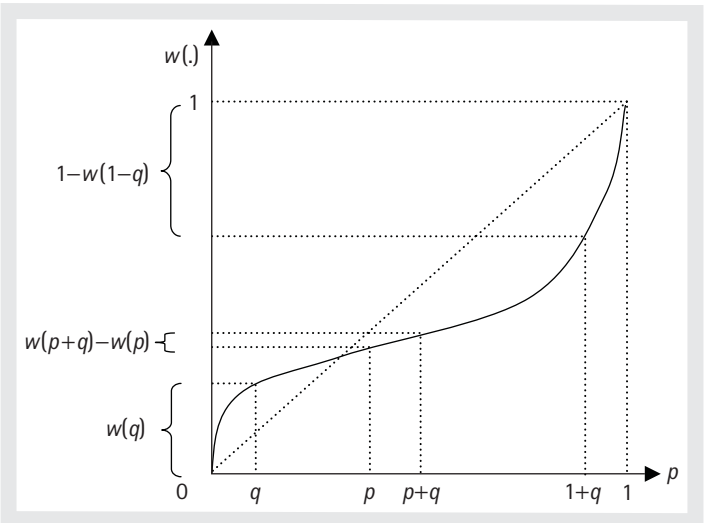

Fig. 2.2. Probability weighting function.

Most experimental studies on probability weighting report an S-shaped pattern both for gains and for losses (Tversky and Kahneman 1992; Gonzalez and Wu 1999; Abdellaoui 2000; Bleichrodt and Pinto 2000). Several parametric specifications of the probability weighting function are used in these experimental studies. Tversky and Kahneman (1992) use the most popular single-parameter form allowing for concave, convex, S-shape, and inverse-S-shape probability weighting functions. The most widely used two-parametric specifications were proposed by Goldstein and Einhorn (1987) and Prelec (1998). Table 2.6 reports some median estimates of the most usual parametric forms.

Table 2.6. Examples of probability weighting function parameter estimates

Experimental study	$w(p)$	Median estimates	
		Gains	Losses
Tversky and Kahneman (1992)	$\frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}$	$\gamma = 0.61$	$\gamma = 0.69$
Wu and Gonzalez (1996)	$e^{(-\ln(p))^\alpha}$ (**)	$\alpha = 0.74$	
Gonzalez and Wu (1999)	$\frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}$ (*)	$\gamma = 0.44$ $\delta = 0.77$	
Abdellaoui (2000)	$\frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}$ (*)	$\gamma = 0.60$ $\delta = 0.65$	$\gamma = 0.65$ $\delta = 0.84$
Bleichrodt and Pinto (2000)	$e^{\beta(-\ln(p))^\alpha}$ (**)	$\alpha = 0.53$ $\beta = 1.08$	

(*) Initially proposed by Goldstein and Einhorn (1987); (**) Initially proposed by Prelec (1998).

For decision under uncertainty (with unknown probabilities), empirical research has shown that the weighting function exhibits diminishing sensitivity, resulting in the overweighting of unlikely events and underweighting of likely events (e.g. Tversky and Fox 1995; Wu and Gonzalez 1999; Kilka and Weber 2001; Abdellaoui, Vossmann, and Weber 2005).

2.5 Conclusion

In this chapter, I have argued that RDU is not just a sound formal generalization of expected utility but that it is one of the most compelling descriptive theories of decision under risk and uncertainty. More specifically, RDU accomplishes two important things. First, it accommodates a wide range of discrepancies from expected utility and allows for non-biased measurement of utility for normative and descriptive purposes. Second, it introduces the useful concepts of nonlinear weighting of chance, and non-additive measures of belief. This makes it possible to relate risk attitude to the way people feel about risk and uncertainty. It also opens a promising avenue for future research on the observation that people's preferences depend not only on the degree of uncertainty but also on the source of uncertainty (Tversky and Wakker 1995; Wakker 2006).

References

Abdellaoui, M. (2000). Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science*, 46, 1497–512.

—— (2002). A Genuine Rank-Dependent Generalization of the von Neumann–Morgenstern Expected Utility Theorem. *Econometrica*, 70, 717–36.

—— and Munier, B. R. (1998). The Risk–Structure Dependence Effect: Experimenting with an Eye to Decision-Aiding. *Annals of Operation Research*, 80, 237–52.

—— and Wakker, P. P. (2005). The Likelihood Method for Decision under Uncertainty. *Theory and Decision*, 58, 3–76.

—— Barrios, C., and Wakker, P. P. (2007). Reconciling Introspective Utility with Revealed Preference: Experimental Arguments Based on Prospect Theory. *Journal of Econometrics*, 138, 356–78.

—— Bleichrodt, H., and Paraschiv, C. (2007). Measuring Loss Aversion under Prospect Theory: A Parameter-Free Approach. *Management Science*, 53, 1659–74.

—— Vossmann, F., and Weber, M. (2005). Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses under Uncertainty. *Management Science*, 51, 1384–99.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21, 503–46.

Baucells, M., and Heukamp, F. H. (2006). Stochastic Dominance and Cumulative Prospect Theory. *Management Science*, 52, 1409–23.

Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica*, 22, 23–36. Translated by L. Sommer from *Specimen theoriae novae de mensura sortis, Commentarii Academiae Scientarum Imperialis Petropolitanae*, 5 (1738), 175–92.

Birnbaum, M., and McIntosh, W. R. (1996). Violations of Branch Independence in Choices between Gambles. *Organizational Behavior and Human Decision Processes*, 67, 91–110.

Bleichrodt, H., and Pinto, J. L. (2000). A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. *Management Science*, 46, 1485–96.

Camerer, C. F. (1989). An Experimental Test of Several Generalized Utility Theories. *Journal of Risk and Uncertainty*, 2, 61–104.

——and Ho, T.-H. (1994). Violations of the Betweenness Axiom and Nonlinearity in Probalility. *Journal of Risk and Uncertainty*, 8, 167–96.

Chateauneuf, A. (1999). Comonotonicity Axioms and Rank-Dependent Expected Utility Theory for Arbitrary Consequences. *Journal of Mathematical Economics*, 32, 21–45.

Chew, S. H. (1985). An Axiomatization of the Rank-Dependent Quasilinear Mean Generalizing the Gini Mean and the Quasilinear Mean. Economics Working Paper no. 156, Johns Hopkins University.

——and MacCrimmon, K. R. (1979). Alpha–nu Choice Theory: An Axiomatization of Expected Utility. University of British Columbia Faculty of Commerce Working Paper no. 669.

Cohen, M., Jaffray, J-Y., and Said, T. (1987). Experimental Comparisons of Individual Behavior under Risk and under Uncertainty for Gains and for Losses. *Organizational Behavior and Human Decision Processes*, 39, 1–22.

Dekel, E. (1989). Asset Demands without the Independence Axiom. *Econometrica*, 57, 163–9.

Diecidue, E., and Wakker, P. P. (2001). On the Intuition of Rank-Dependent Utility. *Journal of Risk and Uncertainty*, 23/3, 281–98.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643–69.

Etchart-Vincent, N. (2004). Is Probability Weighting Sensitive to the Magnitude of Consequences? An Experimental Investigation on Losses. *Journal of Risk and Uncertainty*, 28, 217–35.

Fennema, H., and van Assen, M. (1999). Measuring the Utility of Losses by Means of the Trade-off Method. *Journal of Risk and Uncertainty*, 17, 277–95.

Fishburn, P. C. (1970). *Utility Theory for Decision Making*. New York: Wiley.

——(1978). On Handa's New Theory of Cardinal Utility and Maximization of Expected Return. *Journal of Political Economy*, 86, 321–4.

——(1988). *Nonlinear Preference and Utility Theory*. Baltimore: Johns Hopkins University Press.

Fox, C. R., and See, K. E. (2003). Belief and Preference in Decision under Uncertainty. In D. Hardman and L. Macchi (eds.), *Reasoning and Decision Making: Current Trends and Perspectives*, 273–314. New York: Wiley.

Gilboa, I. (1987). Expected Utility with Purely Subjective Non-Additive Probabilities. *Journal of Mathematical Economics*, 16, 65–88.

Goldstein, W. M., and Einhorn, H. J. (1987). Expression Theory and the Preference Reversal Phenomenon. *Psychological Review*, 94, 236–54.

Gonzalez, R., and Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38, 129–66.

Green, J. R. (1987). Making Book against Oneself, the Independence Axiom, and Nonlinear Utility Theory. *Quarterly Journal of Economics*, 102, 785–96.

Gul, F. (1991). A Theory of Disappointment Aversion. *Econometrica*, 59, 667–86.

Handa, J. (1977). Risk, Probabilities, and a New Theory of Cardinal Utility. *Journal of Political Economy*, 85, 97–122.

Harless, D. W., and Camerer, C. F. (1994). The Predictive Utility of Generalized Expected Utility Theories. *Econometrica*, 62, 1251–89.

Herstein, I. N., and Milnor, J. (1953). An Axiomatic Approach to Measurable Utility. *Econometrica*, 21, 291–7.

Kachelmeier, S. J., and Shehata, M. (1992). Examining Risk Preferences under High Monetary Incentives: Experimental Evidence from the People's Republic of China. *American Economic Review*, 82, 1120–41.

Kahnemann, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263–91.

Karmarkar, U. S. (1978). Subjective Weighted Utility: A Descriptive Extension of the Expected Utility Model. *Organizational Behavior and Human Performance*, 21, 61–72.

Kilka, M., and Weber, M. (2001). What Determines the Shape of the Probability Weighting Function under Uncertainty. *Management Science*, 47, 1712–26.

Loomes, G., and Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. *Economic Journal*, 92, 805–24.

Machina, M. J. (1989). Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty. *Journal of Economic Literature*, 27, 1622–88.

Nakamura, Y. (1995). Rank Dependent Utility for Arbitrary Consequence Spaces. *Mathematical Social Sciences*, 29, 103–29.

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*. 66, 497–527.

Preston, M. G., and Baratta, P. (1948). An Experimental Study of the Auction Value of an Uncertain Outcome. *American Journal of Psychology*, 61, 183–93.

Quiggin, J. (1981). Risk Perception and Risk Aversion among Australian Farmers. *Australian Journal of Agricultural Economics*, 25, 160–9.

—— (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization*, 3, 323–43.

Ramsey, F. P. (1931). Truth and Probability. In *The Foundations of Mathematics and Other Logical Essays*, 156–98. London: Routledge & Kegan Paul.

Savage, L. (1954). *The Foundations of Statistics*. New York: Wiley.

Schmeidler, D. (1989). Subjective Probability and Expected Utility without Additivity. *Econometrica*, 57, 571–87.

Schunk, D., and Betsch, C. (2006). Explaining Heterogeneity in Utility Functions by Individual Differences in Decision Modes. *Journal of Economic Psychology*, 27, 386–401.

Segal, U. (1989). Anticipated Utility: A Measure Representation Approach. *Annals of Operations Research*, 19, 359–73.

—— (1990). Two-Stage Lotteries without the Reduction Axiom. *Econometrica*, 58, 349–77.

Tversky, A. (1969). Intransitivity of Preferences. *Psychological Review*, 76, 31–48.

——— and Fox, C. R. (1995). Weighing Risk and Uncertainty. *Psychological Review*, 102, 269–83.
——— and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5, 297–23.
——— and Wakker, P. P. (1995). Risk Attitudes and Decision Weights. *Econometrica*, 63, 1255–80.
von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.
Wakker, P. P. (1994). Separating Marginal Utility and Probabilistic Risk Aversion. *Theory and Decision*, 36, 1–44.
——— (2001). Testing and Characterizing Properties of Nonadditive Measures through Violations of the Sure-Thing Principle. *Econometrica*, 69, 1039–59.
——— (2006). Uncertainty. In Lawrence Blume and Steven N. Durlauf (eds.), *The New Palgrave: A Dictionary of Economics*, 6780–91. London: Macmillan.
——— and Deneffe, D. (1996). Eliciting von Neumann–Morgenstern Utilities when Probabilities are Distorted or Unknown. *Management Science*, 42, 1131–50.
——— and Tversky, A. (1993). An Axiomatization of Cumulative Prospect Theory. *Journal of Risk and Uncertainty*, 7, 147–76.
Wu, G., and Gonzalez, R. (1996). Curvature of the Probability Weighting Function. *Management Science*, 42, 1676–90.
——— (1999). Nonlinear Decision Weights in Choice under Uncertainty. *Management Science*, 45, 74–85.
Zank, H. (2004). Deriving Rank-Dependent Expected Utility through Probabilistic Consistency. Unpublished paper, School of Economic Studies, University of Manchester, Manchester, UK.

CHAPTER 3

APPLICATIONS OF NON-EXPECTED UTILITY

HAN BLEICHRODT
ULRICH SCHMIDT

3.1 Introduction

Since the work of von Neumann and Morgenstern (1944) expected utility (EU) has been the dominant framework for analyzing decision situations under risk and uncertainty. Starting with the well-known paradoxes of Allais (1953) and Ellsberg (1961), however, a large body of experimental evidence has been gathered which indicates that individuals systematically violate the assumptions underlying EU. This empirical evidence has motivated researchers to develop alternative theories of choice under risk and uncertainty able to accommodate the observed patterns of behavior. These models are usually referred to as non-expected utility models. Nowadays the rank-dependent models, in particular prospect theory, have become the most prominent alternative, and, accordingly, these models will also be the main focus of our paper.

If the decisions of subjects are not in line with EU, applied models which rest on EU may make wrong predictions. Therefore, applications of non-expected utility models may lead to a better accommodation of real-world data. In general, applications of non-expected utility can be regarded as part of behavioral economics,

a research stream which integrates psychological concepts into economics analysis and has received increasing attention in recent years. Non-expected utility models can in principle be applied to every economic setting involving risk. Due to this fact, it is impossible to cover all fields of applications in the present chapter. We have decided to focus on three fields: insurance economics, auctions, and health economics. Health economics is treated more extensively than the two other fields because it has recently become an important research topic, and to our knowledge no review of applications of non-expected utility in the health domain exists.

The chapter is organized as follows. Section 3.2 gives notation and basic concepts. Section 3.3 describes expected utility, and Section 3.4 describes important non-expected utility models. Sections 3.5 to 3.7 are devoted to a discussion of applications of non-expected utility. Section 3.5 discusses applications in insurance, Section 3.6 in auctions, and Section 3.7 surveys applications of non-expected utility in the health domain.

3.2 Notation and Basic Concepts

Let X denote a set of *outcomes*, which can be quantitative—for example, money amounts or life durations—but also qualitative—e.g. states of health. The set of all probability measures or *prospects* over X will be denoted by P. A prospect $p \in P$ assigns a nonnegative probability p_i to outcome $x_i \in X$, and we have $p(X) = 1$. The set P includes all *riskless prospects*, i.e. prospects that assign probability 1 to one of the outcomes. The probability measure which assigns probability 1 to outcome x is denoted by δ_x. For convenience we restrict attention to probability measures with finite support, i.e. for all $p \in P$ there exists a finite $W \subset X$ with $p(W) = 1$.

Models of decision-making analyze the preference of a decision-maker between prospects which will be formalized by the binary relation $\succcurlyeq \subset P \times P$. For $p, q \in P$, $p \succcurlyeq q$ means that p is at least as good as q (weak preference). The strict preference relation $\succ$ and the indifference relation $\sim$ are defined as usual. By restricting attention to riskless prospects, the preference relation $\succcurlyeq$ defines a preference relation over outcomes, i.e. for all outcomes $x, y \in X$, $x \succcurlyeq y$ iff $\delta_x \succcurlyeq \delta_y$. A real-valued function V on P is called a utility function if it *represents* $\succcurlyeq$ on P, i.e.

$$p \succcurlyeq q \Leftrightarrow V(p) \geq V(q) \text{ for all } p, q \in P. \tag{1}$$

We will denote prospects giving outcome x_i with probability p_i, $i = 1, \ldots, n$, as $(p_1, x_1; \ldots; p_n, x_n)$. Within this notation we implicitly assume that outcomes are rank-ordered from best to worst, i.e. $x_1 \succcurlyeq \ldots \succcurlyeq x_n$. *Binary prospects*, i.e. prospects that yield just two outcomes x_1 and x_2 with positive probabilities p_1 and $p_2 = 1 - p_1$ will be denoted $(p_1, x_1; x_2)$ for short.

3.3 Expected Utility

Expected utility (EU) holds if the utility function representing preference can be restricted to being of the following form:

$$V(p) = \sum_{i=1}^{n} u(x_i)p_i, \tag{2}$$

i.e. the utility of a prospect equals the expected value of the utility of the single outcomes. The central condition in EU is the well-known independence axiom: for all $\lambda \in [0, 1]$ and for all $r \in P$, $p \succcurlyeq q \Rightarrow \lambda p + (1 - \lambda)r \succcurlyeq \lambda q + (1 - \lambda)r$. The independence axiom has intuitive appeal and is accepted as a principle of rational choice by most authors. However, it is often violated in empirical studies. This empirical evidence has motivated the development of non-expected utility models which usually rely on weakened variants of the independence axiom.

A decision-maker is defined to be *risk-averse* if she dislikes mean-preserving spreads in risk. A mean-preserving spread results from increasing one outcome and decreasing a worse outcome without affecting the expected value of a prospect. Consequently, risk aversion holds if $(p_1, x_1; \ldots; p_i, x_i; \ldots; p_j, x_j; \ldots; p_n, x_n) \succcurlyeq (p_1, x_1; \ldots; p_i, x_i + \epsilon/p_i; \ldots; p_j, x_j - \epsilon/p_j; \ldots; p_n, x_n)$ for all positive ϵ. It follows that risk aversion in EU is equivalent to a concave utility function u.

Two common graphical representations, the two-outcome diagram and the triangle diagram, may help to clarify some properties of EU (see Figure 3.1). The two-outcome diagram restricts attention to binary prospects with outcomes x_1 and x_2, which occur with probabilities p_1 and $1 - p_1$. Setting the total differential of

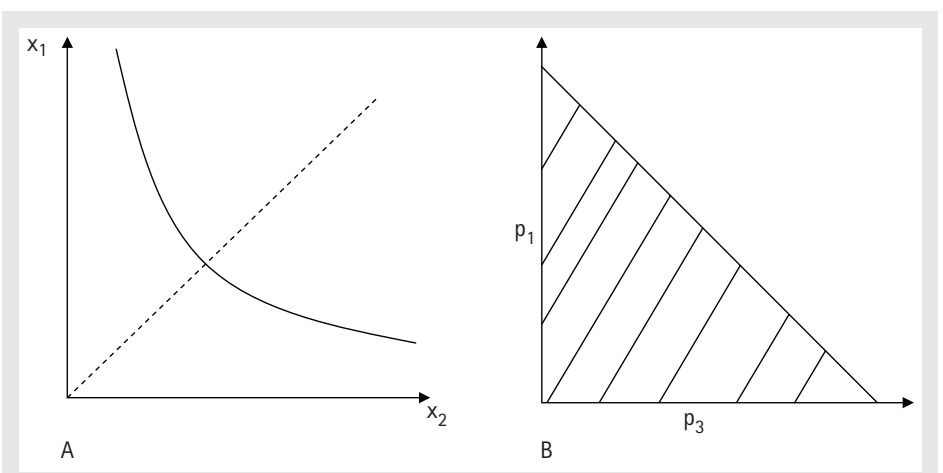

Fig. 3.1. Graphical representations of EU. A: the two-outcome diagram. B: the triangle diagram.

a constant utility $V = p_1u(x_1) + (1- p_1)u(x_2)$ equal to zero yields the slope of indifference curves

$$\frac{dx_1}{dx_2} = \frac{1 - p_1}{p_1}\frac{u'(x_2)}{u'(x_1)}. \tag{3}$$

Indifference curves have a negative slope and are convex if risk aversion is assumed. Moreover, their slope equals the negative probability ratio along the 45° axis which is also called the certainty line since we have $x_1 = x_2$, i.e. the individual is in a riskless position along this line.

In the triangle diagram there are three fixed outcomes, $x_1 > x_2 > x_3$, with varying probabilities. Taking $p_2 = 1- p_1 - p_3$, a fixed utility level is given by $V = p_1u(x_1) + (1 - p_1 - p_3)u(x_2) + p_3u(x_3)$. Solving for the probability of the best outcome, we get the equation for an indifference curve:

$$p_1 = \frac{V - u(x_2)}{u(x_1) - u(x_2)} + p_3\frac{u(x_2) - u(x_3)}{u(x_1) - u(x_2)}. \tag{4}$$

It follows, that indifference curves are upward-sloping parallel lines since the slope is independent of the utility level V. Note that a higher degree of risk aversion leads to steeper indifference curves. Consequently, parallel indifference curves mean that the degree of risk aversion is constant for all prospects consisting of these three outcomes.

The descriptive validity of the independence axiom of EU was first questioned by Allais (1953). A well-known experimental design is the so-called common-ratio effect: there is a choice between p = (1, \$3000) and q = (0.8, \$4000; 0.2, \$0) and a second choice between p^* = (0.25, \$3000; 0.75, \$0) and q^* = (0.2, \$4000; 0.8, \$0). It turns out that many people choose p in the first problem and q^* in the second one, which violates EU. If we normalize utility by u(4000) = 1 and u(0) = 0, choosing p implies $u(3000) > 0.8$ while choosing q^* implies $0.25u(3000) < 0.2$, which is obviously a contradiction.

3.4 Non-Expected Utility

The experimental evidence against the independence axiom has motivated the development of various alternative models; for overviews see Starmer (2000), Schmidt (2004), and Sugden (2004). Two classes of models can be distinguished: utility theories with the betweenness property and rank-dependent models. These classes are disjoint in the sense that only the EU belongs to both. Utility theories with the betweenness property generalize EU by implying that indifference curves in the triangle diagram are also linear but not necessarily parallel. By this generalization the Allais paradox can be explained if indifference curves become steeper for higher

utility levels. This means that subjects are more risk-averse when choosing between "good" prospects than when the choice is between "bad" prospects, which is also called the fanning-out hypothesis (Machina 1982). Formally, betweenness is defined by $p \succ (\sim)\ q \Rightarrow p \succ (\sim)\ \lambda p + (1-\lambda)q \succ (\sim)\ q$ for all $1 > \lambda > 0$.

In the following we will focus on the family of rank-dependent models, since these are currently the most popular in applications. One central model within this family is prospect theory (Tversky and Kahneman 1992, Wakker and Tversky 1993).[1] Prospect theory (PT) differs from EU in three important ways. First, it assumes that decision-makers evaluate outcomes not as final wealth levels but rather as deviations from a status quo, i.e. as gains and losses relative to a reference point. Second, decision-makers are *loss-averse*, which means that a given loss has a greater impact on the desirability of a prospect than a gain of equal size. Third, people do not evaluate probabilities linearly, as in EU, but transform probabilities. Compared with EU, probabilities are replaced by decision weights π_i in all rank-dependent models. These decision weights are constructed by transforming probabilities through a weighting function w. In prospect theory, probability weighting can be different for gains and losses. Altogether, for a prospect consisting of k gains and $n - k$ losses, we have in PT the following representation of preferences (recall that outcomes are rank-ordered from best to worst):

$$V(p) = \sum_{i=1}^{k} v(x_i)\pi_i^+ + \sum_{i=k+1}^{n} v(x_i)\pi_i^-. \tag{5}$$

In this equation, the outcomes x_i are gains and losses relative to a reference point and not final wealth positions as in EU. The decision weights are defined as follows:

$$\pi_i^+ = w^+\Big(\sum_{j=1}^{i} p_j\Big) - w^+\Big(\sum_{j=1}^{i-1} p_j\Big) \tag{6}$$

and

$$\pi_i^- = w^-\Big(\sum_{j=i}^{n} p_j\Big) - w^-\Big(\sum_{j=i+1}^{n} p_j\Big), \tag{7}$$

with both weighting functions strictly increasing and satisfying $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. Note that in the domain of gains decumulative probabilities are transformed, whereas in the domain of losses cumulative probabilities are transformed.

The value function v plays the same role as the utility function u in EU and is strictly increasing. The hypothesis of loss aversion can be captured by assuming that v is steeper in the domain of losses than in the domain of gains. A second important

[1] Prospect theory is sometimes referred to as "cumulative prospect theory", to distinguish it from the original version of prospect theory proposed by Kahneman and Tversky (1979).

hypothesis is diminishing sensitivity, according to which marginal utility decreases as one moves away from the reference point. Consequently, the value function is concave for gains and convex for losses. This leads to the reflection effect, according to which people are often risk-averse for gains and risk-seeking for losses.

Empirical evidence confirms concave utility for gains and convex utility for losses (Tversky and Kahneman 1992; Abdellaoui 2000; Abdellaoui, Bleichrodt, and Paraschiv 2007). There is also a lot of evidence supporting loss aversion, both in the laboratory and in field studies (Camerer 2000; Schmidt and Traub 2002; Pennings and Smidts 2003; Abdellaoui, Bleichrodt, and Paraschiv 2007). Empirical evidence on probability weighting indicates that w has an inverse-S-shaped form, indicating that people are sensitive to changes in probability around 0 (the impossibility effect) and 1 (the certainty effect) and much less so for intermediate probabilities (Tversky and Kahneman 1992; Tversky and Fox 1995; Wu and Gonzalez 1996; Gonzalez and Wu 1999; Abdellaoui 2000; Bleichrodt and Pinto 2000).

The development of PT was influenced by the rank-dependent utility (RDU) model of Quiggin (1982), which differs from EU only by probability weighting. Formally, RDU is given by

$$V(p) = \sum_{i=1}^{k} u(x_i)\pi_i. \tag{8}$$

The construction of the decision weights is identical to that in Eq. 6 with a weighting function w. An interesting special case of RDU is the dual theory (DT) of Yaari (1987) which is given by Eq. 8 with the restriction $u(x_i) = x_i$. Although utility is linear, due to probability weighting we may also have risk aversion in DT. More precisely, a decision-maker in DT exhibits risk aversion if the weighting function is convex. This is because a convex weighting function, compared to untransformed probability, underweighs the probabilities of the best outcomes and overweighs the probabilities of the worst outcomes. In RDU risk aversion can be produced by either a convex weighting function or a concave utility function or both (Chateauneuf and Cohen 1994).

The utility of prospects in a two-outcome diagram for RDU is given by

$$\begin{aligned} V &= w(p_1)u(x_1) + (1 - w(p_1)u(x_2) \quad \text{if } x_1 \succcurlyeq x_2 \\ V &= w(1 - p_1)u(x_2) + (1 - w(1 - p_1))u(x_1) \quad \text{if } x_2 \succ x_1 \end{aligned} \tag{9}$$

Calculating the slope of indifference curves yields

$$\begin{aligned} \frac{dx_1}{dx_2} &= \frac{1 - w(p_1)}{w(p_1)} \frac{u'(x_2)}{u'(x_1)} \qquad \text{if } x_1 \succcurlyeq x_2 \\ \frac{dx_1}{dx_2} &= \frac{w(1 - p_1)}{1 - w(1 - p_1)} \frac{u'(x_2)}{u'(x_1)} \quad \text{if } x_2 \succ x_1 \end{aligned} \tag{10}$$

In case of a convex weighting function we have $1-w(p_1) > w(1-p_1)$ and $w(p_1) < 1-w(1-p_1)$. Consequently, risk aversion implies that indifference curves have a kink at the 45° line. This kink is an important difference between EU and RDU (as well as PT and DT) and can be characterized by the concepts of first-order and second-order risk aversion. Consider a random variable ϵ with $E(\epsilon) = 0$ and variance σ_ϵ^2. From Pratt (1964) it is known that the risk premium RP for avoiding $t\epsilon$ in the case of EU with differentiable utility function can for a sufficiently small t be approximated by $RP \cong -(t^2/2)\,\sigma_\epsilon^2 u''(x)/u'(x)$. The risk premium is thus proportional to t^2 and approaches zero faster than t, which means that for small risks no risk premium will be demanded. This behavior has been termed second-order risk aversion by Segal and Spivak (1990). Second-order risk aversion holds not only for EU with differentiable utility function but for all non-expected utility models which are smooth in the sense of Fréchet-differentiability. If we have a kink along the 45° axis, however, Segal and Spivak (1990) have shown that RP is proportional to t, which is called first-order risk aversion and yields $dRP/dt\,|_{t=0+} \neq 0$. First-order risk aversion implies that a decision-maker will demand a risk premium also for infinitesimally small risks.

3.5 Applications of Non-Expected Utility in Insurance Economics

Insurance economics is a straightforward field for applying non-expected utility. In an important article Machina (1995) has shown that all classical results in insurance economics derived under EU carry over to non-expected utility as long as the representing utility function exhibits second-order risk aversion. In the case of first-order risk aversion, however, some differences occur. We will show this by considering classical results by Mossin (1968), Arrow (1971), and Borch (1960).

Consider an individual with wealth $y > 0$ which is subject to a random loss $\tilde{L}$. If the individual insures the loss, she will receive an indemnity $I(L)$ for paying a premium $P(I(L))$. In the case of coinsurance the indemnity is given by $I(L) = \alpha L$, where α can be chosen by the insured and is between zero and unity. The premium is usually given by $P(I(L)) = (1+\lambda)\alpha E(\tilde{L})$, where $\lambda \geq 0$ is a loading factor for profits and fixed costs of the insurer. A well-known theorem by Mossin (1968) now states that the insured will choose full coverage ($\alpha = 1$) if and only if insurance is fair ($\lambda = 0$). For $\lambda > 0$, partial coverage ($\alpha < 1$) will be chosen. Mossin's theorem can easily be explained in the case where $\tilde{L}$ has only two possible realizations, $L > 0$ with probability p_1 and $L = 0$ with probability $1-p_1$. Final wealth x is now given by $x_1 = y - L + \alpha L - (1+\lambda)\alpha p_1 L$ and $x_2 = y - (1+\lambda)\alpha p_1 L$, where $(1+\lambda)\alpha p_1 L$

is the premium for a coverage of α. Taking differentials with respect to α yields $dx_1 = (1 - (1 + \lambda)p_1)Ld\alpha$ and $dx_2 = -(1 + \lambda)p_1 Ld\alpha$. Consequently, the slope of the budget line in a two-outcome diagram is

$$\frac{dx_1}{dx_2} = -\frac{1 - (1 + \lambda)p_1}{(1 + \lambda)p_1}. \tag{11}$$

If full coverage is demanded, the individual is at a position on his certainty line. We know from Section 3.3 that the slope of indifference curves along the certainty line equals $-(1 - p_1)/p_1$. Full coverage is thus only optimal if $\lambda = 0$, since only then does the slope of the budget line equal the slope of indifference curves along the certainty line.

The demand for coinsurance with non-expected utility preferences was analyzed by, among others, Doherty and Eeckhoudt (1995), Schmidt (1996), Schlesinger (1997), and Segal and Spivak (1990). Recall from Eq. 10 that the slope of indifference curves at the certainty line for RDU equals $-w(1 - p_1)/(1 - w(1 - p))$ for the case $x_2 > x_1$, which is the only relevant case if overinsurance is ruled out. Therefore, full coverage is optimal as long as the indifference curve is flatter than the budget line, i.e.

$$-\frac{1 - (1 + \lambda)p_1}{(1 + \lambda)p_1} \geq \frac{-w(1 - p_1)}{1 - w(1 - p_1)}, \tag{12}$$

which yields $1 + \lambda \leq (1 - w(1 - p_1))/p_1$. In the case of risk aversion w is convex, and therefore $(1 - w(1 - p_1))/p_1 > 1$. Consequently, due to the kink of indifference curves in the case of first-order risk aversion full coverage is also optimal for strictly positive loading factors. This means that Mossin's (1968) theorem carries over only partly to non-expected utility preferences. Note that in the case of DT, indifference curves are linear. Therefore the individual either demands full coverage in case Eq. 12 holds or no coverage at all.

An alternative to coinsurance is deductible insurance. For deductible insurance the indemnity is given by

$$I(L) = \begin{cases} L - d & \text{if } L \geq d \\ 0 & \text{if } L < d. \end{cases} \tag{13}$$

Arrow (1971) has shown that for a given loading factor, deductible insurance is the most preferred form of insurance contract for every insured who is a risk-averse expected utility maximizer. Moreover, it can be shown that the optimal deductible d equals zero if and only if the loading factor equals zero.

Deductible insurance with non-expected utility has been analyzed by Karni (1992), Schlee (1995), and Schlesinger (1997). Karni (1995) has shown that the optimality of deductible insurance carries over to all non-expected utility models which satisfy second-order risk aversion. Schlesinger (1997) has generalized this result by showing that final wealth under every possible insurance contract is a

mean-preserving spread in risk of final wealth under deductible insurance. Consequently, every risk-averse non-expected utility maximizer will also prefer deductible insurance. In contrast to EU, however, the optimal deductible under first-order risk aversion may be equal to zero even for strictly positive loading factors.

Finally, let us analyze efficient risk sharing. Consider n individuals who have to share state-dependent outcomes. For simplicity we assume that there exist only two possible outcomes x_1 and $x_2$with $x_1 > x_2$. Borch (1960) has shown that efficient risk sharing under EU can be characterized by equal marginal rates of substitution; i.e. for any two individuals i and j it must be true that

$$-\frac{1-p_1}{p_1}\frac{u_i'(x_2)}{u_i'(x_1)} = -\frac{1-p_1}{p_1}\frac{u_j'(x_2)}{u_j'(x_1)}, \tag{14}$$

where p_1 is the probability of final wealth level x_1. An important conclusion from this equation is that an efficient risk-sharing agreement leaves each individual with some residual wealth uncertainty. This can be explained as follows: suppose i is in a riskless position. This means that her marginal rate of substitution equals $-(1-p_1)/p_1$. According to Eq. 14, the marginal rate of substitution for all other individuals also has to equal $-(1-p_1)/p_1$; i.e. they are also in a riskless position. But this is impossible if there is aggregate risk.

Schmidt (1996, 1999*a*) has shown that this result does not carry over to first-order risk aversion; i.e. it is possible that some individuals are in a riskless position. Recall from Eq. 10 that the marginal rate of substitution at a riskless position is, in contrast to EU, not identical for all individuals, but determined the probability weighting function. If one individual has a rather convex weighting function, while the weighting function of the other individual is nearly linear, indifference curves may have identical slope at the certainty line of the first individual. In the case of DT, indifference curves are linear, and their slope is determined solely by the weighting function. It turns out that efficient risk sharing here assigns all risk to the least risk-averse individual, while all others enjoy a riskless position.

For further applications of non-expected utility to insurance economics, the reader is referred to Konrad and Skaperdas (1993); Wang, Young, and Panjer (1997); and Schmidt (1999*b*).

3.6 Applications of Non-Expected Utility in Auctions

Due to the increasing importance of auctions in the real world, the literature on auctions has grown rather rapidly in recent years. Since the analysis of many auction

Table 3.1. The standard auction formats

	first-price	second-price
open bids	descending bid auction	ascending bid auction
sealed bids	first-price sealed-bid auction	second-price sealed-bid auction

designs like combinatorial auctions is rather complex even for risk-neutral bidders, applications of non-expected utility are rare in this context. In the present chapter we will focus on auctions of a single object and stick to the independent private values framework. In this framework each bidder i has a private valuation v_i of the auctioned object Z which is not known by the other bidders. Formally, we have $[Z - v_i] \sim \delta_0$, where $[Z - v_i]$ denotes receiving the object Z through the payment of v_i and, as before, δ_0 denotes receiving 0 with certainty. All valuations are drawn from the same distribution over an interval $[v^+, v^-]$.

The basic literature (see Engelbrecht-Wiggans 1980 for a review) assumes risk neutrality of bidders and analyzes four auction formats stated in Table 3.1. In the ascending bid auction, open bidding prevails until no bidder is willing to raise the last bid. It is obvious that the optimal maximal bid of bidder i is v_i, since on the one hand it does not make sense to bid more than v_i and on the other hand it is always possible to make a gain if the highest bid among the other bidders is below v_i. In the second-price sealed-bid auction, each bidder secretly submits a bid to the auctioneer, and the highest bidder wins the auction and has to pay the second-highest bid. Note that the own bid does not determine the price one has to pay (because this is determined by the second-highest bid) but only whether one will buy the object for a given price or not. Since bidder i is willing to buy the object for all prices which do not exceed v_i, the optimal bid is v_i. Consequently, the two second-price auctions are demand revealing, and the revenue of the auctioneer is the second-highest valuation in both cases as long as bid increments are infinitesimally small in the ascending bid auction.

The first-price sealed-bid auction equals the second-price sealed-bid auction, but the highest bidder has to pay his own bid and not only the second-highest bid. In this auction there is no dominant bidding strategy, as the optimal bid is a trade-off between winning probability and profit. If valuations are distributed uniformly, optimal bids in the unique Nash equilibrium are given by $v_i(n-1)/n$ where n is the number of participating bidders. This is also the Nash equilibrium of the descending bid auction, which runs as follows: the auctioneer starts by announcing a prohibitively high price for the object and then continuously decreases the price until one bidder accepts to buy the object for the current price. It is obvious that both first-price auctions are strategically equivalent, because in both cases the bidder determines his bid without knowing any bid of his competitors.

Moreover, the bid is also the price in both cases. Since also the two second-price auctions are strategically equivalent, it remains to compare first-price with second-price auctions. Note that $v_i(n-1)/n$ is precisely the expected value of the second-highest valuation if v_i is the highest valuation. Consequently, the expected price for the bidder and, therefore, the revenue of the auctioneer are identical in all four standard auctions. This well-known revenue equivalence theorem first established by Vickrey (1961) is also valid if valuations are not uniformly distributed. However, if bidders are not risk-neutral but risk-averse expected utility maximizers, optimal bids in the first-price auctions exceed those in second-price auctions (Milgrom and Weber 1982). This can be explained as follows: the optimal strategy in second-price auctions is independent of risk attitude. In first-price auctions, however, the optimal bid is determined in a trade-off between winning probability and potential profit. Risk-averse bidders are willing to solve this trade-off at a higher bid, which increases the winning probability but decreases potential profit.

In the case of non-expected utility let us first analyze the ascending bid auction. Suppose a bidder $j \neq i$ was bidding $v_i - \epsilon$ for an infinitesimally small ϵ, so bidder i has to choose between bidding v_i or quitting the auction. If he quits the auction, the consequence is obviously δ_0. If he bids v_i, he will win the auction with some probability λ and get $[Z - v_i]$. However, with probability $1 - \lambda$, another bidder bids more, and i will also get δ_0. Consequently, a maximal bid of v_i is optimal if $\delta_0 \sim \lambda[Z - v_i] + (1 - \lambda)\delta_0$. Suppose that the auctioned object Z is a lottery, which is often the case in the real world due to uncertainty of the precise quality of the object. Then this indifference is obviously satisfied only if betweenness holds (see the definition of betweenness in Section 3.4). Consider now quasi-concave preferences defined by $p \sim q \Rightarrow \lambda p + (1 - \lambda)q \succ p$. Since $\delta_0 \sim \lambda[Z - v_i]$, quasi-concavity implies $\lambda[Z - v_i] + (1 - \lambda)\delta_0 \succ \delta_0$, and thus the optimal bid exceeds v_i. Quasi-concavity can be interpreted as preference for randomization. This preference causes bidders to stay in the auction even for prices above v_i, since doing so yields a random consequence compared to quitting the auction. This result was first obtained by Karni and Safra (1989*a*, *b*). Analogously, the optimal bid is lower than v_i in the case of quasi-convex preferences defined by $p \sim q \Rightarrow p \succ \lambda p + (1 - \lambda)q$. Karni and Safra (1989*a*, *b*) also analyzed the second-price sealed-bid auction. In this auction the maximal bid is determined without knowing that another bidder will continue bidding until $v_i - \epsilon$. Thus there is a chance that the bidder will get the object for a price much lower than v_i, which is no longer the case in the ascending bid auction if another bidder is already bidding $v_i - \epsilon$. Thus, at the optimal bid the bidder reaches a higher indifference curve than in the ascending bid auction. If, in contrast to EU, the degree of risk aversion may vary for different utility levels, the evaluation of the auctioned lottery, and consequently the optimal bid, may change. More precisely, Grimm and Schmidt (2000) have shown that the optimal bid in the second-price sealed-bid auction is lower that in the ascending

bid auction if the preferences satisfy betweenness and fanning out. Altogether it turns out that the advantage of second-price auctions, i.e. the fact that they elicit true valuations, does not carry over to non-expected utility if the auctioned object is a lottery. In other words, there does not exist an incentive-compatible mechanism to elicit certainty equivalents for non-expected utility models. If, however, a deterministic object is auctioned, second-price auctions elicit true valuations for all preferences which are consistent with first-order stochastic dominance.

First-price auctions with non-expected utility were analyzed by Weber (1982), Karni (1988), and Grimm and Schmidt (2000). Consider a bidder in a descending bid auction whose valuation is already larger than the actual price b. The choice between accepting the actual price and waiting slightly longer is a choice between a sure gain of $[Z - b]$ and the lottery $\lambda[Z - b - \epsilon] + (1 - \lambda)\delta_0$, where λ is the probability that another bidder will accept the price $b - \epsilon$. The optimal bid is thus determined by $[Z - b] \sim \lambda[Z - b - \epsilon] + (1 - \lambda)\delta_0$. In the first-price sealed-bid auction bidder i lacks information that he can make a sure profit, because there is the possibility that another bidder will place a bid higher than v_i. Consequently, at the optimal bid the bidder is on a lower indifference curve. Since only for EU preferences is the degree of risk aversion equal on different indifference curves, we can conclude that optimal bids in the descending bid auction and in the first-price sealed-bid auction are always identical if and only if EU holds. Suppose, in contrast, that preferences are consistent with the fanning-out hypothesis. This means that the degree of risk aversion is higher for the decision in the descending bid auction, which leads to a higher bid.

3.7 Applications of Non-Expected Utility in the Health Domain

3.7.1 Health Insurance

In almost all developed countries health-care expenditures constitute an ever growing share of gross domestic product. In an attempt to curb the rise in health-care expenditures, many countries are considering reforms of their health-care systems. Health insurance has a central role in these reform plans. In the USA, for example, the discussion centers on how universal coverage can be achieved. The Netherlands witnessed a drastic reform of the health insurance system in 2006. The purpose of this reform was to foster competition between health insurers. Analyses of the effects of such reforms are commonly based on expected utility. Two studies have

shown, however, that the predictions of the effects of reforms can be substantially improved when insights from non-expected utility are taken into account.

Marquis and Holmer (1996) reanalyzed the data from the RAND Health Insurance Experiment (HIE), a randomized trial in alternative health insurance plans conducted between 1975 and 1982 (for details about the design of the experiment see Manning *et al.* 1987). Families in the HIE were randomly assigned to one of fourteen different fee-for-service health insurance plans that varied across two dimensions: the coinsurance rate and an upper limit on annual out-of-pocket expenses. They analyzed families' choices for health insurance plans under expected utility and under several non-expected utility hypotheses. They found that allowing for loss aversion, the asymmetric valuation of gains and losses, and risk-seeking behavior for losses significantly improved the fit of the model as compared with expected utility. Marquis and Holmer (1996) did not allow for probability weighting.

The results of Marquis and Holmer (1996) have important implications for the reform of health insurance. Policies like the promotion of increased consumer choice of insurance plans to stimulate competition and the expansion of the choice of employer-provided plans are unlikely to bear much fruit due to the presence of loss aversion. Loss aversion entails inertia in plan choice as consumers tend to stay with their insurance plan. It also implies that a decrease in the price of insurance (e.g. through tax breaks) will increase the demand for generous coverage by more than the demand for generous coverage will fall in response to an increase in the price.

Wakker, Timmermans, and Machielse (2007) analyzed the effects of providing statistical information on insurance purchases using in-depth individual interviews of a large representative sample of the general public. Their data are consistent with prospect theory. They observed much less risk aversion for losses than for gains. Interestingly, they also observed that people became more risk-averse when additional information about probabilities of medical expenditures were provided even if there were no apparent increases in the likelihood of losses. This is consistent with ambiguity-seeking behavior and violates not only expected utility but also the common assumption in the theoretical economics literature that people are ambiguity-averse; i.e. they do not like decisions in which probabilities are unknown. The data of Wakker, Timmermans, and Machielse (2007) suggest that no special aversion to unknown probabilities holds in real-life situations as long as uncertainty about the probabilities is natural in a decision situation.

3.7.2 Medical Decision-Making

Utility theory is widely applied in medical decision-making. The common way to evaluate new medical technologies is through cost–utility analysis in which the

benefits of these technologies are expressed in terms of utility. The most popular utility model is the quality-adjusted life-years (QALY) model. The QALY model combines the two dimensions of health, life duration and health status, in a single index number and claims that the utility of T years in health state Q is equal to

$$U(Q,T) = V(Q) \cdot T, \tag{15}$$

where V(Q) denotes a weight that reflects the utility (or attractiveness) of health state Q. QALYs play an important role in health policy in many countries. For example, in the UK the National Institute for Clinical Excellence (NICE) requires a cost–utility analysis based on QALYs before a treatment is eligible for inclusion in the National Health Service (NHS). In the Netherlands, the Council for Public Health and Care, the main advisory board of the Dutch government on health policy, recently recommended that only treatments that cost less than €80,000 per QALY gained should be included in the basic insurance package. Treatments costing more can be insured only through supplementary insurance.

Sometimes a more general form of the QALY model is proposed in which the utility for life duration is not linear, as in Eq. 15, but can be curved:

$$U(Q,T) = V(Q) \cdot W(T), \tag{16}$$

where W(T) denotes the utility for life duration. In what follows we will refer to Eq. 16 as the *nonlinear QALY model*, to distinguish it from Eq. 15.

QALYs have two important advantages: they are tractable, which makes them attractive for practical applications, and they are intuitive—one QALY can be interpreted as one year in good health—which makes them easy to communicate to policymakers. There are, however, also important methodological questions surrounding QALYs. In this chapter we will focus on two of these and, in particular, on the insights that non-expected utility has offered to solve these methodological questions. The first question relates to the validity of the QALY model. The QALY model is a simple model, which, as explained above, has clear advantages. But this simplicity may also have a price: the QALY model could be too simple and may misrepresent people's preferences for health. To obtain insight into the descriptive validity of the QALY models, Eqs. 15 and 16, we need behavioral foundations that identify the conditions on which the models depend. These conditions can then be tested in experimental studies. As we will explain, non-expected utility has been very useful in designing robust tests of the QALY models.

A second question on which we will focus is the estimation of the utilities V(Q) and W(T). The measures commonly used to measure V(Q) and W(T) yield systematically different results, which cannot be explained under expected utility. The insights from non-expected utility, most notably prospect theory, can help to reconcile some of these differences, as we will show in what follows.

3.7.2.1 *Non-Expected Utility and Tests of the Descriptive Validity of QALYs*

Pliskin, Shepard, and Weinstein (1980) were the first to give a behavioral foundation for the QALY model. Their model was later simplified by Bleichrodt, Wakker, and Johannesson (1997) and Miyamoto *et al.* (1998). Bleichrodt *et al.* (1997) and Miyamoto *et al.* (1998) showed that the crucial condition of the QALY model is that people be risk-neutral with respect to life duration. That is, for a given health quality Q they should be indifferent between a risky treatment that gives life duration T_1 with probability p and life duration T_2 with probability $1 - p$ and $p^*T_1 + (1 - p)^*T_2$ for sure. Empirical evidence has generally shown that people do not behave according to this condition, but are risk-averse with respect to life duration. For example, the median subject in Stiggelbout *et al.* (1994) was indifferent between 4 years for sure and a risky treatment, giving 10 years with probability 1/2 and 0 years (death) with probability 1/2.

The analyses of Pliskin *et al.* (1980), Bleichrodt *et al.* (1997), and Miyamoto *et al.* (1998) relied crucially on the assumption that people behave according to expected utility. Without this assumption, their behavioral foundations are no longer true. For example, under rank-dependent utility it is very possible that people have linear utility and are risk-neutral with respect to life duration. Consider, for example, the median preference observed by Stiggelbout *et al.* (1994). If $w(1/2) = 0.40$, then this response is consistent with a linear utility for life duration.

As mentioned before in Section 3.3, evidence abounds that people violate expected utility. These violations of expected utility cast doubt on the validity of previous tests of the QALY model. Several authors have derived tests of the QALY model that are robust to violations of expected utility. Bleichrodt and Quiggin (1997) derived a test of the QALY model that is valid under a large class of non-expected utility models, including rank-dependent utility. Recall that $(p, (Q_1, T_1); (Q_2, T_2))$ denotes the risky prospect that gives (Q_1, T_1) with probability p and (Q_2, T_2) with probability $1-$ p. As before, we assume that all prospects are *rank-ordered,* i.e. $(Q_1, T_1) \succcurlyeq (Q_2, T_2)$. The condition that Bleichrodt and Quiggin (1997) imposed, *constant marginal* utility, says that for all Q and for all ϵ small enough that the prospects involved are still rank-ordered and the life durations do not exceed the maximum possible life duration, $(p,(Q, T_1); (Q, T_2)) \sim (p, (Q, T_3); (Q, T_4))$ iff $(p, (Q, T_1 + \epsilon); (Q, T_2)) \sim (p, (Q, T_3 + \epsilon); (Q, T_4))$ and $(p, (Q, T_1); (Q, T_2)) \sim (p, (Q, T_3); (Q, T_4))$ iff $(p, (Q, T_1); (Q, T_2 + \epsilon)) \sim (p, (Q, T_3); (Q, T_4 + \epsilon))$. Constant marginal utility was tested and rejected by Bleichrodt and Pinto (2005). Bleichrodt and Miyamoto (2003) extended the analysis of Bleichrodt and Quiggin (1997) to prospect theory, where outcomes can be both gains and losses. Miyamoto (1999) proposed another condition, constant proportional coverage, that allows testing of the QALY model under rank-dependent utility. *Constant proportional coverage* holds if for all Q and for all $T_1 > T_2 > T_3$ and $T_1' > T_2' > T_3'$,

whenever $(Q, T_2) \sim (p, (Q, T_1); (Q, T_3))$, $(Q, T'_2) \sim (p', (Q, T'_1); (Q, T('_3)))$, and $(T_2 - T_3)/(T_1 - T_3) = (T'_2 - T'_3)/(T'_1 - T'_3)$, then $p = p'$. Doctor *et al.* (2004) showed that the condition is also valid under prospect theory if a plausible assumption about the location of the reference point is made. They tested constant proportional coverage and obtained support for it.

Miyamoto and Eraker (1988) were the first to test the nonlinear QALY model under a general utility theory, and obtained support for it. Bleichrodt and Pinto (2005) considered an even more general utility theory and also obtained support for the nonlinear QALY model.

In summary, the insights from non-expected utility have helped to perform more robust tests of the QALY model. It is more plausible that people have linear utility for life duration under non-expected utility models because under these models risk attitude is captured not only by the utility function, as in expected utility, but also by other functions and parameters. For example, in rank-dependent utility part of people's risk attitude is captured by the probability weighting, and in prospect theory loss aversion plays an important part in the explanation of attitudes towards risk as well. The available evidence on QALYs under non-expected utility is still limited, but it indicates that the QALY model may have been too easily dismissed. The QALY model may describe people's preferences for health better than is commonly thought. This is of course an important finding given the dominant role that the QALY model plays in practical health economics research.

3.7.2.2 *Non-Expected Utility and the Measurement of Health State Utilities*

Non-expected utility theory also has implications for the main methods to determine $V(Q)$ and $W(T)$. One of the most widely used methods to measure the health state utilities $V(Q)$ is the *standard gamble*. The standard gamble is, for example, used in the SF-6D, a very popular valuation method in applied health economics (Brazier, Roberts, and Deverill 2002). In the standard gamble method people face a choice between an impaired health state Q for T years for sure and a risky treatment option that gives full health with probability p and death with probability $1 - p$.[2] The purpose of the standard gamble is to determine the probability p that leads to indifference between these two options. Under expected utility and the nonlinear QALY model, it then follows that $V(Q) = p$.

It is well known that the standard gamble gives systematically higher utilities than other methods to determine health state utilities that do not involve risk. This obviously raises the question as to which method should be preferred. The traditional view was that the standard gamble should be the preferred method,

[2] In principle it is not necessary to use full health and death as the outcomes of the risky treatment, but this is the way the standard gamble is commonly asked.

as it is based on expected utility. The idea was that medical decision analysis is a prescriptive exercise, and that expected utility is a prescriptive theory, and hence health utility measurement should be based on expected utility. The problem with this point of view is that the measurement of utility is a descriptive exercise, and that if people do not behave according to expected utility, then utilities that are derived under expected utility will be biased. Using biased utilities in cost–utility analysis runs the risk of wrong recommendations for health policy.

Conclusive evidence of such biases was observed by Llewellyn-Thomas *et al.* (1982). They used two different ways to ask the standard gamble. The first way was as described above, with full health and death as endpoints in the risky treatment option. We will refer to this format as the *direct method*. In the *chained method* they first established indifference between (Q, T) for sure and a risky treatment (q, (full health, T); (Q′, T)), where Q′ is a worse health state than Q, and then they established indifference between (Q′, T) for sure and a risky treatment (r, (full health, T); death). Under expected utility the first indifference in the chained method entails

$$V(Q) = q + (1 - q)^{*}V(Q'). \tag{17}$$

The second indifference implies $V(Q') = r$, and hence the chained method gives $V(Q) = q + (1 - q)^{*}r$. Except for random error, we should therefore observe that $p = q + (1 - q)^{*}r$ when expected utility holds. However, Llewellyn-Thomas *et al.* observed that the chained method led to systematically higher utilities V(Q) (see also Rutten-van Mölken *et al.* 1995; Bleichrodt 2001), a clear violation of expected utility casting doubt on the validity of standard gamble measurements.

To determine W(T), two methods are commonly used. In the *probability equivalence (PE) method*, people are asked for the probability p that makes them indifferent between T_2 years in some health state Q for sure and a risky treatment that gives probability p of T_1 years in health state Q and probability $1 - p$ of T_3 years in health state Q. In the *certainty equivalence (CE) method*, people are asked for the number of years S_2 in some health state Q for sure and a risky treatment that gives probability q of T_1 years in health state Q and probability $1 - q$ of T_3 years in health state Q. Of course, if we use the same T_1 and T_3 in the PE and the CE, and if we substitute the response from the PE in the CE (i.e. $p = q$),[3] then we should observe that $S_2 = T_2$ under expected utility. In fact, what is typically observed is that $S_2 > T_2$,[4] which leads to a more concave utility for life duration under the PE method than under the CE method, and which obviously violates expected utility (Bleichrodt, Pinto, and Wakker 2001).

Several authors have tried to solve the above inconsistencies by using non-expected utility. Wakker and Stiggelbout (1995) explored the impact of correcting

[3] Or for that matter substitute the response from the CE in the PE (i.e. $T_2 = S_2$).

[4] For money outcomes this was already observed by Hershey and Schoemaker (1985).

the standard gamble for probability weighting as in rank-dependent utility. They showed that if people have an inverse-S-shaped probability weighting function, then the resulting utilities are generally pushed downwards, leading to more consistency with other utility measurement methods. Their conjectures were confirmed by two empirical studies. Bleichrodt, van Rijn, and Johannesson (1999) observed that correcting the standard gamble for inverse-S-shaped probability weighting leads to utilities that are more consistent with people's preferences than using the uncorrected standard gamble utilities. Bleichrodt and Pinto (2000) developed a new parameter-free method to examine the shape of the probability weighting function and tested their method in a medical setting. They found that the probability weighting function was indeed inverse-S-shaped. On the other hand, Bleichrodt (2001) showed that correcting the standard gamble for inverse-S-shaped probability weighting did not resolve the difference between the direct version of the standard gamble and the chained version of the standard gamble that was first observed by Llewellyn-Thomas *et al.* (1982) but, instead, exacerbated this difference. Correcting for probability weighting cannot resolve the difference between the PE and the CE either, even though the difference tends to be mitigated (Bleichrodt, Pinto, and Wakker 2001). Finally, Stalmeier and Bezembinder (1999) observed that correction for probability weighting reduced the difference between risky and riskless utilities for life duration but was not sufficient to resolve the difference.

The missing element in the corrections by Wakker and Stiggelbout (1995) was that they did not account for loss aversion, the other main deviation from expected utility modeled by prospect theory. Bleichrodt, Pinto, and Wakker (2001) derived new formulae that allow the measurement of health state utilities under prospect theory. The crucial step in their analysis was the location of the reference point. Bleichrodt, Pinto, and Wakker (2001) conjectured that in the standard gamble and in the PE method people take the sure outcome as their reference point, because this outcome is given. In the CE method, however, the sure outcome has to be determined, and it is therefore unlikely to serve as the reference point. People will adopt a reference point in the CE, but it will not be the sure outcome. Their data confirmed these hypotheses. Bleichrodt, Pinto, and Wakker (2001) tested whether their formulae could explain the discrepancy between the PE and the CE methods, and their results were encouraging: the corrections made the discrepancy vanish. Further support for the hypotheses in Bleichrodt, Pinto, and Wakker (2001) comes from several studies in the literature, some of which recorded people's thought processes in responding to PE and CE questions (Stalmeier and Bezembinder 1999; Morrison 2000; Robinson, Loomes, and Jones-Lee 2001; van Osch *et al.* 2004; van Osch, van den Hout, and Stiggelbout 2006). Bleichrodt *et al.* (2007) extended the analysis of Bleichrodt, Pinto, and Wakker (2001) to other utility measurement procedures and found that prospect theory clearly performed better than expected utility and rank-dependent utility and solved many inconsistencies that were observed under expected utility.

Oliver (2003) observed, however, that the corrections of Bleichrodt, Pinto, and Wakker (2001) could not entirely explain the differences between the direct version of the standard gamble and the chained version. He found that the best-performing model was a version of prospect theory in which there was loss aversion but no probability weighting. Nevertheless, Stalmeier (2002) found evidence that the difference between the direct version of the standard gamble and the chained version is produced by a more elementary violation of rationality than probability weighting or loss aversion. He observed that people tended to give the same probability in the direct version and to both questions in the chained version, i.e. $p = q = r$. That is, people anchor on their response and do not adjust it for differences in health quality. Such basic violations of rationality are hard to accommodate in any formal theory of decision under risk.

In summary, the insights from prospect theory seem to have improved the measurement of utility in the health domain. Inconsistencies are significantly reduced when corrections for probability weighting and loss aversion are applied. An important implication of these findings is that the standard gamble as it is commonly used will result in utilities that are far too high. As was shown by Bleichrodt, Pinto, and Wakker (2001), under the common findings in the literature (inverse-S-shaped probability weighting and losses looming larger than gains) uncorrected standard gamble utilities are clearly biased upwards. In particular, there are serious reasons to suspect that the widely used SF-6D method produces biased utilities, and care should be taken in applying this algorithm.

References

Abdellaoui, M. (2000). Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science*, 46, 1497–512.

—— Bleichrodt, H., and Paraschiv, C. (2007). Measuring Loss Aversion under Prospect Theory: A Parameter-Free Approach. *Management Science*, 53, 1659–74.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica*, 21, 503–46.

Arrow, K. J. (1971). *Essays in the Theory of Risk-Bearing*. Amsterdam: North-Holland.

Bleichrodt, H. (2001). Probability Weighting in Choice under Risk: An Empirical Test. *Journal of Risk and Uncertainty*, 23, 185–98.

—— and Miyamoto, J. (2003). A Characterization of Quality-Adjusted Life-Years under Cumulative Prospect Theory. *Mathematics of Operations Research*, 28, 181–93.

—— and Pinto, J. L. (2000). A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. *Management Science*, 46, 1485–96.

—— —— (2005). The Validity of QALYs under Non-Expected Utility. *Economic Journal*, 115, 533–50.

—— and Quiggin, J. (1997). Characterizing QALYs under a General Rank Dependent Utility Model. *Journal of Risk and Uncertainty*, 15, 151–65.

——— Pinto, J. L., and Wakker, P. P. (2001). Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility. *Management Science*, 47, 1498–514.
——— van Rijn, J., and Johannesson, M. (1999). Probability Weighting and Utility Curvature in QALY Based Decision Making. *Journal of Mathematical Psychology*, 43, 238–60.
——— Wakker, P. P., and Johannesson, M. (1997). Characterizing QALYs by Risk Neutrality. *Journal of Risk and Uncertainty*, 15, 107–14.
——— Abellan, J. M., Pinto, J. L., and Mendez., I. (2007). Resolving Inconsistencies in Utility Measurement under Risk: Tests of Generalizations of Expected Utility. *Management Science*, 53, 469–82.
Borch, K. (1960). The Safety Loading of Reinsurance Premiums. *Skandinavisk Aktuarietskrift*, 153–84.
Brazier, J., Roberts, J., and Deverill, M. (2002). The Estimation of a Preference-Based Measure of Health from the SF-36. *Journal of Health Economics*, 21, 271–92.
Camerer, C. F. (2000). Prospect Theory in the Wild: Evidence from the Field. In D. Kahneman and A. Tversky (eds.), *Choices, Values and Frames*, 288–300. New York: Cambridge University Press.
Chateauneuf, A., and Cohen, M. (1994). Risk Seeking with Diminishing Marginal Utility in a Non-Expected Utility Model. *Journal of Risk and Uncertainty*, 9, 77–91.
Doctor, J. N., Bleichrodt, H., Miyamoto, J., Temkin, N. R., and Dikmen, S. (2004). A New and More Robust Test of QALYs. *Journal of Health Economics*, 23, 353–67.
Doherty, N. A., and Eeckhoudt, L. (1995). Optimal Insurance without Expected Utility: The Dual Theory and the Linearity of Insurance Contracts. *Journal of Risk and Uncertainty*, 10, 157–79.
Ellsberg, D. (1961). Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643–69.
Engelbrecht-Wiggans, R. (1980). Auctions and Bidding Models: A Survey. *Management Science*, 26, 119–42.
Gonzalez, R., and Wu, G. (1999). On the Form of the Probability Weighting Function. *Cognitive Psychology*, 38, 129–66.
Grimm, V., and Schmidt, U. (2000). Equilibrium Bidding without the Independence Axiom: A Graphical Analysis. *Theory and Decision*, 49, 361–74.
Hershey, J. C., and Schoemaker, P. J. H. (1985). Probability versus Certainty Equivalence Methods in Utility Measurement: Are they Equivalent? *Management Science*, 31, 1213–31.
Kahneman, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263–91.
Karni, E. (1988). On the Equivalence between Descending Bid Auctions and First Price Sealed Bid Auctions. *Theory and Decision*, 25, 211–17.
——— (1992). Optimal Insurance: A Nonexpected Utility Analysis. In G. Dionne (ed.), *Contributions to Insurance Economics*, 217–38. Boston: Kluwer.
——— (1995). Non-Expected Utility and the Robustness of the Classical Insurance Paradigm: Discussion. *Geneva Papers on Risk and Insurance Theory*, 20, 51–6.
——— and Safra, Z. (1989*a*). Ascending Bid Auctions with Behaviorally Consistent Bidders. *Annals of Operations Research*, 19, 435–46.
——— ——— (1989*b*). Dynamic Consistency, Revelations in Auctions and the Structure of Preferences. *Review of Economic Studies*, 56, 421–34.
Konrad, K. A., and Skaperdas, S. (1993). Self-Insurance and Self-Protection: A Nonexpected Utility Analysis. *Geneva Papers on Risk and Insurance Theory*, 18, 131–46.

Llewellyn-Thomas, H., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., and Boyd., N. F. (1982). The Measurement of Patients' Values in Medicine. *Medical Decision Making*, 2, 449–62.

Machina, M. (1982). 'Expected Utility' Analysis without the Independence Axiom. *Econometrica*, 50, 277–323.

—— (1995). Non-Expected Utility and the Robustness of the Classical Insurance Paradigm. *Geneva Papers on Risk and Insurance Theory*, 20, 9–50.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A., and Marquis, M. S. (1987). Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. *American Economic Review*, 77, 251–77.

Marquis, M. S., and Holmer, M. R. (1996). Alternative Models of Choice under Uncertainty and Demand for Health Insurance. *Review of Economics and Statistics*, 78, 421–7.

Milgrom, P. R., and Weber, R. J. (1982). A Theory of Auctions and Competitive Bidding. *Econometrica*, 50, 1089–122.

Miyamoto, J. M. (1999). Quality-Adjusted Life-Years (QALY) Utility Models under Expected Utility and Rank Dependent Utility Assumptions. *Journal of Mathematical Psychology*, 43, 201–37.

—— and Eraker, S. A. (1988). A Multiplicative Model of the Utility of Survival Duration and Health Quality. *Journal of Experimental Psychology: General*, 117, 3–20.

—— Wakker, P. P., Bleichrodt, H., and Peters, H. J. M. (1998). The Zero-Condition: A Simplifying Assumption in QALY Measurement and Multiattribute Utility. *Management Science*, 44, 839–49.

Morrison, G. C. (2000). The Endowment Effect and Expected Utility. *Scottish Journal of Political Economy*, 47, 183–97.

Mossin, J. (1968). Aspects of Rational Insurance Purchasing. *Journal of Political Economy*, 76, 553–68.

Oliver, A. J. (2003). The Internal Consistency of the Standard Gamble: Tests after Adjusting for Prospect Theory. *Journal of Health Economics*, 22, 659–74.

Pennings, J. M. E., and Smidts, A. (2003). The Shape of Utility Functions and Organizational Behavior. *Management Science*, 49, 1251–63.

Pliskin, J. S., Shepard, D. S., and Weinstein, M. C. (1980). Utility Functions for Life Years and Health Status. *Operations Research*, 28, 206–23.

Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica*, 32, 83–98.

Quiggin, J. (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization*, 3, 323–43.

Robinson, A., Loomes, G., and Jones-Lee, M. (2001). Visual Analog Scales, Standard Gambles, and Relative Risk Aversion. *Medical Decision Making*, 21, 17–27.

Rutten-van Mölken, M. P., Bakker, C. H., van Doorslaer, E. K. A., and van der Linden, S. (1995). Methodological Issues of Patient Utility Measurement. Experience from Two Clinical Trials. *Medical Care*, 33, 922–37.

Schlee, E. E. (1995). The Comparative Statics of Deductible Insurance in Expected and Non-Expected Utility Theories. *Geneva Papers on Risk and Insurance Theory*, 20, 57–72.

Schlesinger, H. (1997). Insurance Demand without the Expected-Utility Paradigm. *Journal of Risk and Insurance*, 64, 19–39.

Schmidt, U. (1996). Demand for Coinsurance and Bilateral Risk-Sharing with Rank-Dependent Utility. *Risk, Decision and Policy*, 1, 217–28.

——(1999*a*). Efficient Risk-Sharing and the Dual Theory of Choice under Risk. *Journal of Risk and Insurance*, 66, 597–608.

——(1999*b*). Moral Hazard and First-Order Risk Aversion. *Journal of Economics*, Supplement 8, 167–79.

——(2004). Alternatives to Expected Utility: Some Formal Theories. In S. Barbéra, P. J. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory*, ii. 757–838. Dordrecht: Kluwer.

—— and Traub, S. (2002). An Experimental Test of Loss Aversion. *Journal of Risk and Uncertainty*, 25, 233–49.

Segal, U., and Spivak, A. (1990). First Order versus Second Order Risk Aversion. *Journal of Economic Theory*, 51, 111–25.

Stalmeier, P. F. M. (2002). Discrepancies between Chained and Classic Utilities Induced by Anchoring with Occasional Adjustments. *Medical Decision Making*, 22, 53–64.

—— and Bezembinder, T. G. G. (1999). The Discrepancy between Risky and Riskless Utilities: A Matter of Framing? *Medical Decision Making*, 19, 435–47.

Starmer, C. (2000). Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk. *Journal of Economic Literature*, 28, 332–82.

Stiggelbout, A. M., Kiebert, G. M., Kievit, J., Leer, J. W. H., Stoter, G., and de Haes, J. C. J. M. (1994). Utility Assessment in Cancer Patients: Adjustment of Time Tradeoff Scores for the Utility of Life Years and Comparison with Standard Gamble Scores. *Medical Decision Making*, 14, 82–90.

Sugden, R. (2004). Alternatives to Expected Utility. In S. Barbéra, P. J. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory*, ii. 685–755. Dordrecht: Kluwer.

Tversky, A., and Fox, C. (1995). Weighing Risk and Uncertainty. *Psychological Review*, 102, 269–83.

—— and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.

van Osch, S. M. C., van den Hout, W. B., and Stiggelbout, A. M. (2006). Exploring the Reference Point in Prospect Theory: Gambles for Length of Life. *Medical Decision Making*, 26, 338–46.

—— Wakker, P. P., van den Hout, W. B., and Stiggelbout, A. M. (2004). Correcting Biases in Standard Gamble and Time Tradeoff Utilities. *Medical Decision Making*, 24, 511–17.

Vickrey, W. (1961). Counter Speculation, Auctions and Competitive Sealed Tenders. *Journal of Finance*, 16, 8–37.

von Neumann, J., and Morgenstern, O. (1944). *The Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wakker, P. P., and Stiggelbout, A. M. (1995). Explaining Distortions in Utility Elicitation through the Rank-Dependent Model for Risky Choices. *Medical Decision Making*, 15, 180–6.

—— and Tversky, A. (1993). An Axiomatization of Cumulative Prospect Theory. *Journal of Risk and Uncertainty*, 7, 147–76.

—— Timmermans, D. R. M., and Machielse, I. (2007). The Effects of Statistical Information on Risk- and Ambiguity-Attitudes, and on Rational Insurance Decisions. *Management Science*, 53, 1770–84.

Wang, S. S., Young, V. R., Panjer, H. H. (1997). Axiomatic Characterization of Insurance Prices. *Insurance: Mathematics and Economics*, 21, 173–83.

Weber, R. (1982). The Allais Paradox, Dutch Auctions, and Alpha-Utility Theory. Working Paper 536, Kellogg Graduate School of Management, Northwestern University, Evanston, IL.

Wu, G., and Gonzalez, R. (1996). Curvature of the Probability Weighting Function. *Management Science*, 42, 1676–90.

Yaari, M. E. (1987). The Dual Theory of Choice under Risk. *Econometrica*, 55, 95–115.

CHAPTER 4

AMBIGUITY

JÜRGEN EICHBERGER
DAVID KELSEY

4.1 Introduction

Most economic decisions are made under uncertainty. Decision-makers are often aware of variables which will influence the outcomes of their actions but which are beyond their control. The quality of their decisions depends, however, on predicting these variables as correctly as possible. Long-term investment decisions provide typical examples, since their success is also determined by uncertain political, environmental, and technological developments over the lifetime of the investment. In this chapter we review recent work on decision-makers' behavior in the face of such risks and the implications of these choices for economics and public policy. Over the past fifty years, decision-making under uncertainty was mostly viewed as choice over a number of prospects each of which gives rise to specified outcomes with known probabilities. Actions of decision-makers were assumed to lead to well-defined probability distributions over outcomes. Hence, choices of actions could be identified with choices of probability distributions. The expected utility paradigm (see Chapter 1) provides a strong foundation for ranking probability distributions over outcomes while taking into account a decision-maker's subjective risk preference. Describing uncertainty by probability distributions, expected utility theory could also use the powerful methods of statistics. Indeed, many of the theoretical achievements in economics over the past five decades are due to the successful application of the expected-utility approach to economic problems in finance and information economics.

At the same time, criticism of the expected utility model has arisen on two accounts. On the one hand, following Allais's seminal (1953) article, more and more experimental evidence was accumulated contradicting the expected utility decision criterion, even in the case where subjects had to choose among prospects with controlled probabilities (compare Chapters 2 and 3). On the other hand, in practice, for many economic decisions the probabilities of the relevant events are not obviously clear. This chapter deals with decision-making when some or all of the relevant probabilities are unknown.

In practice, nearly all economic decisions involve unknown probabilities. Indeed, situations where probabilities are known are relatively rare and are confined to the following cases:

1. *Gambling*. Gambling devices, such as dice, coin-tossing, roulette wheels, etc., are often symmetric, which means that probabilities can be calculated from relative frequencies with a reasonable degree of accuracy.[1]
2. *Insurance*. Insurance companies usually have access to actuarial tables which give them fairly good estimates of the relevant probabilities.[2]
3. *Laboratory experiments*. Researchers have artificially created choices with known probabilities in laboratories.

Many current policy questions concern ambiguous risks: for instance, how to respond to threats from terrorism and rogue states, and the likely impact of new technologies. Many environmental risks are ambiguous, due to limited knowledge of the relevant science and because outcomes will be seen only many decades from now. The effects of global warming and the environmental impact of genetically modified crops are two examples. The hurricanes which hit Florida in 2004 and the tsunami of 2004 can also be seen as ambiguous risks. Although these events are outside human control, one can ask whether the economic system can or should share these risks among individuals.

Even if probabilities of events are unknown, this observation does not preclude that individual decision-makers may hold beliefs about these events which can be represented by a subjective probability distribution. In a path-breaking contribution to the theory of decision-making under uncertainty, Savage (1954) showed that one can deduce a unique subjective probability distribution over events with unknown probabilities from a decision-maker's choice behavior if it satisfies certain axioms. Moreover, this decision-maker's choices maximize an expected utility functional of state-contingent outcomes, where the expectation is taken with respect to this subjective probability distribution. Savage's (1954) Subjective Expected Utility (SEU) theory offers an attractive way to continue working with

[1] The fact that most people prefer to bet on symmetric devices is itself evidence for ambiguity aversion.

[2] However, it should be noted that many insurance contracts contain an 'act of God' clause declaring the contract void if an ambiguous event happens. This indicates some doubts about the accuracy of the probability distributions gleaned from the actuarial data.

the expected utility approach even if the probabilities of events are unknown. SEU can be seen as a decision model under uncertainty with unknown probabilities of events where, nevertheless, agents whose behavior satisfies the Savage axioms can be modeled as expected utility maximizers with a subjective probability distribution over events. Using the SEU hypothesis in economics, however, raises some difficult questions about the consistency of subjective probability distributions across different agents. Moreover, the behavioral assumptions necessary for a subjective probability distribution are not supported by evidence, as the following section will show.

Before proceeding, we shall define terms. The distinction of *risk* and *uncertainty* can be attributed to Knight (1921). The notion of *ambiguity,* however, is probably due to Ellsberg (1961). He associates it with the lack of information about relative likelihoods in situations which are characterized neither by risk nor by complete uncertainty. In this chapter, *uncertainty* will be used as a generic term to describe all states of information about probabilities. The term *risk* will be used when the relevant probabilities are known. *Ambiguity* will refer to situations where some or all of the relevant information about probabilities is lacking. Choices are said to be *ambiguous* if they are influenced by events whose probabilities are unknown or difficult to determine.

4.2 Experimental Evidence

There is strong evidence which indicates that, in general, people do not have subjective probabilities in situations involving uncertainty. The best-known examples are the experiments of the *Ellsberg paradox.*[3]

Example 4.2.1. (Ellsberg 1961) *Ellsberg paradox I: three-color urn experiment*
There is an urn which contains ninety balls. The urn contains thirty red balls (**R**), and the remainder are known to be either black (**B**) or yellow (**Y**), but the number of balls which have each of these two colors is unknown. One ball will be drawn at random.

Consider the following bets: (**a**) "Win 100 if a red ball is drawn", (**b**) "Win 100 if a black ball is drawn", (**c**) "Win 100 if a red or yellow ball is drawn", (**d**) "Win 100 if a black or yellow ball is drawn". This experiment may be summarized as follows:

[3] Notice that these experiments provide evidence not just against SEU but against all theories which model beliefs as additive probabilities.

		30	60	
		R	**B**	**Y**
Choice 1: "Choose either bet **a** or bet **b**".	**a**	100	0	0
	b	0	100	0
Choice 2: "Choose either bet **c** or bet **d**".	**c**	100	0	100
	d	0	100	100

Ellsberg (1961) offered several colleagues these choices. When faced with them most subjects stated that they preferred **a** to **b** and **d** to **c**.

It is easy to check algebraically that there is no subjective probability, which is capable of representing the stated choices as maximizing the expected value of any utility function. In order to see this, suppose to the contrary that the decision-maker does indeed have a subjective probability distribution. Then, since (s)he prefers **a** to **b** (s)he must have a higher subjective probability for a red ball being drawn than for a black ball. But the fact that (s)he prefers **d** to **c** implies that (s)he has a higher subjective probability for a black ball being drawn than for a red ball. These two deductions are contradictory.

It is easy to come up with hypotheses which might explain this behavior. It seems that the subjects are choosing gambles where the probabilities are "better known". Ellsberg (1961, p. 657) suggests the following interpretation:

> Responses from confessed violators indicate that the difference is not to be found in terms of the two factors commonly used to determine a choice situation, the relative desirability of the possible pay-offs and the relative likelihood of the events affecting them, but in a third dimension of the problem of choice: the nature of one's information concerning the relative likelihood of events. What is at issue might be called the *ambiguity* of information, a quality depending on the amount, type, reliability and "unanimity" of information, and giving rise to one's degree of "confidence" in an estimate of relative likelihoods.

The Ellsberg experiments seem to suggest that subjects avoid the options with unknown probabilities. Experimental studies confirm a preference for betting on events with information about probabilities. Camerer and Weber (1992) provide a comprehensive survey of the literature on experimental studies of decision-making under uncertainty with unknown probabilities of events. Based on this literature, they view ambiguity as "uncertainty about probability, created by missing information that is relevant and could be known" (Camerer and Weber 1992, p. 330).

The concept of the *weight of evidence*, advanced by Keynes (2004[1921]) in order to distinguish the probability of an event from the evidence supporting it, appears closely related to the notion of ambiguity arising from

known-to-be-missing information (Camerer 1995, p. 645). As Keynes (2004[1921], p. 71) wrote: "New evidence will sometimes decrease the probability of an argument, but it will always increase its *weight*." The greater the weight of evidence, the less ambiguity a decision-maker experiences.

If ambiguity arises from missing information or lack of evidence, then it appears natural to assume that decision-makers will dislike ambiguity. One may call such attitudes *ambiguity-averse*. Indeed, as Camerer and Weber (1992) summarize their findings, "ambiguity aversion is found consistently in variants of the Ellsberg problems" (p. 340).

There is a second experiment supporting the Ellsberg paradox which sheds additional light on the sources of ambiguity.

Example 4.2.2. (Ellsberg 1961) *Ellsberg paradox II: two-urn experiment*
There are two urns which contain 100 black (**B**) or red (**R**) balls. Urn 1 contains 50 black balls and 50 red balls. For Urn 2 no information is available. From both urns one ball will be drawn at random.

Consider the following bets: (**a**) "Win 100 if a black ball is drawn from Urn 1", (**b**) "Win 100 if a red ball is drawn from Urn 1", (**c**) "Win 100 if a black ball is drawn from Urn 2", (**d**) "Win 100 if a red ball is drawn from Urn 2". This experiment may be summarized as follows:

	Urn 1			Urn 2	
	50	50		100	
	B	**R**		**B**	**R**
a	100	0	**c**	100	0
b	0	100	**d**	0	100

Faced with the choices "Choose either bet **a** or bet **c**" (Choice 1) and "Choose either bet **b** or bet **d**" (Choice 2), most subjects stated that they preferred **a** to **c** and **b** to **d**.

As in Example 4.2.1, it is easy to check that there is no subjective probability which is capable of representing the stated choices as maximizing expected utility.

Example 4.2.2 also confirms the preference of decision-makers for known probabilities. The psychological literature (Tversky and Fox 1995) tends to interpret the observed behavior in the Ellsberg two-urn experiment as evidence "that people's preference depends not only on their degree of uncertainty but also on the *source of uncertainty*" (Tversky and Wakker 1995, p. 1270). In the Ellsberg two-urn experiment subjects preferred any bet on the urn with known proportions of black and red balls, the first source of uncertainty, to the equivalent bet on the urn where this information is not available, the second source of uncertainty. More generally, people prefer to bet on a better-known source.

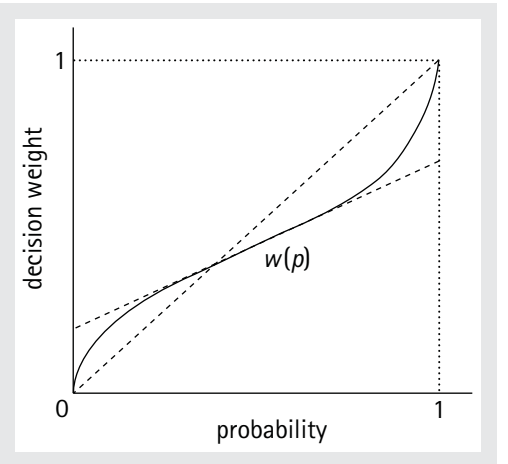

Fig. 4.1. Probability weighting function.

Sources of uncertainty are sets of events which belong to the same context. Tversky and Fox (1995), for example, compare bets on a random device with bets on the Dow Jones index, on football and basketball results, or temperatures in different cities. In contrast to the Ellsberg observations in Example 4.2.2, Heath and Tversky (1991) report a preference for betting on events with unknown probabilities compared to betting on the random devices for which the probabilities of events were known. Heath and Tversky (1991) and Tversky and Fox (1995) attribute this *ambiguity preference* to the *competence* which the subjects felt towards the source of the ambiguity. In the study by Tversky and Fox (1995) basketball fans were significantly more often willing to bet on basketball outcomes than on chance devices, and San Francisco residents preferred to bet on San Francisco temperatures rather than on a random device with known probabilities.

Whether subjects felt a preference for or an aversion against betting on the events with unknown probabilities, the experimental results indicate a systematic difference between the decision weights revealed in choice behavior and the assessed probabilities of events. There is a substantial body of experimental evidence that deviations are of the form illustrated in Figure 4.1. If the decision weights of an event would coincide with the assessed probability of this event as SEU suggests, then the function $w(p)$ depicted in Figure 4.1 should equal the identity. Tversky and Fox (1995) and others[4] observe that decision weights consistently exceed the probabilities of unlikely events and fall short of the probabilities near certainty. This S-shaped weighting function reflects the distinction between certainty and possibility which was noted by Kahneman and Tversky (1979). While the decision weights are almost linear for events which are possible but neither certain nor impossible, they deviate substantially for small-probability events.

[4] Gonzalez and Wu (1999) provide a survey of this psychological literature.

Decision weights can be observed in experiments. They reflect a decision-maker's ranking of events in terms of willingness to bet on the event. In general, they do not coincide, however, with the decision-maker's assessment of the probability of the event. Decision weights capture both a decision-maker's perceived ambiguity and the attitude towards it. Wakker (2001) interprets the fact that small probabilities are overweighted as *optimism* and the underweighting of almost certain probabilities as *pessimism*. The extent of these deviations reflects the *degree of ambiguity* held with respect to a subjectively assessed probability.

The experimental evidence collected on decision-making under ambiguity documents consistent differences between betting behavior and reported or elicited probabilities of events. While people seem to prefer risk over ambiguity if they feel unfamiliar with a source, this preference can be reversed if they feel competent about the source. Hence, we may expect to see more optimistic behavior in situations of ambiguity where the source is familiar, and more pessimistic behavior otherwise.

Actual economic behavior shows a similar pattern. Faced with Ellsberg-type decision problems, where an obvious lack of information cannot be overcome by personal confidence, most people seem to exhibit ambiguity aversion and choose among bets in a pessimistic way. In other situations, where the rewards are very uncertain, such as entering a career or setting up a small business, people may feel competent enough to make choices with an optimistic attitude. Depending on the source of ambiguity, the same person may be ambiguity-averse in one context and ambiguity-loving in an another.

4.3 Models of Ambiguity

The leading model of choice under uncertainty, subjective expected utility theory (SEU), is due to Savage (1954). In this theory, decision-makers know that the outcomes of their actions will depend on circumstances beyond their control, which are represented by a set of states of nature S. The states are mutually exclusive and provide complete descriptions of the circumstances determining the outcomes of the actions. Once a state becomes known, all uncertainty will be resolved, and the outcome of the action chosen will be realized. *Ex ante* it is not known, however, which will be the true state. *Ex post* precisely one state will be revealed to be true. An act a assigns an outcome $a(s) \in X$ to each state of nature $s \in S$. It is assumed that the decision-maker has preferences $\succsim$ over all possible acts. This provides a way of describing uncertainty without specifying probabilities.

If preferences over acts satisfy some axioms which attempt to capture reasonable behavior under uncertainty, then, as Savage (1954) shows, the decision-maker will

have a utility function over outcomes and a subjective probability distribution over the states of nature. Moreover, (s)he will choose so as to maximize the expected value of his or her utility with respect to his or her subjective probability. SEU implies that individuals have beliefs about the likelihood of states that can be represented by subjective probabilities. Savage (1954) can be, and has been, misunderstood as transforming decision-making under ambiguity into decision under risk. Note, however, that beliefs, though represented by a probability distribution, are purely subjective. Formally, people whose preference order $\succsim$ satisfies the axioms of SEU can be described by a probability distribution p over states in S and a utility function u over outcomes such that

$$a \succsim b \Leftrightarrow \int u(a(s))\, dp(s) \geqslant \int u(b(s))\, dp(s).$$

SEU describes a decision-maker who behaves like an expected utility maximizer whose uncertainty can be condensed into a subjective probability distribution, even if there is no known probability distribution over states. Taking up an example by Savage (1954), an individual satisfying the SEU axioms would be able to assign an exact number, such as 0.42 to the event described by the proposition "The next president of the United States will be a Democrat".

There are good reasons, however, for believing that SEU does not provide an adequate model of decision-making under ambiguity. It seems unreasonable to assume that the presence or absence of probability information will not affect behavior. In unfamiliar circumstances, when there is little evidence concerning the relevant variables, subjective certainty about the probabilities of states appears a questionable assumption. Moreover, as the Ellsberg paradox and the literature in Section 4.2 make abundantly clear, SEU is not supported by the experimental evidence.[5]

This section surveys some of the leading theories of ambiguity and discusses the relations between them. The two most prominent approaches are Choquet expected utility (CEU) and the multiple prior model (MP). CEU has the advantage of having a rigorous axiomatic foundation. MP does not have an overall axiomatic foundation, although some special cases of it have been axiomatized.

4.3.1 Multiple Priors

If decision-makers do not know the true probabilities of events, it seems plausible to assume that they might consider several probability distributions. The multiple prior approach suggests a model of ambiguity based on this intuition. Suppose an individual considers a set $\mathcal{P}$ of probability distributions as possible. If there is no information at all, the set $\mathcal{P}$ may comprise all probability distributions. More

[5] This does not preclude that SEU provides a good normative theory, as many researchers believe.

generally, the set $\mathcal{P}$ may reflect partial information. For example, in the Ellsberg three-urn example $\mathcal{P}$ may be the set of all probability distributions where the probability of a red ball being drawn equals $\frac{1}{3}$. For technical reasons $\mathcal{P}$ is assumed to be closed and convex.

An ambiguity-averse decision-maker may be modeled by preferences which evaluate an ambiguous act by the worst expected utility possible, given the set of probability distributions $\mathcal{P}$: i.e.

$$a \succsim b \Leftrightarrow \min_{p\in\mathcal{P}} \int u(a(s))\, dp(s) \geqslant \min_{p\in\mathcal{P}} \int u(b(s))\, dp(s).$$

These preferences provide an intuitive way to model a decision-maker with a pessimistic attitude towards ambiguity. They are axiomatized in Gilboa and Schmeidler (1989) and often referred to as *minimum expected utility* (MEU). Similarly, one can model an ambiguity-loving decision-maker by a preference order, which evaluates acts by the most optimistic expected utility possible with the given set of probability distributions $\mathcal{P}$,

$$a \succsim b \Leftrightarrow \max_{p\in\mathcal{P}} \int u(a(s))\, dp(s) \geqslant \max_{p\in\mathcal{P}} \int u(b(s))\, dp(s).$$

Preferences represented in this way are capable only of representing optimistic or pessimistic attitudes towards ambiguity (ambiguity aversion or ambiguity preference). Attitudes towards ambiguity which are optimistic for low-probability events and at the same time pessimistic for high-probability events are precluded in these cases. The following modified version, however, is capable of modeling ambiguity preference as well as ambiguity aversion.

A preference relation $\succsim$ on the set of acts is said to model *multiple priors* (α-MP) if there exists a closed and convex set of probability distributions $\mathcal{P}$ on S such that:

$$\begin{aligned} a \succsim b \Leftrightarrow\ & \alpha \min_{p\in\mathcal{P}} \int u(a(s))\, dp(s) + (1-\alpha) \max_{p\in\mathcal{P}} \int u(a(s))\, dp(s) \\ \geqslant\ & \alpha \min_{p\in\mathcal{P}} \int u(b(s))\, dp(s) + (1-\alpha) \max_{p\in\mathcal{P}} \int u(b(s))\, dp(s). \end{aligned}$$

These preferences provide an intuitive way to model a decision-maker whose reaction to ambiguity displays a mixture of optimism and pessimism. It is natural to associate the set of probability distributions $\mathcal{P}$ with the decision-maker's information about the probabilities of events, and the parameter α with the attitude towards ambiguity. For $\alpha = 1$, respectively $\alpha = 0$, the reaction is pessimistic (respectively optimistic), since the decision-maker evaluates any given act by the least (respectively, most) favorable probability distribution. Notice that the purely pessimistic case ($\alpha = 1$) coincides with MEU.

4.3.2 Choquet Integral and Capacities

A second related way of modeling ambiguity is to assume that individuals do have subjective beliefs, but that these beliefs do not satisfy all the mathematical properties of a probability distribution. In this case, decision weights may be defined by a *capacity*, a kind of non-additive subjective probability distribution. Choquet (1953) has proposed a definition of an expected value with respect to a capacity, which coincides with the usual definition of an expected value when the capacity is additive.[6]

For simplicity, assume that the set of states S is finite. A *capacity* on S is a real-valued function ν on the subsets of S such that $A \subseteq B$ implies $\nu(A) \leqslant \nu(B)$. Moreover, one normalizes $\nu(\varnothing) = 0$ and $\nu(S) = 1$. If, in addition, $\nu(A \cup B) = \nu(A) + \nu(B)$ for disjoint events A, B holds, then the capacity is a *probability distribution*. Probability distributions are, therefore, special cases of capacities. Another important example of a capacity is the *complete-uncertainty capacity* defined by $\nu(A) = 0$ for all $A \subsetneqq S$.

If S is finite, then one can order the outcomes of any act a from lowest to highest, $a_1 < a_2 < \ldots < a_{m-1} < a_m$. The Choquet expected utility (CEU) of an action a with respect to the capacity ν is given by the following formula,

$$\int u(a)\, d\nu = \sum_{r=1}^{m} u(a_r)\left[\nu(\{s \,|\, a(s) \geq a_r\}) - \nu(\{s \,|\, a(s) \geq a_{r+1}\})\right],$$

where we put $\{s \,|\, a(s) \geq a_{m+1}\} = \varnothing$ for notational convenience.

It is easy to check that for an additive capacity, i.e. a probability distribution, one has $\nu(\{s \,|\, a(s) \geq a_r\}) = \nu(\{s \,|\, a(s) = a_r\}) + \nu(\{s \,|\, a(s) \geq a_{r+1}\})$ for all r. Hence, CEU coincides with the expected utility of the act. For the *complete-uncertainty capacity*, the Choquet expected utility equals the utility of the worst outcome of this act, $\int u(a)\, d\nu = \min_{s \in S} u(a(s))$.

Preferences over acts for which there is a unique capacity ν and a utility function u such that

$$a \succsim b \Leftrightarrow \int u(a)\, d\nu \geq \int u(b)\, d\nu$$

will be referred to as *Choquet expected utility* (*CEU*) *preferences*. This representation has been derived axiomatically by Schmeidler (1989), Gilboa (1987) and Sarin and Wakker (1992). It is easy to see that the Ellsberg paradox can be explained by the CEU hypothesis.

[6] The theory and properties of capacities and the Choquet integral have been extensively studied. We will present here only a simple version of the general theory, suitable for our discussion of ambiguity and ambiguity attitude. For excellent surveys of the more formal theory, see Chateauneuf and Cohen (2000) and Denneberg (2000).

4.3.3 Choquet Expected Utility (CEU) and Multiple Priors (MP)

CEU preferences do not coincide with α-MP preferences. These preference systems have, however, an important intersection characterized by *convex capacities* and the *core of a capacity*. A capacity is said to be *convex* if $\nu(A \cup B) \geqslant \nu(A) + \nu(B) - \nu(A \cap B)$ holds for any events A, B in S. In particular, if two events are mutually exclusive, i.e. $A \cap B = \varnothing$, then the sum of the decision weights attached to the events A and B does not exceed the decision weight associated with their union $A \cup B$.

For any capacity ν on S, one can define a set of probability distributions called the *core* of the capacity ν, $\text{core}(\nu)$. The core of a capacity ν is the set of probability distributions which yield a higher probability for each event than the capacity ν,

$$\text{core}(\nu) = \left\{ p \in \Delta(S) | \; p(A) \geqslant \nu(A) \text{ for all } A \subseteq S \right\},$$

where we write $\Delta(S)$ for the set of all probability distributions on S and $p(A)$ for $\sum_{s \in A} p(s)$. If the capacity satisfies $\nu(A \cup B) = \nu(A) + \nu(B) - \nu(A \cap B)$ for all events A and B, then it is a probability distribution, and the core consists of only this probability distribution.

If the core of a capacity is nonempty, then it defines a set of probability distributions associated with the capacity. The capacity may be viewed as a set of constraints on the set of probability distributions which a decision-maker considers possible. These constraints may arise from the decision-maker's information about the probability of events. If a decision-maker faces no ambiguity, the capacity will be additive, i.e. a probability distribution, and the core will consist of this single probability distribution.

Example 4.3.1. In Example 4.2.1, for example, one could consider the state space $S = \{R, B, Y\}$ and the capacity ν defined by

$$\nu(E) = \begin{cases} \frac{1}{3} & \text{if } \{R\} \subseteq E \\ \frac{2}{3} & \text{if } \{B, Y\} \subseteq E \\ 0 & \text{otherwise} \end{cases}$$

for any event $E \neq S$. This capacity ν is convex and its core is the set of probability distributions p with $p(R) = \frac{1}{3}$, $\text{core}(\nu) = \{p \in \Delta(S) | \; p(R) = \frac{1}{3}\}$.

It is natural to ask when a capacity will define a set of priors such that the representations of CEU and α-MP coincide. Schmeidler (1989) proved that for a convex capacity, the Choquet integral for any act a is equal to the minimum of the expected utility of a, where the minimum is taken over the probabilities in the core. If ν is a

convex capacity on S, then

$$\int u(a)dv = \min_{p\in \text{core}(v)} \int u(a(s))dp(s).$$

Since the core of a convex capacity is never empty, this result provides a partial answer to our question. It shows that the α-MP preference representation equals the CEU preference representation if $\alpha = 1$ holds and if the capacity v is convex.

Jaffray and Philippe (1997) show a more general relationship between α-MP preferences and CEU preferences.[7] Let μ be a convex capacity on S, and for any $\alpha \in [0, 1]$ define the capacity

$$v(A) := \alpha\mu(A) + (1 - \alpha)\,[1 - \mu(S\backslash A)]\,,$$

which we will call *JP capacity*. JP capacities allow preferences to be represented in both the α-MP and CEU forms. For $\alpha \in [0, 1]$ and a convex capacity μ, let v be the associated JP capacity, then one obtains

$$\int u(a)dv = \alpha \min_{p\in \text{core}(\mu)} \int u(a(s))dp(s) + (1 - \alpha) \max_{p\in \text{core}(\mu)} \int u(a(s))dp(s).$$

The CEU preferences with respect to the JP capacity, v, coincide with the α-MP preferences, where the set of priors is the core of the convex capacity μ on which the JP capacity depends, $\mathcal{P} = \text{core}\,(\mu)$. As in the case of α-MP preferences, it is natural to interpret α as a parameter related to the ambiguity attitude and the core of μ, the set of priors, as describing the ambiguity of the decision-maker.

A special case of a JP capacity, which illustrates how a capacity constrains the set of probability distributions in the core is the *neo-additive capacity*.[8] A neo-additive capacity is a JP capacity with a convex capacity μ defined by $\mu(E) = (1 - \delta)\pi(E)$ for all events $E \neq S$, where π is a probability distribution on S. In this case,

$$\mathcal{P} = \text{core}\,(\mu) = \{p \in \Delta\,(S) \mid p(E) \geq (1 - \delta)\,\pi(E)\}$$

is the set of priors. A decision-maker with beliefs represented by a neo-additive capacity may be viewed as holding ambiguous beliefs about an additive probability distribution π. The parameter δ determines the size of the set of probabilities

[7] Recently, Ghirardato, Maccheroni, and Marinacci (2004) have axiomatized a representation

$$V(f) = \alpha(f) \min_{p\in\mathcal{P}} \int u(f(s))dp(s) + (1 - \alpha(f)) \max_{p\in\mathcal{P}} \int u(f(s))dp(s),$$

where the set of probability distributions $\mathcal{P}$ is determined endogenously, and where the weights $\alpha(f)$ depend on the act f. Nehring (2007) axiomatizes a representation where the set of priors can be determined partially exogenously and partially endogenously.

[8] Neo-additive capacities are axiomatized and carefully discussed in Chateauneuf, Eichberger, and Grant (2007).

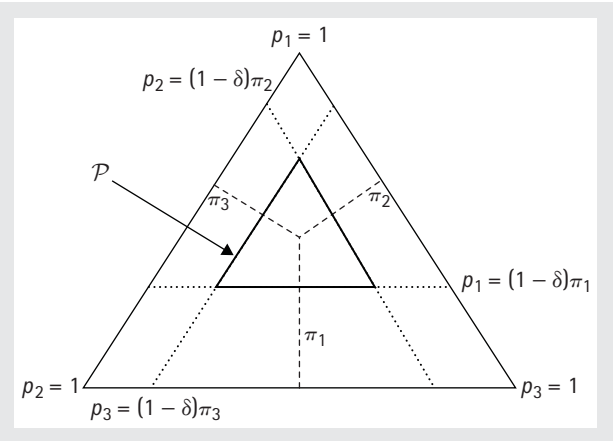

Fig. 4.2. Core of a neo-additive capacity.

around π which the decision-maker considers possible. It can be interpreted as a measure of the decision-maker's ambiguity.

Figure 4.2 illustrates the core of a neo-additive capacity for the case of three states. The outer triangle represents the set of all probability distributions[9] over the three states $S = \{s_1, s_2, s_3\}$. Each point in this triangle represents a probability distribution $p = (p_1, p_2, p_3)$. The set $\mathcal{P}$ of probability distributions in the core of the neo-additive capacity μ is represented by the inner triangle with the probability distribution $\pi = (\pi_1, \pi_2, \pi_3)$ as its center.

4.4 Ambiguity and Ambiguity Attitude

A central, yet so far still not completely resolved, problem in modeling ambiguity concerns the separation of ambiguity and ambiguity attitude. As discussed in Section 4.2, early experiments, e.g. Ellsberg (1961), suggested an aversion of decision-makers to ambiguity arising from the lack of information about the probability of events. The negative attitude towards ambiguity seems not to hold in situations where the decision-maker has no information about the probabilities of

[9] Figure 4.2 is the projection of the three-dimensional simplex onto the plane. Corner points correspond to the degenerate probability distributions assigning probability $p_i = 1$ to state s_i and $p_j = 0$ to all other states. Points on the edge of the triangle opposite the corner $p_i = 1$ assign probability of zero to the state s_i. Points on a line parallel to an edge of the triangle, e.g. the ones marked $p_i = (1 - \delta)\pi_i$, are probability distributions for which $p_i = (1 - \delta)\pi_i$ holds. Moreover, if one draws a line from a corner point, say $p_1 = 1$, through the point $\pi = (\pi_1, \pi_2, \pi_3)$ in the triangle to the opposing edge, then the distance from the opposing edge to the point π represents the probability π_1.

events, but feels competent about the situation. Experimental evidence suggests that a decision-maker who feels expert in an ambiguous situation is likely to prefer an ambiguous act to an unambiguous one, e.g. Tversky and Fox (1995).

Separating ambiguity and ambiguity attitude is important for economic models, because attitudes towards ambiguity of a decision-maker may be seen as stable personal characteristics, whereas the experienced ambiguity varies with the information about the environment. Here, information should not be understood in the Bayesian sense of evidence which allows one to condition a given probability distribution. Information refers to evidence which in the decision-maker's opinion may have some impact on the likelihood of decision-relevant events. For example, one may reasonably assume that an entrepreneur who undertakes a new investment project feels ambiguity about the chances of success. Observing success and failure of other entrepreneurs with similar, but different, projects is likely to affect ambiguity. Information about the success of a competitor's investment may reduce ambiguity, while failure of it may have the opposite effect. Hence, the entrepreneur's degree of ambiguity may change with such information. In contrast, it seems reasonable to assume that optimism or pessimism, understood as the underlying propensity to take on uncertain risks, is a more permanent feature of the decision-maker's personality.

Achieving such a separation is complicated by two additional desiderata.

(i) In the spirit of Savage (1954), one would like to derive all decision-relevant concepts purely from assumptions about the preferences over acts.
(ii) The distinction of ambiguity and ambiguity attitude should be compatible with the notion of risk attitudes in cases of decision-making under risk.

The second desideratum is further complicated since there are differing notions of *risk attitudes* in SEU and *rank-dependent expected utility* (RDEU),[10] as introduced by Quiggin (1982).

The three approaches outlined here differ in these respects. Ultimately, the answer to the question as to how to separate ambiguity from ambiguity attitude may determine the choice among the different models of decision-making under ambiguity discussed in Section 4.3.

4.4.1 Ambiguity Aversion and Convexity

The Ellsberg paradox suggests that people dislike the ambiguity of not knowing the probability distribution over states, e.g. the proportions of balls in the urn. In an effort to find preference representations which are compatible with the behavior observed in this paradox, most of the early research assumed *ambiguity aversion*

[10] Chapters 2 and 3 of this Handbook deal with rank-dependent expected utility and other non-expected utility theories.

and attributed all deviations between decision weights and probabilities to the ambiguity experienced by the decision-maker.

Denote by $\mathcal{A}$ the set of acts. Schmeidler (1989) and Gilboa and Schmeidler (1989) assume that acts yield lotteries as outcomes.[11] Hence, for constant acts, decision-makers choose among lotteries. In this framework, one can define (pointwise) convex combinations of acts. An act with such a convex combination of lotteries as outcomes can be interpreted as a reduction in ambiguity, because there is a state-wise diversification of lottery risks. A decision-maker is called *ambiguity-averse* if any $\frac{1}{2}$-convex combination of two indifferent acts is considered at least as good as these acts, formally,

$$\text{for all acts } a, b \in \mathcal{A} \quad \text{with} \quad a \sim b \quad \text{holds} \quad \left(\tfrac{1}{2}\right) a + \left(\tfrac{1}{2}\right) b \succsim b. \qquad \text{(Ambiguity aversion)}$$

For preferences satisfying ambiguity aversion, Schmeidler (1989) shows that the capacity of the CEU representation must be *convex*. Moreover, for the derivation of the MEU representation, Gilboa and Schmeidler (1989) include ambiguity aversion as an axiom.

In a recent article, Ghirardato, Maccheroni, and Marinacci (2004) provide a useful exposition of the axiomatic relationship among representations. For a given utility function u over lotteries, one can treat the act a as a parameter and denote by $u_a : S \to \mathbb{R}$ the function $u_a(s) := u(a(s))$ which associates with each state s the utility of the lottery assigned to this state by the act $a \in \mathcal{A}$. Five standard assumptions[12] on the preference order $\succsim$ on $\mathcal{A}$ characterize a representation by a positively homogeneous and constant additive[13] functional $I(f)$ on the set of real-valued functions f and a non-constant affine function $u : X \to \mathbb{R}$ such that, for any acts $a, b \in \mathcal{A}$,

$$a \succsim b \Leftrightarrow I(u_a) \geq I(u_b).$$

If the preference order satisfies in addition *ambiguity aversion*, then there is a *unique* nonempty, compact, convex set of probabilities $\mathcal{P}$ such that

$$I(u_a) = \min_{p \in \mathcal{P}} \int u(a(s))\, dp(s). \qquad \text{(MEU)}$$

The CEU and SEU representations can now be obtained by extending the independence axiom to larger classes of acts.

[11] Anscombe and Aumann (1963) introduced this notion of an act in order to simplify the derivation of SEU.

[12] The five axioms are *weak order*, *certainty independence*, *Archimedean axiom*, *monotonicity*, and *nondegeneracy*. For more details, compare Ghirardato, Maccheroni, and Marinacci (2004, p. 141).

[13] The functional I is *constant-additive* if $I(f + c) = I(f) + c$ holds for any function $f : S \to \mathbb{R}$ and any constant $c \in \mathbb{R}$. The functional I is *positively homogeneous* if $I(\lambda f) = \lambda I(f)$ for any function $f : S \to \mathbb{R}$ and all $\lambda \geq 0$.

(i) *CEU*: If the preference order satisfies in addition *comonotonic independence,* i.e.

for all *comonotonic*[14] acts $a, b \in \mathcal{A}$ with $a \sim b$ holds $\left(\frac{1}{2}\right) a + \left(\frac{1}{2}\right) b \sim b$, (*Comonotonic independence*)

then there is a *convex capacity* ν on S such that

$$I(u_a)) = \int u(a)\, d\nu. \tag{CEU}$$

(ii) *SEU*: If the preference order satisfies *independence*, i.e.

for all acts $a, b \in \mathcal{A}$ with $a \sim b$ holds $\left(\frac{1}{2}\right) a + \left(\frac{1}{2}\right) b \sim b$, (*Independence*)

then there is a probability distribution π on S such that

$$I(u_a) = \int u(a(s))\, d\pi(s). \tag{SEU}$$

Given ambiguity aversion, the CEU model is more restrictive than the MEU model, since it requires also comonotonic independence. As explained in Section 4.3.3, for convex capacities, the core is nonempty and represents the set of priors. Imposing independence for all acts makes SEU the most restrictive model. In this case, strict ambiguity aversion is ruled out. Only the limiting case of a unique additive probability distribution π remains, which coincides with the capacity in CEU and forms the trivial set of priors $\mathcal{P} = \{\pi\}$ for MEU.

A priori, this approach allows only for a negative attitude towards ambiguity. Any deviation from expected utility can, therefore, be interpreted as ambiguity. Hence, absence of ambiguity coincides with SEU preferences.

4.4.2 Comparative Ambiguity Aversion

In the context of decision-making under risk, Yaari (1969) defines a decision-maker A as *more risk-averse* than decision-maker B if A ranks a certain outcome higher than a lottery whenever B prefers the certain outcome over this lottery. If one defines as *risk-neutral* a decision-maker who ranks lotteries according to their expected value, then one can classify decision-makers as risk-averse and risk-loving according to whether they are more, respectively less, risk-averse than a risk-neutral decision-maker. Note that the reference case of risk neutrality is arbitrarily chosen.

In the spirit of Yaari (1969), a group of articles[15] propose comparative notions of "more ambiguity-averse". Epstein (1999) defines a decision-maker A as *more*

[14] Two acts $a, b \in \mathcal{F}$ are *comonotonic* if there exists no $s, s' \in S$ such that $a(s) \succ a(s')$ and $b(s') \succ b(s)$. This implies that comonotonic acts rank the states in the same way.

[15] Kelsey and Nandeibam (1996), Epstein (1999), Ghirardato and Marinacci (2002), and Grant and Quiggin (2005) use the comparative approach for separating ambiguity and ambiguity attitude.

ambiguity-averse[16] than decision-maker B if A prefers an *unambiguous act* over another arbitrary act whenever B ranks these acts in this way. For this definition the notion of an "unambiguous act" has to be introduced. Epstein (1999) assumes that there is a set of *unambiguous events* for which decision-makers can assign probabilities. Acts which are measurable with regard to these unambiguous events are called *unambiguous acts*.

Epstein uses probabilistically sophisticated preferences as the benchmark to define ambiguity neutrality. Probabilistically sophisticated decision-makers assign a unique probability distribution to all events such that they can rank all acts by ranking the induced lotteries over outcomes, (see Machina and Schmeidler 1992). SEU decision-makers are probabilistically sophisticated, but there are other non-SEU preferences which are also probabilistically sophisticated.[17]

Decision-makers are *ambiguity-averse,* respectively *ambiguity-loving*, if they are more, respectively less, ambiguity-averse than a probabilistically sophisticated decision-maker. Hence, *ambiguity-neutral* decision-makers are probabilistically sophisticated. Ambiguity-neutral decision-makers do not experience ambiguity. Though they may not know the probability of events, their beliefs can be represented by a subjective probability distribution.

If a decision-maker has pessimistic MEU preferences and if all prior probability distributions coincide on the unambiguous events, then the decision-maker is ambiguity-averse in the sense of Epstein (1999). A CEU preference order is ambiguity-averse if there is an additive probability distribution in the core of the capacity with respect to which the decision-maker is probabilistically sophisticated for unambiguous acts. Hence, convexity of the capacity is neither a necessary nor a sufficient condition for ambiguity aversion in the sense of Epstein (1999). Ambiguity neutrality coincides with the absence of perceived ambiguity, since an ambiguity-neutral decision-maker has a subjective probability distribution over all events. Hence, risk preferences reflected by the von Neumann–Morgenstern utility in the case of SEU are independent of the ambiguity attitude. A disadvantage of Epstein's (1999) approach is, however, the assumption that there is an exogenously given set of unambiguous events.[18]

Ghirardato and Marinacci (2002) also suggest a comparative notion of ambiguity aversion. They call a decision-maker A *more ambiguity-averse* than decision-maker

[16] Epstein (1999) calls such a relation "more uncertainty-averse". Since we use uncertainty as a generic term, which covers also the case where a decision-maker is probabilistically sophisticated, we prefer the dubbing of Ghirardato and Marinacci (2002).

[17] *Probabilistical sophistication* is a concept introduced by Machina and Schmeidler (1992) in order to accommodate experimentally observed deviations from expected utility in the context of choice over lotteries. A typical case of probabilistically sophisticated preferences is *rank-dependent expected utility (RDEU)* proposed by Quiggin (1982) for choice when the probabilities are known.

[18] In Epstein and Zhang (2001), unambiguous events are defined based purely on behavioral assumptions. See, however, Nehring (2006*b*), who raises some questions about the purely behavioral approach.

B if A prefers a constant act over another act whenever B ranks these acts in this way. In contrast to Epstein (1999), Ghirardato and Marinacci (2002) use constant acts, rather than unambiguous acts, in order to define the relation "more ambiguity-averse". The obvious advantage is that they do not need to assume the existence of unambiguous acts. The disadvantage lies in the fact that this comparison does not distinguish between attitudes towards risk and attitudes towards ambiguity. Hence, for two decision-makers with SEU preferences holding the same beliefs, i.e. the Yaari case, A will be considered more ambiguity-averse than B simply because A has a more concave von Neumann–Morgenstern utility function than B. A disadvantage of this theory is that it implies that the usual preferences in the Allais paradox exhibit ambiguity aversion. However, most researchers do not consider ambiguity to be a significant factor in the Allais paradox.

Ghirardato and Marinacci (2002), therefore, restrict attention to preference orders which allow for a CEU representation over binary acts. They dub such preferences "biseparable". The class of biseparable preferences comprises SEU, CEU, and MEU and is characterized by a well-defined von Neumann–Morgenstern utility function. In this context it is possible to control for risk preferences as reflected in the von Neumann–Morgenstern utility functions. Biseparable preferences which have (up to an affine transformation) the same von Neumann–Morgenstern utility function are called *cardinally symmetric.*

As the reference case of *ambiguity neutrality,* Ghirardato and Marinacci (2002) take cardinally symmetric SEU decision-makers. Hence, decision-makers are *ambiguity-averse* (respectively, *ambiguity-loving*) if they have cardinally symmetric biseparable preferences and if they are more (respectively, less) ambiguity-averse than a SEU decision-maker.

Ghirardato and Marinacci (2002) show that CEU decision-makers are ambiguity-averse if and only if the core of the capacity characterizing them is non-empty. In contrast to Epstein (1999), convexity of the preference order is sufficient for ambiguity aversion but not necessary. MEU individuals are ambiguity-averse in the sense of Ghirardato and Marinacci (2002).

Characterizing ambiguity attitude by a comparative notion, as in Epstein (1999) and Ghirardato and Marinacci (2002), it is necessary to identify (i) acts as more or less ambiguous and (ii) a preference order as ambiguity-neutral. In the case of Epstein (1999), unambiguous acts, i.e. acts measurable with respect to unambiguous events, are considered less ambiguous than other acts, and probabilistically sophisticated preferences were suggested as ambiguity-neutral. For Ghirardato and Marinacci (2002), constant acts are less ambiguous than other acts, and SEU preferences are ambiguity-neutral.

It is possible to provide other comparative notions of ambiguity by varying either the notion of the less ambiguous acts or the type of reference preferences which

are considered ambiguity-neutral. Grant and Quiggin (2005) suggest a concept of "more uncertain" acts. For ease of exposition, assume that acts map states into utilities. Comparing two acts a and b, consider a partition of the state space in two events, B_a and W_a such that $a(s) \gtrsim a(t)$ for all $s \in B_a$ and all $t \in W_a$. Then Grant and Quiggin (2005) call act b an *elementary increase in uncertainty* of act a if there are positive numbers α and β such that $b(s) = a(s) + \alpha$ for all $s \in B_a$ and $b(s) = a(s) - \beta$ for all $s \in W_a$. Act b has outcomes which are higher by a constant α than those of act a for states yielding high outcomes, and outcomes which are lower by β than those of act a for states with low outcomes. In this sense, exposure to ambiguity is higher for act b than for act a. A decision-maker A is *at least as uncertainty-averse* as decision-maker B if A prefers an act a over act b whenever b is an elementary increase in uncertainty of a and B prefers a over b. For the reference case of uncertainty neutrality they use SEU preferences.

In contrast to Ghirardato and Marinacci (2002), Grant and Quiggin (2005) do not control for risk preferences reflected by the von Neumann–Morgenstern utility function. Hence, an SEU decision-maker A is more uncertainty-averse than another SEU decision-maker B if both have the same beliefs, represented by an additive probability distribution over states and if A's von Neumann–Morgenstern utility function is a concave transformation of B's von Neumann–Morgenstern utility function. Using concepts introduced by Chateauneuf, Cohen, and Meilijson (2005), Grant and Quiggin (2005) characterize more uncertainty-averse CEU decision-makers by a pessimism index exceeding an index of relative concavity of the von Neumann–Morgenstern utility functions.

4.4.3 Optimism and Pessimism

Inspired by the Allais paradox, Wakker (2001) suggests a notion of *optimism* and *pessimism* based on choice behavior over acts. These notions do not depend on a specific form of representation. The appeal of this approach lies in its immediate testability in experiments and its link to properties of capacities in the CEU model. For the CEU representation, Wakker (2001) shows that optimism corresponds to concavity and pessimism to convexity of a capacity. Moreover, Wakker (2001) provides a method behaviorally to characterize decision-makers who overweight events with extreme outcomes, a fact which is often observed in experiments.[19] For ease of exposition, assume again that acts associate real numbers with states (see matrix below). The matrix shows four acts a_1, a_2, a_3, a_4 defined on a partition of the state space $\{H, A, I, L\}$ with outcomes $M > m > 0$.

[19] Compare Fig. 4.1.

	H	A	I	L
a_1	m	m	0	0
a_2	M	0	0	0
a_3	m	m	m	0
a_4	M	0	m	0

For given M, assume that m is chosen such that the decision-maker is indifferent between acts a_1 and a_2, i.e. $a_1 \sim a_2$. Wakker (2001) calls a decision-maker *pessimistic* if a_3 is preferred to a_4, i.e. $a_3 \gtrsim a_4$, and *optimistic* if the opposite preference is revealed, i.e. $a_3 \precsim a_4$.

The intuition is as follows. Conditional on the events H or A occurring, m is the certainty equivalent to the partial act yielding M on H and 0 on A. In acts a_3 and a_4 the outcome on the "irrelevant" event I has been increased from 0 to m. Of course, a SEU decision-maker will be indifferent also between acts a_3 and a_4. For a pessimistic decision-maker, the increase in the outcome on the event I makes the partial certainty equivalent more attractive. In contrast, an optimistic decision-maker will now prefer the act a_4, because the increase in the outcome on the event I makes the partial act M on H and 0 on A more attractive.

A key result of Wakker (2001) shows that for CEU preferences, pessimism implies a convex capacity, and optimism a concave capacity. Moreover, for CEU preferences, one can define a weak order on events, which orders any two events as one being *revealed more likely* than the other. This order allows one to define intervals of events. It is possible to restrict optimism or pessimism to nondegenerate intervals of events. Hence, if there is an event E such that the decision-maker is optimistic for all events which are revealed less likely than E, and pessimistic for all events which are revealed more likely than E, then this decision-maker will overweight events with extreme outcomes. For a CEU decision-maker, in this case, the capacity will be partially concave and partially convex.

One may be inclined to think that a decision-maker who is both pessimistic and optimistic, i.e. with $a_3 \sim a_4$, will have SEU preferences. This is, however, not true. For example, a CEU decision-maker with preferences represented by the capacity $\nu(E) = (1-\delta)\pi(E)$ for all $E \neq S$, where π is an additive probability distribution on S and $\delta \in (0,1)$, will rank acts according to $\int u(a)d\nu = \delta \cdot \min_{s \in S} u(a(s)) + (1-\delta) \cdot \int u(a)d\pi$. Straightforward computations show that $\int u(a_1)d\nu = \int u(a_2)d\nu$ holds if and only if $\pi(H \cup A) \cdot u(m) = \pi(H) \cdot u(M) + \pi(A) \cdot u(0)$. Hence,

$$\int u(a_3)d\nu - \int u(a_4)d\nu$$
$$= (1-\delta)\left[\pi(H \cup A) \cdot u(m) - \pi(H) \cdot u(M) - \pi(A) \cdot u(0)\right] = 0.$$

This CEU decision-maker behaves like an SEU decision-maker *as long as the minimum of acts remains unchanged*. For acts with varying worst outcome, however, the behavior would be quite distinct. It is easy to check that the capacity ν is convex.[20] Hence, a decision-maker who evaluates acts a_3 and a_4 as indifferent need not have SEU preferences.

4.5 Economic Applications

Important economic insights depend on the way in which decision-making under uncertainty is modeled. Despite the obvious discrepancies between choice behavior predicted based on SEU preferences and actual behavior in controlled laboratory experiments, SEU has become the most commonly applied model in economics. SEU decision-makers behave like Bayesian statisticians. They update beliefs according to Bayes's rule and behave consistently with underlying probability distributions. In particular, in financial economics, where investors are modeled who choose portfolios, and in contract theory, where agents design contracts suitable to share risks and to deal with information problems, important results depend on this assumption. Nevertheless, in both financial economics and contract theory, there are phenomena which are hard or impossible to reconcile with SEU preferences. Therefore, there is growing research into the implications of alternative models of decision-making under uncertainty. Applications range from auctions, bargaining, and contract theory to liability rules. There are several surveys of economic applications, e.g. Chateauneuf and Cohen (2000), Mukerji (2000), and Mukerji and Tallon (2004). We will describe here only two results of general economic importance relating to financial economics and risk sharing.

4.5.1 Financial Economics

If ambiguity aversion is assumed, then CEU and α-MP preferences have kinks at points of certain consumption. Thus they are not even locally risk-neutral. The model of financial markets of Dow and Werlang (1992) shows that SEU yields the paradoxical result that an individual should either buy or short-sell every asset. This follows from local risk neutrality. Apart from the knife-edge case where all assets have the same expected return, every asset either offers positive expected returns,

[20] Note, however, that the capacity does not satisfy the solvability condition imposed by Wakker (2001, assumption 5.1, p. 1047), which is required for the full characterizations in thms. 5.2 and 5.4.

in which cases it should be purchased, or negative expected returns, in which case it should be short sold. Assuming CEU preferences and ambiguity aversion, Dow and Werlang (1992) show that there is a range of asset prices for which an investor may not be induced to trade. In particular, ambiguity-averse investors will not turn from investing into assets to short-sales by a marginal change of asset prices as SEU models predict. Kelsey and Milne (1995) study asset pricing with CEU preferences and show that many conventional asset pricing results may be generalized to this context.

Epstein and Wang (1994) extend the Dow–Werlang result to multiple time periods. They show that there is a continuum of possible values of asset prices in a financial market equilibrium. Thus, ambiguity causes prices to be no longer determinate. They argue that this is a formal model of Keynes's intuition that ambiguity would cause asset prices to depend on a conventional valuation rather than on market fundamentals.

In a related paper Epstein (2001) shows that differences in the perception of ambiguity can explain the consumption home bias paradox. This paradox refers to the fact that domestic consumption is more correlated with domestic income than theory would predict. Epstein (2001) explains this by arguing that the individual perceives foreign income to be more ambiguous.

Mukerji and Tallon (2001) use the CEU to show that ambiguity can be a barrier to risk sharing through diversified portfolios. There are securities which could, in principle, allow risk to be shared. However, markets are incomplete, and each security carries some idiosyncratic risk. If this idiosyncratic risk is perceived as sufficiently ambiguous, it is possible that ambiguity aversion may deter people from trading it. The authors show that ambiguous risks cannot be diversified in the same way as standard risks. This has the implication that firms as well as individuals may be ambiguity-averse.

4.5.2 Sharing Ambiguous Risks

Consider an economy with one physical commodity and multiple states of nature. If all individuals have SEU preferences, and if there is no aggregate uncertainty, then in a market equilibrium each individual has certain consumption. An individual's consumption is proportional to the expected value of his or her endowment. If there is aggregate uncertainty, then risk is shared between all individuals as an increasing function of their risk tolerance. Individuals' consumptions are comonotonic with one another and with the aggregate endowment.

Chateauneuf, Dana, and Tallon (2000) consider risk sharing when individuals have CEU preferences. In the case where all individuals have beliefs represented by *the same* convex capacity, they show that the equilibrium is the same as would

be obtained if all individuals had SEU preferences and beliefs represented by a particular additive probability distribution. The reason for this is that in an economy with one good, no production, and aggregate uncertainty, all Pareto optimal allocations are comonotonic. CEU preferences evaluate comonotonic acts with the same set of decision weights. These decision weights can be treated as if they are a probability distribution. Hence, any competitive equilibrium coincides with an equilibrium of the economy where SEU decision-makers have a probability distribution equal to these decision weights. In such an equilibrium the optimal degree of risk sharing obtains.

Dana (2004) extends this result by investigating the comparative statics of changes in the endowment. She shows that while any given equilibrium is similar to an equilibrium without ambiguity, the comparative statics of changes in the endowment is different in an economy with ambiguity. In the presence of ambiguity small changes in the endowment can cause large changes in equilibrium prices. The price ratio is always significantly higher in states in which the endowment is relatively scarce. As a consequence, individuals who have larger endowments in such states get higher utility.

4.6 Concluding Remarks

In this chapter, our focus has been on purely behavioral approaches to decision-making under ambiguity. In particular, we have reviewed the literature which takes the Savage, or Anscombe–Aumann, framework as the basis of the analysis.[21] Hence, ambiguity and ambiguity attitudes of a decision-maker have to be inferred from choices based on preferences over acts alone. We have seen that such a separation has not been achieved so far. The difficulty derives from the fact that choice behavior over acts reveals the decision weights of a decision-maker. It does not reveal, however, how much of the decision weight has to be attributed to ambiguity and how much to ambiguity attitude. In these concluding remarks, we would like to mention two other approaches, which start from different premises, in order to obtain a separation of ambiguity and ambiguity attitude. We will also point out another unresolved issue, which is related to the distinction of ambiguity and ambiguity attitude.

[21] There are also other approaches to model ambiguity and to explain the Ellsberg paradox. For example, Segal (1987) shows that one can explain the Ellsberg paradox as choice over compound lotteries without the reduction axiom. Halevy (2007) studies experimentally to what extent one can distinguish these approaches.

A separation of ambiguity and ambiguity attitude can be achieved if one allows for additional a priori information. Klibanoff, Marinacci, and Mukerji (2005) take *two types of acts* and *two preference orders* as primitives. The representation over second-order acts is assumed to model ambiguity and ambiguity attitude. Here, exogenously specified preferences achieve the separation of ambiguity and ambiguity attitude. It is not clear, however, whether one can identify these two types of preference orders from the observed choices over acts.

Nehring (2006*a*) considers partial information about probabilities which characterize a set of probability distributions consistent with this information. If a decision-maker's preferences over acts are compatible with this information, then one can obtain a multiple prior representation with this set of probability distributions. If one takes this set of priors as representing the ambiguity, a decision-maker's ambiguity attitude may be derived from the decision weights. Finally, we would like to point out a problem which is related to the issue of separating ambiguity from ambiguity attitude. If beliefs of a decision-maker are modeled by capacities or sets of probability distributions, it is no longer clear what is an appropriate support notion. This problem becomes important if one considers games where players experience ambiguity about the strategy choice of their opponent. Dow and Werlang (1994), Lo (1996), Eichberger and Kelsey (2000), and Marinacci (2000) study games with players who hold ambiguous beliefs about their opponent's behavior. Eichberger, Kelsey, and Schipper (2007) provide experimental evidence for ambiguity aversion of players in a game. This extends previous research by showing that ambiguity aversion could also be present in games.

In an equilibrium of a game, understood as a situation in which players have no incentives to deviate unilaterally from their strategy choices, the information generated by the equilibrium behavior of the players must be consistent with their beliefs. In traditional game-theoretic analysis, where players' beliefs about their opponents' behavior was modeled by probability distributions, such consistency was guaranteed by a Nash equilibrium in mixed strategies. In a mixed-strategy Nash equilibrium, the support of the equilibrium mixed strategies contains only best-reply strategies.

With ambiguity, there is no obvious support notion. For capacities or sets of probability distributions, there are many support concepts.[22] If one assumes that players play best-reply strategies given some ambiguity about the opponents' strategy choice, then the support notion should reflect a player's perceived ambiguity. In contrast, a player's attitude towards ambiguity appears more as a personal characteristic.

[22] Ryan (2002) provides epistemic conditions for support notions if decision-makers are uncertainty-averse. Haller (2000) studies implications of different support concepts for equilibria in games.

References

Allais, M. (1953). The So-Called Allais Paradox and Rational Decision under Uncertainty. *Econometrica*, 21, 503–46.

Anscombe, F. J., and Aumann, R. J. (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics*, 34, 199–205.

Camerer, C. (1995). Individual Decision Making. In J. H. Kagel and A. E. Roth (eds.), *The Handbook of Experimental Economics*, 587–703. Princeton: Princeton University Press.

—— and Weber, M. (1992). Recent Developments in Modelling Preferences: Uncertainty and Ambiguity. *Journal of Risk and Uncertainty*, 5, 325–70.

Chateauneuf, A., and Cohen, M. (2000). Choquet Expected Utility Model: A New Approach to Individual Behavior under Uncertainty and Social Welfare. In M. Grabisch, T. Murofushi, and M. Sugeno (eds.), *Fuzzy Measures and Integrals*, 289–313. Berlin: Physica-Verlag.

—— —— and Meilijson, I. (2005). More Pessimism than Greediness: A Characterization of Monotone Risk Aversion in the Rank-Dependent Expected Utility Model. *Economic Theory*, 25/3, 649–67.

—— Dana, R.-A., and Tallon, J.-M. (2000). Optimal Risk-Sharing Rules and Equilibria with Choquet Expected Utility. *Journal of Mathematical Economics*, 34, 191–214.

—— Eichberger, J., and Grant, S. (2007). Choice under Uncertainty with the Best and Worst in Mind: Neo-Additive Capacities. *Journal of Economic Theory*, 137/1, 538–67.

Choquet, G. (1953). Theory of Capacities. *Annales Institut Fourier*, 5, 131–295.

Dana, R.-A. (2004). Ambiguity, Uncertainty-Aversion and Equilibrium Welfare. *Economic Theory*, 23, 569–88.

Denneberg, D. (2000). Non-Additive Measure and Integral, Basic Concepts and their Role for Applications. In M. Grabisch, T. Murofushi, and M. Sugeno (eds.), *Fuzzy Measures and Integrals*, 289–313. Berlin: Physica-Verlag.

Dow, J., and Werlang, S. R. C. (1992). Uncertainty Aversion, Risk Aversion, and the Optimal Choice of Portfolio. *Econometrica*, 60, 197–204.

—— —— (1994). Nash Equilibrium under Uncertainty: Breaking Down Backward Induction. *Journal of Economic Theory*, 64, 305–24.

Eichberger, J., and Kelsey, D. (2000). Non-Additive Beliefs and Strategic Equilibria. *Games and Economic Behavior*, 30, 183–215.

—— —— and Schipper, B. (2007). Granny versus Game Theorist: Ambiguity in Experimental Games. *Theory and Decision*, 64, 333–62.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643–69.

Epstein, L. G. (1999). A Definition of Uncertainty Aversion. *Review of Economic Studies*, 66/3, 579–608.

—— (2001). Sharing Ambiguity. *American Economic Review*, *Papers and Proceedings*, 91, 45–50.

—— and Wang, T. (1994). Intertemporal Asset Pricing under Knightian Uncertainty. *Econometrica*, 62, 283–322.

—— and Zhang, J.-K. (2001). Subjective Probabilities on Subjectively Unambiguous Events. *Econometrica*, 69, 265–306.

Ghirardato, P., and Marinacci, M. (2002). Ambiguity Made Precise: A Comparative Foundation. *Journal of Economic Theory*, 102, 251–89.

GHIRARDATO, P., MACCHERONI, F., and MARINACCI, M. (2004). Differentiating Ambiguity and Ambiguity Attitude. *Journal of Economic Theory*, 118, 133–73.

GILBOA, I. (1987). Expected Utility with Purely Subjective Non-Additive Probabilities. *Journal of Mathematical Economics*, 16, 65–88.

—— and SCHMEIDLER, D. (1989). Maxmin Expected Utility with a Non-Unique Prior. *Journal of Mathematical Economics*, 18, 141–53.

GONZALEZ, R., and WU, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38, 129–66.

GRANT, S., and QUIGGIN, J. (2005). Increasing Uncertainty: A Definition. *Mathematical Social Sciences*, 49/2, 117–41.

HALEVY, Y. (2007). Ellsberg Revisited: An Experimental Study. *Econometrica*, 75/2, 503–36.

HALLER, H. (2000). Non-Additive Beliefs in Solvable Games. *Theory and Decision*, 49, 313–38.

HEATH, C., and TVERSKY, A. (1991). Preference and Belief: Ambiguity and Competence in Choice under Uncertainty. *Journal of Risk and Uncertainty*, 4, 5–28.

JAFFRAY, J.-Y., and PHILIPPE, F. (1997). On the Existence of Subjective Upper and Lower Probabilities. *Mathematics of Operations Research*, 22, 165–85.

KAHNEMAN, D., and TVERSKY, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263–91.

KELSEY, D., and MILNE, F. (1995). Arbitrage Pricing Theorem with Non-Expected Utility Preferences. *Journal of Economic Theory*, 65, 557–74.

—— and NANDEIBAM, S. (1996). On the Measurement of Uncertainty Aversion. University of Birmingham Discussion Paper.

KEYNES, J. M. (2004[1921]). *A Treatise on Probability*. New York: Dover Publications.

KLIBANOFF, P., MARINACCI, M., and MUKERJI, S. (2005). A Smooth Model of Decision Making under Ambiguity. *Econometrica*, 73/6, 1849–92.

KNIGHT, F. H. (1921). *Risk, Uncertainty, and Profit*. New York: Houghton Mifflin.

LO, K. C. (1996). Equilibrium in Beliefs under Uncertainty. *Journal of Economic Theory*, 71, 443–84.

MACHINA, M., and SCHMEIDLER, D. (1992). A More Robust Definition of Subjective Probability. *Econometrica*, 60, 745–80.

MARINACCI, M. (2000). Ambiguous Games. *Games and Economic Behavior*, 31, 191–219.

MUKERJI, S. (2000). A Survey of Some Applications of the Idea of Ambiguity Aversion in Economics. *International Journal of Approximate Reasoning*, 24, 221–34.

—— and TALLON, J.-M. (2001). Ambiguity Aversion and Incompleteness of Financial Markets. *Review of Economic Studies*, 68, 883–908.

—— —— (2004). An Overview of Economic Applications of David Schmeidler's Models of Decision Making under Uncertainty. In I. Gilboa (ed.), *Uncertainty in Economic Theory: A Collection of Essays in Honor of David Schmeidler's 65th Birthday*. London: Routledge.

NEHRING, K. (2006*a*). Decision-Making in the Context of Imprecise Probabilistic Beliefs. Working Paper.

—— (2006*b*). Is it Possible to Define Subjective Probabilities in Purely Behavioral Terms? A Comment on Epstein–Zhang (2001). Working Paper.

—— (2007). Imprecise Probabilistic Beliefs as a Context for Decision-Making under Ambiguity. Working Paper.

QUIGGIN, J. (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization*, 3, 323–34.

Ryan, M. J. (2002). What Do Uncertainty-Averse Decision-Makers Believe?. *Economic Theory*, 20, 47–65.

Sarin, R., and Wakker, P. (1992). A Simple Axiomatization of Non-Additive Expected Utility. *Econometrica*, 60, 1255–72.

Savage, L. J. (1954). *Foundations of Statistics*. New York: Wiley.

Schmeidler, D. (1989). Subjective Probability and Expected Utility without Additivity. *Econometrica*, 57, 571–87.

Segal, U. (1987). The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach. *International Economic Review*, 28/1, 175–202.

Tversky, A., and Fox, C. R. (1995). Weighting Risk and Uncertainty. *Psychological Review*, 102, 269–83.

—— and Wakker, P. (1995). Risk Attitudes and Decision Weights. *Econometrica*, 63, 1255–80.

Wakker, P. (2001). Testing and Characterizing Properties of Nonadditive Measures through Violations of the Sure-Thing Principle. *Econometrica*, 69, 1039–59.

Yaari, M. (1969). Some Remarks on Measures of Risk Aversion and their Uses. *Journal of Economic Theory*, 1, 315–22.

CHAPTER 5

THE NORMATIVE STATUS OF THE INDEPENDENCE PRINCIPLE

EDWARD F. MCCLENNEN

5.1 Introduction

One of the most important developments in decision theory in the second half of the twentieth century was the idea of evaluating prospects in terms of expected utility, of assigning a utility to each component of a gamble (either an outcome that involved no risk or itself a gamble) and then taking the utility of the gamble itself as equivalent to discounting the utility of each component by the probability with which it would occur and summing these products together. The utility of a gamble, then, is a linear function of its probabilistically discounted components. In the original theory, probabilities were presumed to be given exogenously (e.g. the probabilities of the sort that could be assigned to the various outcomes of a spin of a roulette wheel or the roll of dice), but subsequently, the theory was recast to provide also for well-defined subjective probabilities.

All of these constructions make appeal to what has come to be known as the independence principle. In what follows I shall survey the arguments that have been offered for what was taken as a fundamental axiom of rational choice. Of course, if the only issue were the extent to which decision-makers *do in fact* evaluate

risky alternatives in a manner consistent with the independence axiom, this would require a careful look at the empirical evidence for this. But my concern here is with the status of the axiom as *normative*: that is, as a principle to which a rational agent *should* adhere. My conclusion is essentially negative: I find none of the arguments for the normative status of the axiom to have much force, and hence conclude that the theory of expected utility and subjective probability is much more problematic than most have assumed.

5.2 Characterization of the Independence Principle

The cornerstone of the theory of expected utility and subjective probability, the independence principle, places a significant restriction on the ordering of options that involve risk or uncertainty (in suitably defined senses of each of these terms). For the case of options involving exogenously defined risk, one version of the axiom is as follows:

The Independence Principle (IND): Let P, P^*, and Q be any three prospects or gambles, and $0 < \lambda \leq 1$; then

$$\textit{if } P \, I \, P^* \textit{ then } \lambda P + (1-\lambda)Q \; I \; \lambda P^* + (1-\lambda)Q,$$

where "I" denotes indifference.

That is, substituting indifferent components for one another preserves indifference. Another version of the axiom to which repeated appeal has been made is:

Let P, Q, S, and T be any four prospects or gambles, and $0 < \lambda \leq 1$; then

$$\lambda P + (1-\lambda)T \; R \; \lambda Q + (1-\lambda)T \quad \textit{iff } \lambda P + (1-\lambda)S \; R \; \lambda Q + (1-\lambda)S,$$

where "R" denotes weak preference.[1]

IND invites particularization and reformulation in a variety of different ways. For the matters to be explored below, perhaps the most important particularization is where the components are not themselves lotteries, but "sure" outcomes

[1] The notations I utilize here are taken from Fishburn and Wakker (1995). That article contains an extremely helpful and for the most part insightful guide to the history of the utilization of various versions of the independence axiom, as well as a very comprehensive bibliography. For the issues to be discussed here, one cannot do better than start with their account. My only complaint is that they pass over the manifold criticisms that have been mounted against the various versions of this axiom. There is, moreover, something of a puzzle about their article: why do they entitle it the "invention" of the independence condition?

(e.g. amounts of money). Since an outcome involving no risk can be viewed as a "gamble" in which one gets that outcome with probability 1, IND yields directly the following:

Independence for Sure Outcomes (**ISO**): Let O_1, O_2, O_3 be any three sure outcomes (monetary prizes, etc.), and $0 < \lambda \leq 1$; then

$$O_1 \; I \; O_2 \Rightarrow \lambda O_1 + (1 - \lambda) O_3 \; I \; \lambda O_2 + (1 - \lambda) O_3.$$

The independence axiom is only one of two ways in which the key axiom of expected utility and subjective probability has been formulated. The other formulation, following Savage (1972[1954]), came to be known as the "sure-thing" principle (STP).[2] It was originally introduced in Friedman and Savage (1952) in the following manner (once again adjusting its formulation to our present notation):

> [S]uppose a physician now knows that his patient has one of several diseases for each of which the physician would prescribe immediate bed rest. We assert that under this circumstance the physician should, and unless confused, will, prescribe immediate bed rest whether he is now, or later, or never, able to make an exact diagnosis.
>
> Much more abstractly, consider a person constrained to choose between a pair of alternatives, R and R^*, without knowing whether a particular event E does (or will) in fact obtain. Suppose that, depending on his choice, and whether E does obtain, he is to receive one of four gambles, according to the following schedule:

	Event	
	E	$-E$
Choice		
R	P	Q
R^*	P^*	Q^*

The principle in sufficient generality for the present purpose asserts: if the person does not prefer P to P^* and does not prefer Q to Q^* then he will not prefer the choice R to R^*. Further, if the person does not prefer R to R^*, he will either not prefer P to P^* or not prefer Q to Q^* (and possibly both). We anticipate that if the reader considers this principle, in

[2] The earliest reference to what came to be known as the sure-thing principle, as far as I have been able to determine, occurs in a discussion by Savage of a decision situation in which risk is not well-defined—what has come to be known as a case of decision-making under conditions of uncertainty. Savage imagines an agent who is interested simply in maximizing *expected* income, but who is faced with a situation in which he cannot appeal to well-defined probabilities. Under such circumstances, Savage argues,

> . . . there is one unquestionably appropriate criterion for preferring some act to some others: If for every possible state, the expected income of one act is never less and is in some cases greater than the corresponding income of another, then the former act is preferable to the latter. This obvious principle is widely used in everyday life and in statistics, but only occasionally does it lead to a complete solution of a decision problem. (Savage 1951, p. 114)

In neither this article nor in the one he wrote with Friedman a year later does Savage characterize the principle in question as the "sure-thing" principle: that term appears to occur for the first time in Savage (1972[1954]).

the light of the illustration that precedes and such others as he himself may invent, he will concede that the principle is not one he would deliberately violate.

(Friedman and Savage 1952, pp. 468–9)

As Savage was to make clear two years later, STP is essentially a dominance principle that had been employed by many statisticians as an admissibility criterion.[3] As formulated above, it should be noted that it applies to cases in which the component entities, P, Q, etc., are themselves risky prospects or gambles. It can also, just like IND, be particularized to the case where outcomes are not gambles but sure amounts of money or other goods. In very general terms, the particular formulation to which appeal is made in a given axiomatic construction typically depends in part on the strength of the other axioms employed and in part on considerations of simplicity and/or formal elegance. But, as Fishburn and Wakker (1995) make clear, some version or other of independence or something that implies independence is invariably to be found in the constructions.[4]

5.3 Arguments for Independence

Broadly speaking, there have been five different arguments that have been used to defend the independence axiom. The first is simply that the independence axiom as a normative principle is intuitively true; that is, self-evident. The epistemological presuppositions of any appeal to a normative condition being self-evident are even more problematic than the axiom itself, and so I will leave this argument to one side and concentrate on the other four arguments. The second is that independence within the context of disjunctive options (gambles) makes sense because the problem of complementarity that arises in the case of conjunctive bundles of goods cannot arise in the context of disjunctive bundles. This is an argument most fully developed by the economist Paul Samuelson (1952). The third, which is due to the statistician Leonard Savage (1951, 1972[1954]), and Friedman and Savage (1952), turns on interpreting the independence axiom as a dominance condition. The fourth, due to John Broome (1991), is that what appear to be cases in which

[3] See Savage (1972[1954], p. 114). The principle can be recast in a form that is applicable to options defined in terms of some partition of n mutually exclusive and exhaustive events, and also applied to cases in which well-defined probabilities can be associated with each event in the partition. In any of its formulations, of course, one must presuppose that the choice of an option does not differentially affect the probabilities of the conditioning events. That is, the conditional probability of E_i, given choice of P, must be equal to the conditional probability of E_i, given choice of Q.

[4] Criticism of the axiom began almost immediately, principally in the form of various counterexamples. See, in particular, Allais (1953) and Ellsberg (1961). See also Allais and Hagen (1979), and MacCrimmon and Larsson (1979). I discuss below an example, due to Kahneman and Tversky (1979), which poses the same issue as was raised by Allais.

independence is violated, can be reinterpreted in such a way that no violation takes place. The fifth, first introduced by Raiffa (1961), defends independence by reference to what can be characterized as a pragmatic argument, in which it is claimed that agents who violate IND or STP can be caught in a "money-pump" situation, in which they will end up making a sequence of choices that will leave them at the end worse off, regardless of the turn of events, than they would have been if they had not violated the independence axiom.

5.4 Independence as Non-Complementarity

The introduction of the term "independence" appears to have been motivated by a perceived analogy with the economic concept of independent goods, in which the value of a conjunctive bundle of various quantities of different goods is an additive function of the value of the quantities of the various separate goods that make up that bundle. It is, of course, a well-known fact that independence with respect to the value of a bundle of commodities can fail. The value of the combination of x amount of one good and y amount of another good may not be equivalent to the sum of the value of x amount of the one good and the value of y amount of the other good. Failure of independence in such cases is said to be due to complementarity or interaction between the constituent commodities. That is, the value of one good may be enhanced or reduced in virtue of its being combined with some other good, as, for example, in the proverbial case in which white wine is said to complement fish, and red wine to complement beef. A given agent might well prefer white wine to red wine (when the two are considered in isolation), but prefer the combination of steak and red wine to steak and white wine.

Starting with von Neumann and Morgenstern (1953[1944]), however, one finds repeated appeal to the argument that such a problem of complementarity cannot arise in the case of what are disjunctive (as distinct from conjunctive) bundles of goods, i.e. lotteries over goods, and hence that the assumption of independence in this context is warranted. Here is how the argument emerges in their work (adjusting for the notation introduced above):

> By a combination of two events we mean this: Let the two events be denoted by P and Q and use, for the sake of simplicity, the probability 50%–50%. Then the "combination" is the prospect of seeing P occur with probability 50% and (if P does not occur) Q with the (remaining) probability of 50%. We stress that the two alternatives are mutually exclusive, so that no possibility of complementarity and the like exists.
>
> (von Neumann and Morgenstern 1953[1944], p. 18)

Table 5.1. Ellsberg's first example

	(30)	(60)		Range of expected monetary return
	Red	Black	Yellow	
P	\$100	0	0	$33\frac{1}{3}$
Q	0	100	0	0 to $66\frac{2}{3}$
P^*	\$100	0	100	$33\frac{1}{3}$ to 1
Q^*	0	100	100	$66\frac{2}{3}$

Samuelson (1952, pp. 672–3) explicitly marks the analogy and, while acknowledging that complementarities can arise in the case of (conjunctive) bundles of goods, insists that the nature of a disjunctive (or stochastic) bundle, in which one is to get just one of the disjuncts, makes it plausible to impose independence as a condition on preferences for gambles.

The argument for non-complementarity in the case of disjunctive bundles, however, is not all that compelling. Disjunctive bundles may not be subject to the problem of commodity complementarity, but this does not rule out the possibility of forms of "complementarity" that are special to disjunctive prospects.[5] In fact, there are two much discussed counterexamples to the independence principle which serve to suggest a distinct type of complementarity that can arise in the concatenation of risky prospects. The examples are due to Ellsberg (1961) and they are directed not at IND but at Savage's version, STP.

In the first example, Ellsberg considers a situation in which the agent is to choose between various gambles based upon drawing a ball at random from an urn that contains red, black, and yellow balls, where one knows that there are thirty red balls in the urn, and that sixty are either black or yellow, but the relative proportion of black and yellow is unknown (see Table 5.1). Since the probabilities of the conditioning events are only partially defined, one cannot associate with every such option an unambiguous expected monetary return. But, as the column to the far right serves to indicate, one can at least specify the possible range of such values.

Ellsberg notes that many people prefer P to Q, while preferring Q^* to P^*. He also notes that the following rule of evaluation generates this preference ordering: rank options in terms of increasing minimum expected monetary return. Now note that the pair of options $\{P, Q\}$ differs from the pair of options $\{P^*, Q^*\}$ only with respect to the payoffs in the event that a yellow ball is drawn. But in each case the amount to be received if a yellow ball is drawn is constant. Thus, once again with an appropriate repartitioning of the states, STP applies, and requires that P

[5] The possibility of complementarity is discussed in Manne (1952), Allais and Hagen (1979, pp. 80–106), McClennen (1983), and Loomes and Sugden (1984).

Table 5.2. Ellsberg's second example

	H	T
PQ	*P*	*Q*
QQ	*Q*	*Q*
PP	*P*	*P*
QP	*Q*	*P*

P = [$100, *E*; $0, *not-E*]
Q = [$0, *E*; $100, *not-E*]

be preferred to Q just in case P^* is preferred to Q^*, contrary to the described preferences.

Once again, one can interpret what is happening here in terms of the notion of complementarities with respect to the value of disjunctions of outcomes. The shift from a situation in which, under the condition of drawing a yellow ball one receives $0, regardless of which act is chosen, to a situation in which, under the same chance conditions, one receives $100, regardless of which act is chosen, results in "contamination" (to use Samuelson's term for complementarity). And, once again, there is no mystery here as to how this happens. The person who adopts Ellsberg's rule can be characterized as uncertainty (or, as Ellsberg himself terms it, "ambiguity") averse: uncertain prospects (as distinct from those whose associated expected return is well-defined) are discounted to their minimum expected return. Although one can think of P^* and Q^* as resulting from a modification of P and Q, respectively—in each case the addition of $100 to the payoff when a yellow ball is drawn—this proportional increase in payoffs has a differential impact with regard to uncertainty or ambiguity. In the case of choice between P and Q it is Q that presents an uncertainty; but given the substitution, it is now the counterpart to P, namely, P^* (and not the counterpart to Q, namely Q^*) that presents the uncertainty.

In the same article Ellsberg offers another counterexample, in which the agent is to choose between a pair of gambles, and then between another pair of gambles (see Table 5.2).

E is some event whose likelihood is completely unknown, and (H, T) are the two possible outcomes of a toss of a fair coin, so that a given subject can be presumed to assign a subjective probability of 1/2 to each of H and T. The pattern of outcomes here is such that by IND or STP, *PQ* is weakly preferred to *QQ* if and only if *PP* is weakly preferred to *QP*. On closer inspection, however, *PQ* and *QP* each offer one a 50–50 chance of getting $100, while *QQ* and *PP* each offer one only completely indeterminate odds of getting $100. Many indicate that they prefer even odds to indeterminate odds, and thus prefer *PQ* to *QQ*, and prefer *QP* to *PP*, in violation of IND and STP.

What has been isolated here is something that is distinct from the complementarity that arises in connection with conjunctive bundles. Here, once again, there is no question of some sort of interaction between prizes, both of which are to be received. It arises within the context of a disjunctive concatenation of prizes or goods, and turns on the implication of combining both well-defined and indeterminate odds. But it bears all the marks of being a type of complementarity. The gambles P and Q, when they are combined in various ways—*PQ, QQ, PP,* and *QP*—sometimes result in a combination that assures one determinate even-chance odds (in the case of *PQ* and *QP*), and sometimes result in complete uncertainty of odds (in the case of *PP* and *QQ*). For one who is uncertainty (or ambiguity) averse, then, it does make a difference as to which of the (indifferent) components are combined with each other. The implication of both of the examples presented above is quite clear. One cannot infer that STP is a plausible condition to impose on the ordering of disjunctive bundles simply by appeal to the consideration that complementarities of the type that arise in connection with conjunctive bundles cannot arise in connection with disjunctive bundles. That argument is a nonstarter, for it ignores a kind of interaction that is unique to disjunctive bundles, and that forms an intelligible basis for discriminating between prospects, if one is concerned with uncertainty (or ambiguity).

5.5 Sure-Thing Reasoning

There is no question that the dominance idea to which STP appeals is intuitively very plausible. Recall, once again, the argument presented by Friedman and Savage (1952) in support of STP. STP mandates preference for P over Q if, no matter what the turn of events, the outcome of choosing P is at least as good as the outcome of choosing Q and, for *some* turn of events, the outcome of choosing P is better than the outcome of choosing Q. And that seems plausible enough. In effect, choice of P over Q promises a "sure thing" with respect to consequences: by choosing P you cannot do worse, and may end up doing better, than if you choose Q.

Despite the fact that many have taken this to be a decisive consideration in favor of STP, this line of reasoning is also flawed. STP is *very* strong. It is framed with respect to the outcomes that can be associated with *arbitrarily selected* partitions of conditioning states. The principle requires that if there exists *any* event partition for which the outcomes of P dominate the outcomes of Q, then P must be preferred to Q. In particular, then, the principle is not limited in its application to outcomes that can be characterized as sure or riskless.

This raises a substantial issue. Consider the following problem, which is due to Kahneman and Tversky (1979):

P [\$2400, E or −E]
Q [\$2500, E; \$0, −E]
P^* [P, F; \$0, −F] = [[\$2400, E or −E], F; \$0, −F]
Q^* [Q, F; \$0, −F] = [[\$2500, E; \$0, −E], F; \$0, −F]

Once again, many report that they prefer P to Q, but Q^* to P^*. In the case of P^* and Q^*, then, we have a partition for which the associated outcomes satisfy the conditions for dominance: P preferred to Q, and \$0 at least as good as \$0. Thus, by STP the agent should rank P^* over Q^*. But what qualifies these outcomes as relevant for the purposes of assessing the choice between P^* and Q^* from a "sure-thing" perspective? Within the framework of a *finer* partitioning of events—and one that is an *explicit* feature of the problem—it is simply not true that one does at least as well by choosing P^* as by choosing Q^*, regardless of the turn of (all relevant) events. By inspection, the outcome of Q^* in the event that both E and F occur is \$2500, which is, by hypothesis, strictly preferred to any of the possible outcomes of P^*. I do not mean to suggest, of course, that application of STP can be undercut in such cases simply by displaying *some* other partition of events such that preferences for the outcomes under that partition fail to satisfy the antecedent condition of STP. Rather, the issue here concerns the propriety of appealing to a partition under which the antecedent conditions are satisfied even though there exists an explicit *refinement* of that very same partition for which the antecedent conditions are *not* satisfied. If there is such an explicit refinement, then by reference to the consequences under that (refined) description, it is no longer clear what the force is of an appeal to dominance considerations.

Savage himself was well aware of the full scope of STP, and explicitly raised the question of whether it might be appropriate to restrict it to cases where the outcomes themselves are not defined in probabilistic terms. Focusing on the case of event-defined gambles over "sure" amounts of money, he rejects this suggestion on the following grounds:

> A cash prize is to a large extent a lottery ticket in that the uncertainty as to what will become of a person if he has a gift of a thousand dollars [for example] is not in principle different from the uncertainty about what will become of him if he holds a lottery ticket.
>
> (Savage 1972[1954], p. 99)

This amounts to denying that there is anything like a bedrock level of certainty. On this account, *it is risk all the way down*. Suppose, however, we grant this, and hence understand that one cannot distinguish a more restrictive version of STP. What makes this an argument for accepting STP rather than rejecting it?

The agent can acknowledge, of course, that if he chooses Q^* over P^*, then he *moves through* a state in which a dominance relation obtains. More specifically, if we

think of the *F*-events as occurring first, followed by the *E*-events, then no matter what the outcome of the *F*-events, and before the *E*-events are run, the prospect the agent then faces, if he has chosen Q^*, is dispreferred to, or no better than, the prospect he would then be facing if he had chosen P^*. He could argue, though, that this is something that he suffers only *en passant*, and since he is concerned only with final outcomes and their probabilities, it is of no consequence to him.

Now, Savage's reply, as reported above, is that, in effect, it is always a matter of what we face *en passant,* since it is risk all the way down. This means, however, that any problem involves choice between gambles, and thus that the agent can never be sure he will always do better choosing one way rather than another. But, then, granting Savage's point, why not turn it upside down and regard it as undercutting the whole idea of an appeal to dominance with respect to outcomes?

Perhaps we need not take such a drastic position. Any principle such as STP must be interpreted as constraining preferences among alternatives, *under a given description* of those alternatives. If the agent has not refined his description of certain component gambles, and treats them as simply possible outcomes over which he has a preference ranking, then it can be argued that it is appropriate to appeal to dominance considerations. Suppose, though, that he has refined his description of those outcomes—recognizing explicitly the nature of the further risks to be encountered. In such a case, since at *that* level of discrimination the principle is revealed not to apply, it is unclear what force there is to an argument that invokes dominance at the coarser level of description.

I conclude, then, that while sure-thing considerations provide a highly plausible basis for a version of STP that is framed with respect to riskless outcomes (relative to some base description), there is little to support the extension of this line of reasoning to the full-blown principle STP. STP, no less than the IND, is subject to serious question.

5.6 Broome's Argument

John Broome has offered a very different defense of the independence axiom, one that neatly handles all of the counterexamples I have explored above. By way of illustration, consider once again the second of Ellsberg's counterexamples in Table 5.2 above. The pattern of outcomes here is such that by IND or STP, *PQ* is weakly preferred to *QQ* if and only if *PP* is weakly preferred to *QP.* On closer inspection, however, *PQ* and *QP* each offer one a 50–50 chance of getting $100, while *QQ* and *PP* each offer one only completely indeterminate odds of getting $100. If one prefers known even odds to indeterminate odds, as many would, then one will prefer *PQ* to *QQ*, and prefer *QP* to *PP*.

Table 5.3. Broome's reinterpretation of Ellsberg's second example

	H	T
PQ	*P* An even chance of $100	*Q* An even chance of $100
QQ	*Q* Uncertain odds of $100	*Q* Uncertain odds of $100
PP	*P* Uncertain odds of $100	*P* Uncertain odds of $100
QP	*Q* An even chance of $100	*P* An even chance of $100

Must this be seen as a violation of IND and STP? Broome would suggest that it need not be seen in that way.[6] Instead we may suppose that the appropriate description of the four options is as given in Table 5.3.

Under this description, *PQ* given H and *PP* given H are now different: both involve being exposed to *P*, but in the former case this also involves an even chance of getting $100, while in the latter case what is involved is an indeterminate chance of getting $100. Similar remarks apply to *QQ* and *QP*. The conflict with IND and STP is resolved, then, by a more detailed description of what happens in each case. That is, they no longer apply given the way in which the distinct possibilities are descriptively individuated.

Broome is careful to note that we would not want to allow just any more detailed description to be invoked in such a situation, for in that case IND and STP would end up having no bite—descriptions could be employed in such a way that these conditions would for all intents and purposes never apply. But whether the odds to which one is exposed are (determinately) even-chance or indeterminate, is plausibly a relevant descriptive item. Now, on first consideration that *PP*, for example, yields even-chance odds of getting $100 is a property of the whole gamble in question, and not a property of one of the states that results from the choice of *PP* and the coin coming up heads. But Broome can deal with this by pointing out that it is still true of the state in which *PP* is selected and heads comes up, that the decision-maker arrived at that state as the result of exposing himself to even-chance odds of getting $100. In effect, Broome can propose to individuate the various state possibilities in such a manner that any features of each gamble taken as a whole can be taken into account in the step-by-step comparison of outcome states. If you will, each outcome carries with it, or includes, the path by which it was generated.

There are two things that must be kept in mind. First, the construction that yields a theory of subjective expected utility (e.g. Savage's version of the

[6] My example here is not the one that Broome uses. He takes on a somewhat parallel example, due to Diamond (1967), that has to do with achieving fairness by flipping a coin to determine who gets a prize.

construction) precludes distinguishing between completely uncertain odds and even-chance odds, as in Ellsberg's second counterexample. The subjective expected utility construction requires that a rational decision-maker assign determinate probabilities to all conditioning events. Thus, introducing an additional justifier that turns on distinguishing between well-defined and uncertain odds, in order to avoid a violation of STP, cannot be squared with subjective expected utility theory.

Second, what is unclear about Broome's analysis is what happens if the method of evaluation involves not just dealing with isolated cases in which, e.g., uncertainty of odds is a feature of some gambles, but where gambles are systematically evaluated in a way that takes into account not just expected return but also some measure of the dispersion of the gamble. For example, a standard approach to the evaluation of portfolios of stocks involves looking not just at the expected return of a portfolio, but in terms of a two-termed function of expected monetary (mean) value and some measure of the dispersion of monetary values. The one that has been discussed the most involves appeal to a two-termed linear function of mean value and variance (Markowitz 1952). The standard argument is that in this case an expected utility function can be constructed, but that utility as a function of monetary amount must be quadratic in form.[7] This result carries with it a number of bothersome implications. In order to ensure that the utility of money is an increasing function of monetary amount, the utility function must be bounded; also, the cash equivalence of a gamble will turn out to decrease as one gets wealthier; and finally, the preference order of any set of options with equal expectations must either minimize or maximize variance (Pollatsek and Tversky 1970, pp. 549–50).

Here the question is not the existence of odd cases that can be explained away by showing that IND or STP are really not violated—that they are just not applicable in the cases in question. Rather, we are forced to work with a utility function with rather questionable features. The alternative, which seems not to have not been given the hearing it deserves, is that one can reject expected utility theory in cases in which mean/variance evaluation is appropriate, and so reject any application of either IND or STP.[8] My own sense is that this is the real issue that needs to be addressed: namely, that expected utility theory fits poorly with a completely plausible mean/variance model of evaluation. As far as I can see, Broome's method of explaining away certain counterexamples leaves this issue untouched.

[7] For a very interesting, and much more formal, exploration of measures of risk that are sensitive to both mean values and dispersion measured in terms of variance, and the problems posed by trying to reconcile such an approach with expected utility theory, see Pollatsek and Tversky (1970).

[8] Pollatsek and Tversky themselves make this suggestion. See Pollatsek and Tversky (1970, p. 550).

5.7 The Dynamic Money-Pump Argument

Consider the following variation on one of the examples discussed above:

P ($2400, 1)
Q ($2500, 33/34 ; $o, 1/34)

and

P^* ($2400, 34/100; $0, 66/100)
Q^* ($2500, 33/100; $0, 67/100)
R^* ($2401, 34/100; $1, 66/100)

And suppose that the agent prefers P to Q, but Q^* to P^*, in violation of the independence axiom. It is plausible to suppose that one could increase the payoffs in P^* just marginally enough that the augmented version of P^* will (obviously) be preferred to P^* but dispreferred to Q^*. Let this augmented version of P^* be, for example, R^* =[$2401, 34/100; $1, 66/100]. Now suppose that someone offers the agent rights to R* for a fair price of $$x$ (what that price is doesn't matter), the agent accepts the trade, and then is offered the sequential choices shown in Figure 5.1.

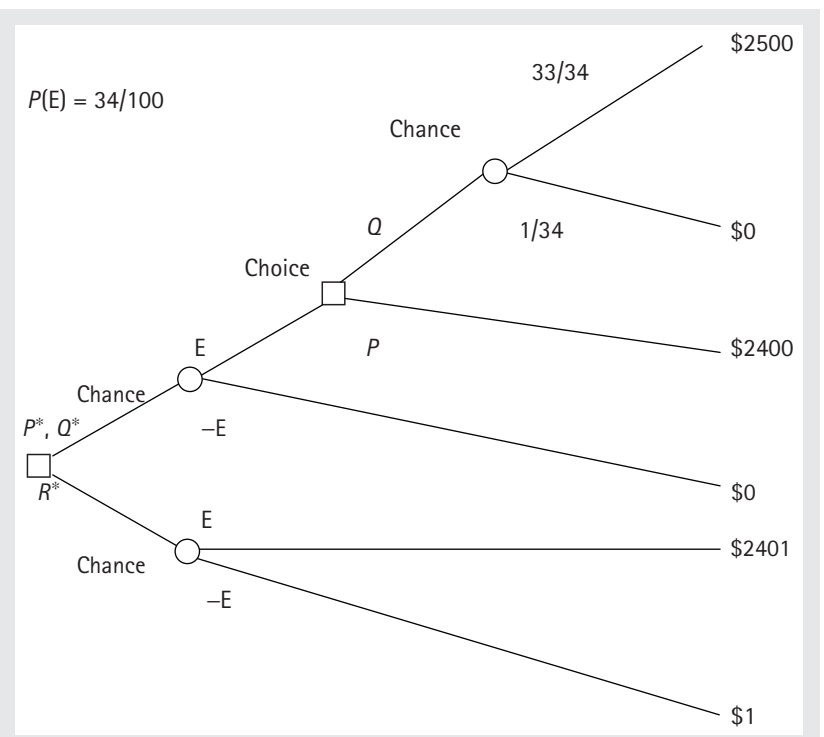

Fig. 5.1. A sequential choice example.

The agent, then, has given up $\$x$, and secured rights to the outcome of R^*. However, he is now offered the opportunity to trade again—to exchange rights to R^* for an opportunity to have either Q^* or P^* in exchange for R^*, and a small amount of money, say $\$e$. We may suppose that the agent will accept this trade, since he prefers Q^* to R^*. Now "nature" makes its move (either E or −E), and this means that the agent either receives \$0 or has a chance at either P or Q. In the latter case, however, by hypothesis, he prefers P to Q, and thus in this event he will choose P, and receive \$2400. So he has now paid out $\$e$, and given up R^* in order to get something which is *dispreferred* to R^*, namely P^*. More specifically, no matter how nature moves, he is out by \$1, and the person trading with him has gained a sure \$1, no matter whether E or –E occurs. But the person with whom he is trading can repeat this cycle of exchanges over and over, each time offering him R^*, and then offering to exchange R^* for the opportunity of Q^*. Thus, he can be "pumped" for all the money he has.[9] Again, however, I think the argument does not go through smoothly. Having exchanged R^* for the opportunity of Q^* or P^*, the agent must be supposed to be planning to choose Q rather than P, if and when he comes to the second choice point. The only way the agent could get into the money-pump situation would be if, on those occasions in which the agent confronts the second choice point, he or she forgets what was planned. If the agent were myopic in the sense of approaching each subsequent choice point as if the plan settled upon earlier no longer meant anything, the agent could be exploited. But if the agent settles upon a *plan* at the first choice point, he or she could *resolutely* execute the balance of the plan if and when the second choice point becomes available, by choosing Q rather than P.[10] The only way that this would not be open to the agent would be if the agent must choose, at each choice point in any decision tree, as if that choice were one to be made *de novo* regardless of what he or she had decided to do in the past. But why must the agent choose in this manner? Surely the agent is capable of choosing Q over P in the context of completing the plan to execute Q^* rather than simply regard himself in the same situation as he would be in, were he faced with a *de novo* choice between P and Q.

5.8 Conclusions

As far as I have been able to determine, the arguments discussed above are all of those that have been offered in defense of the independence axiom. Since I

[9] Variations on this kind of critique are to be found in Raiffa (1961, 1968). For a fuller discussion, see McClennen (1990, sects. 5.5 and 10.4).

[10] See McClennen (1990, chs. 9–12), for a full exposition of the concept of the resolute execution of a plan one has adopted.

find them all questionable, I conclude that the appeal to the independence axiom is questionable. This point, moreover, is more significant than just an abstract theoretical conclusion. Those who are engaged in making recommendations to investors, and those who are contemplating taking or not taking the advice of their advisors, have understood that a sound approach to investing involves balancing risk against expected monetary return, where the amount of risk one is willing to bear typically varies from one individual to the next. It is this approach that deserves much more exploration, rather than continuing the dogged attempt to defend the independence principle in any way possible.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21/4, 503–46.

—— and Hagen, O. (eds.) (1979). *Expected Utility Hypothesis and the Allais Paradox*. Boston: D. Reidel Publishing Company.

Broome, John (1991). Rationality and the Sure-Thing Principle. In Gay Meeks (ed.), *Thoughtful Economic Man*, 74–102. Cambridge: Cambridge University Press.

Diamond, P. A. (1967). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility: Comment. *Journal of Political Economy*, 75, 765–6.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643–69.

Fishburn, P., and Wakker, P. (1995). The Invention of the Independence Condition for Preferences. *Management Science*, 41/7, 1130–44.

Friedman, M., and Savage, L. J. (1952). The Expected-Utility Hypothesis and the Measurability of Utility. *Journal of Political Economy*, 60/6, 463–74.

Kahneman, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47/2, 263–91.

Loomes, G., and Sugden, R. (1984). The Importance of What Might Have Been. In O. Hagen and F. Wenstop (eds.), *Progress in Utility and Risk Theory*, 219–35. Boston: D. Reidel Publishing Company.

MacCrimmon, K. R., and Larsson, S. (1979). Utility Theory: Axioms Versus "Paradoxes". In M. Allais and O. Hagen (eds.), *Expected Utility and the Allais Paradox*, 333–409. Boston: D. Reidel Publishing Company.

Manne, A. S. (1952). The Strong Independence Assumption—Gasoline Blends and Probability Mixtures (with discussion). *Econometrica*, 20, 665–9.

Markowitz, H. M. (1952). Portfolio Selection. *Journal of Finance*, 7/1, 77–91.

McClennen, E. F. (1988). Sure-Thing Doubts. In P. Gärdenfors and N.-E. Sahlin (eds.), *Decision, Probability and Utility*: *Selected Readings*, 166–82. Cambridge: Cambridge University Press.

—— (1990). *Rationality and Dynamic Choice: Foundational Explorations*. New York: Cambridge University Press.

Pollatsek, A., and Tversky, A. (1970). A Theory of Risk. *Journal of Mathematical Psychology*, 7, 540–53.

Raiffa, H. (1961). Risk, Ambiguity, and the Savage Axioms: Comment. *Quarterly Journal of Economics*, 75, 690–4.

—— (1968). *Decision Analysis*. Reading, MA: Addison-Wesley.

Samuelson, P. A. (1952). Probability, Utility, and the Independence Axiom. *Econometrica*, 20, 670–8.

Savage, L. J. (1951). The Theory of Statistical Decision. *Journal of the American Statistical Association*, 46, 55–67.

—— (1972[1954]). *The Foundations of Statistics*, 2nd edn. New York: Dover.

von Neumann, J., and Morgenstern, O. (1953[1944]). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

CHAPTER 6

RATIONALITY AND INTRANSITIVE PREFERENCE

FOUNDATIONS FOR THE MODERN VIEW

PAUL ANAND

6.1 Introduction

Once considered a cornerstone of rational choice theory, the status of transitivity has been dramatically reevaluated by economists and philosophers in recent years. For although normative, technical, and empirical issues used to stand or fall together in justifications of axiomatic assumptions, it has become increasingly clear that these are distinct issues, which call for different kinds of resolutions. We know that many of the conventional mathematical results in economics are robust to relaxing traditional assumptions like transitive preference (see, for instance, Gale

I am indebted for comments over many years to a number of colleagues in Oxford and elsewhere and especially to participants of the Foundations of Decision Theory seminars organized by Michael Bacharach and Susan Hurley. In addition I want to thank Clemens Puppe, Alan Hájek, and especially Robin Cubitt for comments on earlier drafts as well as contributors to a London School of Economics Choice Group seminar in 2007. The chapter draws heavily on papers collected together in my monograph, and the usual caveat applies.

and Mas-Collel (1975); Shafer (1976); Kim and Richter (1986); and Sugden (2003)); but that tells us little, if anything, about whether *rational* agents should be allowed to have preferences that are sometimes intransitive. Similarly, we know that experimental subjects sometimes exhibit preferences which are not transitive; but even when such violations are systematic, replicable, and made by informed, intelligent people, we do not take this as *proof* of the normative status of such behavior even if we allow that it might be suggestive.[1]

Whilst our primary interest here will be in questions of rationality, it is worth noting the role that transitivity plays in formal theories of choice. A natural starting point can be found in the representation and uniqueness theorems of von Neumann and Morgenstern (1944), which axiomatize expected utility and show that if certain assumptions hold, of which transitivity is one, then there exists a utility function, u(.), which maps each choice object (in a set C of such objects) onto the real line, Re, such that Pab if and only if $u(a) > u(b)$.[2] In this case, u: C $\rightarrow$ Re. Fishburn's innovation is to allow for context at the axiomatic level by defining what in effect is a generalized utility function (at the level of observed behavior) on a product space of alternatives. Although his so-called skew-symmetric bilinear theorem has a comparatively limited domain of application, as it is defined for only two alternatives, his approach does demonstrate the significant fact that it is possible to axiomatize intransitive behavior. The formal results, Fishburn (1982, 1984), which assume continuity, convexity, and symmetry, show the existence of a mapping u: C x C $\rightarrow$ Re, where u is defined on the product space of lottery pairs, C x C, which is then mapped onto the real-number line. To identify the utility of an option, one needs additional information about the opportunity set. Fishburn's generalized "utility" function is skew-symmetric, so that $u(a, b) = -u(b, a)$, and is bilinear just in the sense that u(. , .) is linear in both arguments, and his work provides a useful, axiomatic complement to substantive theories of decision-making which argue for the importance of contextual factors, like regret, disappointment, and comparisons in the determination of choice and the evaluation of welfare.

Whether rational preferences must always be transitive is a normative question best addressed directly, so this chapter assesses the most significant arguments in the debate. The position I explore is what has been called "the modern view" and holds, simply, that it is perfectly possible for rational agents to have intransitive preferences—albeit for reasons that may be different from those advanced when explaining empirical violations of transitivity if and when they occur.[3] The structure of the chapter is as follows. Section 6.2 considers the logical equation of rationality with transitivity, whilst Section 6.3 evaluates the constitutional

[1] See also Guala (2000) on the evaluation of normative claims.

[2] Pab denotes "a is preferred to b" and u(.) is the utility of an operator.

[3] Further support for the view can be found in Bauman (2005), Bell (1982), Fishburn (1988), Gendin (1996), Hansson (2001), Hughes (1980), Machina (1989), Mandler (2005), May (1954), McClennen (1988), Mongin (2000), Putnam (1995), Rambo (1999), Sen (1997), Schick (1986), and Walsh (1996).

(semantic) argument as outlined by Donald Davidson. We then go on to examine the money-pump argument dealing with the point of time constraint interpretation of intransitive preference in Section 6.4. Section 6.5 provides an overview of examples that have been developed to demonstrate the reasonableness of intransitive choice behavior. How we describe the choice setup plays a significant role in the context of such thought-experiments, and the section therefore also considers the conditions under which the question of whether preference should be transitive could just be a matter of language choice. Finally, Section 6.6 returns to the money-pump argument and offers an overview of arguments that attempt to persuade us that rational agents should have transitive preferences in sequential choices.

6.2 Intransitivity as Inconsistency—The Logical Case

There is a sense in which intransitive choice appears to be a form of error. Transitive choices are just a matter of logic, and it is natural to think that their opposite is illogical. Indeed the very title of Georg von Wright's (1963) *The Logic of Preference* makes this association explicit.[4] One of the earliest explorations of the idea, and one that helps to clarify the difficulties can be found in Tullock (1964, p. 403):

> The proof of intransitivity is a simple example of *reductio ad absurdum*. If the individual is alleged to prefer A to B, B to C, and C to A, we can enquire which he would prefer from the collection of A, B and C. *Ex-hypothesi* he must prefer one, say he prefers A to B or C. This however contradicts the statement that he prefers C to A, and hence the alleged intransitivity must be false.

If the agent is assumed to prefer B or C, then a similar sort of contradiction can be produced. But what really has been proved? It seems that we have (using lower case for entities and upper case for the relation) a proof which holds that a statement of choice from the feasible set {a, b, c} contradicts a statement in the set of possible binary preference rankings {Bab, Bbc, Bac} where B denotes binary preference.[5] Strictly speaking, any choice from the threesome would indicate something about ternary relations (which we denote with predicate T). Any such ranking would belong to the set {Tabc, Tacb, Tbca, Tbac, Tcab, Tcba}, none of which is an element of {¬Bab, ¬Bbc, ¬Bac}, which one must be if the contradiction is to be established.

[4] See also Broome's (1991) book in which he uses such ideas to argue for a structure of "the good".

[5] Strict preference is assumed for sake of exposition: nothing turns on this, but the argument is more cumbersome when taking care of possible indifference.

The problem here, and elsewhere as we shall see, is that the attempted proof confounds two- and three-place predicates. In general, such relations are different: m divided n and "Lucifer is the friend of Eve" are propositions involving binary relations, whilst 'r, the integral part of m divided by n' and 'Lucifer is the mutual acquaintance of Adam and Eve" are propositions involving ternary relations. If it is easy to use $\succ$ as a symbol for strict preference, then it is easy to think of the relation $>$ defined over real numbers as an analog, but this is suggestive of properties in a way that predicate notation is not. To accept such properties without further examination amounts to question-begging, though it may be that a link between binary and ternary preferences can be constructed, and that this link serves to deliver the *demonstrandum*. So let us examine the plausibility of such a link.

One principle which would complete the proof is known as contraction consistency (property α; see Sen 1970). The idea is that if you have a preference between two choice objects, this should be unaffected if other items are removed from the choice set. Formally, we might state the assumption thus:

$$\text{Tabc} \Rightarrow \text{Bab} \wedge \text{Bbc} \wedge \text{Bac}.$$

Contraction consistency lets one infer transitive binary preferences immediately, and this leads to the not entirely obvious position that the appeal of transitivity may hinge on our acceptance of contraction consistency. If it is rational to have complete ternary preferences and to abide by contraction consistency, then it would be rational to have transitive pairwise preferences. The normative question has not been solved—it has merely been relocated: must rational agents order their preferences so as to abide by contraction consistency (ignoring the usual objections to completeness)? Anand (1987, 1993*a*, 1993*b*), Arrow (1979), Kirchsteiger and Puppe (1996), Levi (1986), Loomes and Sugden (1982), Sen (1985), and Sugden (1985) all provide reasons why rational agents should not be so bound, and one reason that recurs turns on noting that choice depends on the context in which it is made (the term "state dependency" is used more frequently perhaps, but "context dependency" is more general and allows for contextual factors that vary with choice objects as well as outcome states).[6] Choices made from differing opportunity sets can have quite different meanings, for reasons to do with the *gestalt*, just as they might draw on very different forms of justification or explanations. Such arguments are well understood by anthropologists, and to see their force in undermining contraction consistency, consider the following scenario. First, suppose that you are being asked to say whether you think clinical patients should normally be able to see their medical records and that your preference is:

$$\neg \text{ access to all records} \succ \text{ access to all records} \qquad (1)$$

[6] See also in Putnam (1995) and Hansson (2001) for recognition of the view.

Say you think that whilst transparency is important, to provide access to all records would accelerate the development of a litigious environment in health care and encourage doctors to behave strategically in a manner that, overall, would be deleterious to patient health care. Now suppose a third option, "access to records electronically stored", is introduced. In that case, you might feel differently. Perhaps access to electronically stored medical records would satisfy the demands of natural rights without impinging unduly on clinicians' time, and so be preferred to the no access option. On the other hand, it might be claimed that access to a subset of medical records would give clinicians an incentive to minimize the amount of information that was stored electronically, and under these circumstances a person might feel it preferable to allow access to all records after all. In that case, the preference would take the following form:

$$\text{access to all records} \succ \text{access to records stored electronically} \succ \neg\text{access to all records} \qquad (2)$$

To be clear, this is not an argument that an agent aspiring to rationality must have the constellation of preferences implied by the conjunction of (1) and (2). All we need claim is that, were a person to have such preferences, for reasons such as those outlined, it would hardly be fanciful to say that they were intelligible and coherent. Indeed, one might hold that the difficulty is in finding any grounds on which (1) and (2) could be deemed irrational. However, if contraction consistency itself is not an axiom of rational choice, then Tullock's attempted logical argument remains a *non sequitur*.

6.3 Intransitivity as Inconsistency—The Semantic Case

There is a second, superficially weaker, though more intuitively appealing, view which suggests that transitivity is constitutionally embedded in the meaning of preference. An elegant expression of this argument can be found in, if it is not due to, Davidson (1980, p. 237) where he writes:

> The theory . . . is so powerful and simple, and constitutive of concepts assumed by further satisfactory theory . . . that we must strain to fit our findings, or interpretation, to fit the theory. If length is not transitive, what does it mean to use a number to measure length at all? We could find or invent an answer, but unless or until we do, we must strive to interpret "longer than" so that it comes out transitive. Similarly for "preferred to".

The core of the argument lies in beliefs about the theory of measurement. Certainly it would be difficult to imagine how something might have more than one

numerical length—even in Lewis Carroll's Wonderland, Alice's different heights occupied different points in time. So if length is the kind of attribute that can be represented by a single real number, and the comparative relation "is longer than" can be modeled by the relation "is greater than" as defined over real numbers, it seems inevitable that length will be transitive.

All this we can admit, and yet ask what it has to do with preference: "similarly" seems to be the essence of Davidson's reply. The crux of his argument is not the embeddedness of transitivity in the relation "is longer than" but rather the assertion of a double analogy, between relations (being preferred to and being longer than), on the one hand, and the way both relations come to have the property of being transitive, on the other. In both cases, transitivity is alleged to be a property embedded in the meaning of the relation, and not, for example, a property that either relation just happens, empirically, to have.

In general, the modern practice is to avoid for purposes of establishment, where we can, the use of analogies, and there is no decisive solution to the problem of how we should evaluate analogies such as the one upon which Davidson draws. One possible evaluation procedure is to see if there are other analogies that might be drawn and either confirm, or question, the claim made. We might suggest that the determination of a consumer's preference between two commodities, say, is like the judgment of a competition in which players or teams compete in pairwise fashion. However, we know that in such competitions, players are not necessarily ranked transitively. For example, in football, a first division team, may beat a second division team who in turn may beat a third division team, and yet the third division team could be "giant-killers" and beat the first division team. This is hardly a fanciful view requiring large leaps of the imagination: I do not wish to suggest that it provides a metaphor for preference which is more appealing than "length measurement" but merely that both metaphors are plausible and intelligible—which is all that is required to undercut the constitutional case as it currently stands.

6.4 The Money Pump

Of all the attempts to argue for transitive preference, the money pump has perhaps been most widely used, and it can be thought of as a popular *reductio ad absurdum* (related to the Dutch Book argument) purporting to show that an intransitive agent will give up some wealth for no reward. It begins by supposing that an agent has an intransitive set of binary strict preferences thus: Pab (3), Pbc (4), and Pca (5).[7] If these preferences can be translated directly into behavior, then a person could be

[7] Pab is read "a is strictly preferred to b", and so on.

made to trade a for c and pay a small positive finite sum by (5), to trade c for b and pay a further sum by (4), and to trade b for a whilst paying a third sum by (3). We could repeat this cycle *ad infinitum* until the person was pumped dry of all their wealth; but even if we did not, going through one cycle to arrive at an identical holding of a with reduced wealth is irrational, so the story goes.

If, however, the agent were not told at the start that the decision problem was one in which a sequence of opportunity sets was to be presented over time, then we would not be forced to judge the undesirable series of actions as irrational. We could then say that they were the unsurprising consequences of having only partial information about future options.[8] One mathematically grounded response to the above is to question time-bound interpretations and defenses of transitivity. Properly, we should think of transitivity as a constraint that applies to preferences only at a single point in time, and it is this simultaneous interpretation from which the axiom derives its appeal.[9] So we now have two interpretations, one dynamic and one simultaneous, both of which must be addressed.

Let us examine the perhaps more appealing simultaneous interpretation first. Useful as it is, the simultaneous interpretation is not entirely unproblematic. If it is taken to mean that the agent is confronted with three different feasible sets simultaneously, why do we not aggregate to create a single feasible set? Perhaps one should read Pab as the counterfactual proposition "if an agent could choose between a and b, she or he would choose a". In the context of the money-pump argument this becomes something like 'if you could choose between a and b, then you would swap b for a and give up a small positive sum, ϵ. Preferences (3) to (5) could then be written:

$$F = \{a, b\} \,\square\!\!\rightarrow \text{ swap } b \text{ for } a \text{ and pay } \epsilon_1 \tag{6}$$

$$F = \{b, c\} \,\square\!\!\rightarrow \text{ swap } c \text{ for } b \text{ and pay } \epsilon_2 \tag{7}$$

$$F = \{a, c\} \,\square\!\!\rightarrow \text{ swap } a \text{ for } c \text{ and pay } \epsilon_3 \tag{8}$$

where F denotes the feasible set; ϵ_1, ϵ_2, and ϵ_3 are all strictly positive; and $\square\!\!\rightarrow$ is the counterfactual connective.[10] The constellation of preferences specified by (6)–(8) indicates how a person would choose given a variety of possible opportunity sets and if the preferences are not transitive. To see how pernicious this is, we must follow the consequences of every possibility. Suppose F turns out to be {a, b}, then the agent will trade b for a and pay a finite sum. Alternatively, suppose the feasible set were {a, c}. In this case, the agent would give up a for c and pay a

[8] One could also argue that preferences change over time—and though I don't rely on this argument in any way here, we might note that it is, if anything, an argument that allows for the observation of intransitive choice even if preferences at any point in time were transitive.

[9] I am grateful to Professor Frank Hahn for suggesting this line of argument.

[10] This should be read: "if the antecedent condition as specified on the left-hand side of the connective were true, then the consequent indicated on the right-hand side would also be true".

different positive finite sum. Again there is a trade and a payment, but no grounds for criticism, and we would draw a similar conclusion if F were {b, c}. Insofar as we understand transitivity as a constraint on preferences defined over possible choice sets at a point in time, it seems that none of the actual outcomes has anything like the disastrous consequences that tellers of the money-pump story claim.

Even to get the money-pump story off the ground, it seems that we need to move towards a more explicitly sequential or dynamic understanding of the problem. This has been discussed in a literature to which we shall return in Section 6.6, though I present one argument here as it is a direct development of that above.

One might argue that if we presented an agent with the opportunity sets of (8), (7), and (6) in turn, then they would swap a for c, c for b, and b for a, giving up $\epsilon_3 + \epsilon_2 + \epsilon_1$. However tempting such a conclusion might be, it is not true that it follows logically as it would if the connective were the standard material implication. Despite the similarities between indicative conditionals and counterfactuals, there are differences to which work by Lewis (1986) and Stalnaker (1968) should make us alert. In propositional logic, if it is the case that $P \rightarrow A$, $Q \rightarrow B$, and $R \rightarrow C$, then it follows that $P \wedge Q \wedge R \mapsto A \wedge B \wedge C$. In this case, if one aggregates antecedents, then the sum of the consequents is a consequent of the aggregated antecedent. However, this does not apply to counterfactuals. To see this, consider the example of a person who goes on holiday and ponders how they might drown their sorrows if they lost some money.

$$\text{Lose travelers' cheques} \;\Box\!\!\rightarrow\; \text{have a beer} \tag{9}$$

$$\text{Lose cash} \;\Box\!\!\rightarrow\; \text{have a beer} \tag{10}$$

$$\text{Lose credit cards} \;\Box\!\!\rightarrow\; \text{have a gin and tonic} \tag{11}$$

A person could argue that if she or he lost cheques or cash, they might have a beer (9) and (10), and that losing credit cards called for having a gin and tonic (11). However, this does not mean that if the person were to lose all three forms of cash, they would logically be required to have two beers and a gin and tonic—for the simple reason that counterfactuals do not aggregate in the same way as material implication. And here is another version of this point. An experimenter asks a subject for preferences on three possible opportunity sets and on finding that they are intransitive observes that they expose the subject to the risk of money pumping. The subject could then perfectly reasonably ask on what grounds preferences relating to specific opportunity sets were being used to infer behavior over the concatenation of those sets. The subject adds that, if the experimenter had asked what choices would be made in the game now being considered, comprising a sequence of binary choices, their choices would be different. Given the intransitive choices that the person has, it is absurd to use them as a basis for the dynamic decision problem now under consideration. Much the same has been crisply put by LaValle and Fishburn (1988, p. 1225) thus: "Sensible people with cyclic preferences

would simply refuse to get involved in the money pump game unless they were deceived into believing it was to their advantage to do so." To some, these points are sufficient, though we shall come back to the literature's treatment of dynamic interpretations in Section 6.6 once we have extended our examination of reasons why rational agents may wish to exhibit intransitive behaviors.

6.5 Examples of Intransitive Preference and Issues of Description

A small number of arguments other than those above can be found in the literature. Attempted defenses of transitivity on the basis of decidability or ratifiability are inadequate for reasons similar to those discussed in the context of Tullock's logical argument. The decidability argument says that intransitive preferences yield no particular ranking of options. However, we saw that there is nothing in logic that requires binary preferences to be easily derivable from ternary (or other higher-order) preferences. Strictly, decidability requires only that preferences relevant to the decision problem are well-defined, so that if the choice is from three objects, only the ternary preference ranking need exist. Similarly with ratifiability. Suppose your preferences are Pab, Pbc, and Pca, and that you have to choose from the set {a, b, c}. However you decide, it seems there will always be something better you could choose, so no decision is ratifiable. But, anyone who claimed that they were satisfied with a particular choice could just point out that only the ternary preferences were relevant.

It is worth asking, more explicitly than hitherto, whether there are situations in which it seems positively desirable to violate transitivity, and I mention two examples. First imagine that you are at a friend's dinner party and that after dinner your host offers you a piece of fruit. We now consider three possible worlds, where the choices are as follows:

Possible world 1 (PW_1): offer = {small apple, orange}

Possible world 2 (PW_2): offer = {orange, large apple}

Possible world 3 (PW_3): offer = {small apple, large apple}.

If your preferences in PW_1 are, using the first letter to represent each object, Pos and in PW_2 Plo, how should you choose in PW_3? Transitivity excludes Psl, though many people would encourage their children to make such a choice out of politeness even if they thought the choices in PW_1 and PW_2 were open. This example is interesting in that it seems to get the force of intransitive preference right: it does not

propose objective, universal preference patterns, nor does it suggest that everyone is committed to making an intransitive choice; but it does argue that for those whose preferences are intransitive, it can be perfectly reasonable that they are so, and as we saw before, it illustrates that there need not be any dire bilking consequences of maintaining such a preference pattern.[11]

The second example arises in the context of a competitive game, and takes the form of a statistical paradox which is discussed by Blyth (1972) and was independently formulated by Packard (1982). Essentially, they show that choice over pairs of lotteries can be intransitive for agents who want to maximize the probability of winning. To see this, suppose that there are three six-sided die, α, β, and γ, with numbers on each face as indicated below.

Die α	1	1	4	4	4	4
Die β	3	3	3	3	3	3
Die γ	5	5	2	2	2	2

The game is played as follows. Two dice are chosen by a third neutral party. One player chooses a die, and the other player uses the die that remains. Each player throws their die, and the winner is the person with the highest number on the upturned face. The die are fair, so the probability of any particular face appearing is 1/6, and outcomes on each die are independent. If the die available are α and β, then there is a 2/3 chance of winning with α, so that is the die the rational agent will choose. By similar reasoning, it is easy to show that in this case the binary preferences will be P$\alpha\beta$, P$\beta\gamma$, and P$\gamma\alpha$.

Some people respond to such examples by saying that the primitive options are defined by the context, so that, for example, die α when the alternative is β is not the same option as die α when the alternative is γ, and we should redescribe the options to reflect this fact. The problem with this move, if one allows it, is that it undermines the content claims for the axiom. Indeed, one can go further and show that it is always possible to give any violation of transitivity a more refined primitive description in which the violation disappears (see proof in Anand 1990). To see this, suppose we have a set of intransitive choices: Cab, Cbc, and Cca. Noting that a when b is the alternative is to be distinguished from a when c is the alternative

[11] Larry Tempkin (1996) and Stuart Rachels (1998) have also developed an example-based approach in which they propose a pattern of choices over options that vary in the amount and duration of pain suffered. They argue that people would prefer to avoid longer durations of pain where the decrements are small but would ultimately prefer to suffer an almost negligible amount of pain for a long time, and show how this could lead to intransitive choices. I agree that it would be perfectly reasonable for people to have the intransitive choice pattern they claim, but I am less certain that absolutely everyone must or does have such choice patterns. However, one might be cautious about accepting Binmore and Voorhoeve's (2003) criticism of these examples, as it applies only to the latter's reconstruction, which is structurally different from the examples as originally developed. An earlier survey of counterexamples can be found in Bar-Hillel and Margalit (1988).

(and so on), we then rewrite these choices as:

C(a out of a and b) (b out of a and b)

C(b out of b and c) (c out of b and c)

C(c out of a and c) (a out of a and c).

We then apply the following new notation: l = a out of a and b, m = b out of a and b, n = b out of b and c, o = c out of b and c, p = c out of a and c, and q = a out of a and c. By substitution, we can now rewrite the intransitive choices as Clm, Cno, and Cpq, which eradicates the intransitive preference pattern. QED. Furthermore, it can be shown that the converse is also true: namely, that any transitive pattern of preference or choice can be given a different description in which transitivity is violated. So in a deep sense, one could argue that the issue of whether behavior is transitive or not ultimately depends on the linguistic conventions one chooses. So there is no reason why rational agents should not have intransitive preferences if, linguistically, that is most convenient. However, being able to fix the language *ex post* means that any behavior will turn out intransitive or transitive as one prefers, which in turn makes it impossible to argue that it has normative behavioral content (it would have no behavioral content). Content depends on fixing the language in advance, but any attempt to do so will be open to the difficulties already discussed.

6.6 Sophistication and Backwards Induction

Whilst I believe that most of the significant moves in this debate can be seen as variants of the arguments discussed so far, I want in this final section to return to money-pump arguments as they have been treated in decision trees, an approach that emerges from McClennen's (1990) analysis of dynamic choice, which was, in turn, motivated by his concerns with Hammond's (1988) account of consequentialism.

Of particular significance is the viewpoint developed by McClennen (1990) and Rabinowicz (2000) which holds that a *sophisticated* decision-maker, who uses backwards induction in a decision tree, will be saved from money pumping even if she or he has intransitive preferences. The basic idea is that sophistication enables the decision-maker to see what is coming, and in that case, there is no reason why an agent should not formulate a plan, in the manner conceived of by Strotz (1956), which takes account of the full choice scenario as they *know it could* unfold.

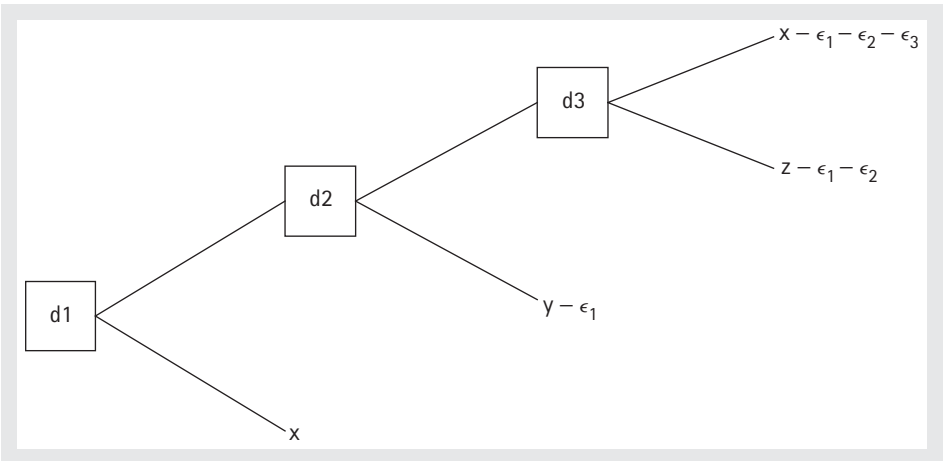

Fig. 6.1. Sophisticated choice and intransitive preference.

To see the basic structure of the argument, consider the decision tree in Figure 6.1. In this tree, small squares denote decision nodes as usual, and we further assume that the agent has the following intransitive preferences Pyx, Pzy, and Pxz. There are, therefore, positive finite amounts ϵ_1, ϵ_2 and ϵ_3, that would be consistent with a person choosing $y - \epsilon_1$ over x, $z - \epsilon_2 - \epsilon_1$ over $y - \epsilon_1$, and $x - \epsilon_3 - \epsilon_2 - \epsilon_1$ over $z - \epsilon_2 - \epsilon_1$. However, the backward induction argument, so the story goes, indicates that the sophisticated chooser would prefer $x - \epsilon_3 - \epsilon_2 - \epsilon_1$ at d3. But Pyx so at d2, which is now a choice between $y - \epsilon_1$ and $x - \epsilon_1 - \epsilon_2 - \epsilon_3$, the agent will choose $y - \epsilon_1$, and the decision problem terminates. This in turn means that d1 would become a choice between x and $y - \epsilon_1$, and so there is some finite positive ϵ_1 at which the person would choose $y - \epsilon_1$. The sophisticated chooser would use backwards induction to make one trade and stop, and there would be no money pumping.

Now Rabinowicz provides an intriguing development of this argument from which he concludes, counter to the earlier view that he shared with McClennen, that even this sophisticated, backwards-inducting agent could be turned into a money pump by a trader who was *persistent*.[12]

In this second tree (see Figure 6.2), we assume that the agent has the same intransitive preferences as before and is faced with a series of choices in which taking the upper branch denotes trading whilst taking the lower branch denotes declining to trade. In this case, the status quo at n1 is owning x, but then, if a person decides not to go to n2, which would guarantee some kind of trade, the persistent

[12] In case there is any doubt, the issue here is not the irrationality of being at a lower level of wealth but simply getting to a state that is dominated or dispreferred. The amounts ϵ_1, ϵ_2, etc. used in the literature are normally put in currency units, but this is only a rhetorical device—all that matters is that they stand for something that the agent values.

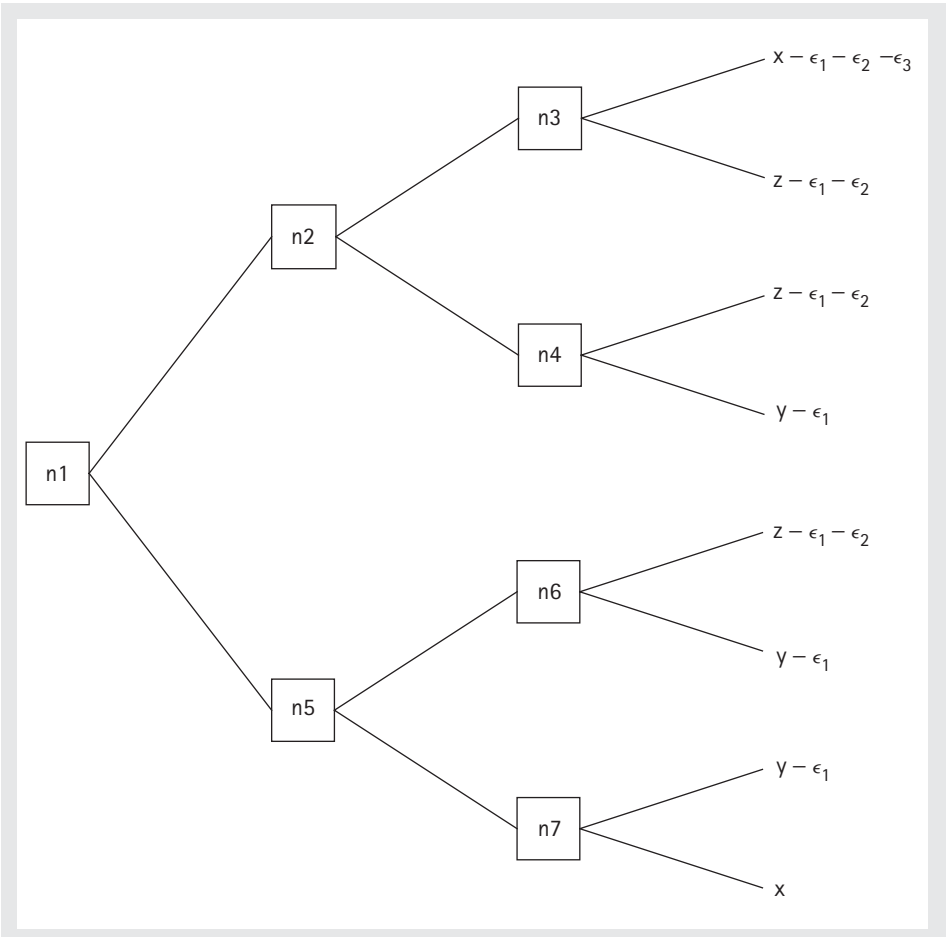

Fig. 6.2. Rabinowicz's persistent trader.

money-pumper offers another choice as specified by n5. Someone who still refused to trade would take the lower branch here and at n7 and end up with x. Rabinowicz shows that in this decision problem, the agent with intransitive pairwise preferences who is sophisticated is in fact vulnerable to money pumping and ends up with a dominated outcome, $x - \epsilon_1 - \epsilon_2 - \epsilon_3$.

This is a neat suggestion, but it raises a question of interpretation. Rabinowicz suggests that his decision problem demonstrates that sophisticated decision-makers with intransitive preferences can, after all, be susceptible to money pumping. However, this interpretation can be questioned on various grounds. Possibly the simplest challenge turns on recognizing that decision nodes can be collapsed where they are inessential, to form a simpler problem. One could represent the sequence of choices refusing to trade at n1, n5, and n7, for example, as a single choice of x. Similarly, the pathways n1, n5, and the lower branch of n6, and n1, n2, and the

lower branch of n4 could be represented as a second single branch leading to $y - \epsilon$. Repeating this process yields a simplified tree in which the agent can choose one of four options from which the rational agent can select any single option. The essence of the sophisticated chooser in Strotz's sense is that they are able to see the problem as a whole and plan accordingly, so it is compatible with that view that the agent assesses the decision problem in its reduced-form tree and then makes an "optimizing plan" for the full (structural) Rabinowicz tree that will deliver the preferred outcome by specifying the decision node route(s) that would lead to the preferred outcome.

A related way of seeing this can be derived from an important paper in which Cubitt and Sugden (2001) formalize sequential choice in a way that further clarifies the compatibility of intransitive choice and maximizing behavior. The new terminology they introduce makes it difficult to summarize their position succinctly; but basically, the authors recognize and demonstrate that money-pump arguments can be located within a formal framework that needs to be able to describe intransitive choices and consistency criteria independently. They highlight the fact that traditional money-pump arguments work by applying unexpected feasible sets, introduce the concept of an encounter as a sequence of trades (a more accurate way of describing choices that previous arguments awkwardly forced into decision trees), and use their formalization to prove a number of propositions which show that people who violate consistency criteria are not vulnerable to money pumping in encounters (when the choice setup is properly described). Transitivity is not one of their explicit criteria, but expansion and contraction consistency are, and, as I have argued above, these are much more closely related to transitivity than the literature sometimes suggests. In their framework, separability (independence of current choices from past decisions) is shown not to be necessary to protect an agent from money pumping. However, the backwards induction argument assumes just this—namely, that avoidance of money pumping necessitates separable preferences. The upshot is that the person who is sophisticated *only* to the point of using backwards induction does not have the defense that Cubitt and Sugden identify as being crucial for the avoidance of money pumping, namely a willingness to violate separability.[13]

To sum up briefly, it is evident that context, mainly in the form of counterfactual considerations about what might happen if one were to choose otherwise, plays a significant role in arguments for the view that rational agents can, coherently, have intransitive preferences. There are deeper issues of language and representation which can be explored too, but they do not appear to support the equation of rational choice with only transitive preference. Substantive and logical aspects of the modern view surveyed here complement Fishburn's formal theory of

[13] This is a view that Rabinowicz allows in his formulation of what he calls "wise" choice, so his ultimate position may in fact be rather similar to that proposed here.

context-dependent choice, but they are also linked to McClennen's analytical work on the rationality of independence violations (see Chapter 5 above) as well as research into state-dependent utility, which is just a special case of context dependence. Expected utility is not just a first-order approximation, we might conclude, but rather a useful exact model of context-free choice, though one that does not possess the conceptual or axiomatic resources to reflect explicitly a range of considerations that normative decision theory needs to model. Elsewhere, I have suggested that the only internal consistent preference axiom in formal rational choice theory that really was "hands off" would be a form of dominance which constrains behavior to match preferences. The doubts about the Dutch Book arguments for axioms concerning belief, to which Hájek draws our attention in Chapter 7, are of a different kind, it seems to me. I find it a little surprising that there are as many potential difficulties with Dutch Book arguments for probability axioms, and agree with Hájek that these do not seem to undermine the classical axioms of probability. However, I also accept that there are concepts of credence (like potential surprise, weight of evidence, and ambiguity) which might be given more prominence when thinking about how rational agents cope with uncertainty. No doubt the axioms of subjective expected utility theory will continue to be recognized as central in the history of economic theory, but their equation with rationality seems less compelling than perhaps it once did, and the arguments concerning are transitivity are illustrative.

References

Anand, P. (1987). Are the Preference Axioms Really Rational? *Theory and Decision*, 23, 189–214.

—— (1990). Interpreting Axiomatic (Decision) Theory. *Annals of Operations Research*, 23, 91–101.

—— (1993*a*). *The Foundations of Rational Choice under Risk*. Oxford: Oxford University Press.

—— (1993*b*). The Philosophy of Intransitive Preference. *Economic Journal*, 103, 337–46.

Arrow, K. J. (1979). Values and Collective Decision-Making. In F. Hahn and M. Hollis (eds.), *Philosophy and Economic Theory*. Oxford: Oxford University Press.

Bar-Hillel, M., and Margalit, A. (1988). How Vicious are Cycles of Intransitive Choice? *Theory and Decision*, 24, 119–45.

Bauman, P. (2005). Theory Choice and the Intransitivity of "Is a Better Theory Than". *Philosophy of Science*, 72, 231–40.

Bell, D. (1982). Regret in Decision-Making under Uncertainty. *Operations Research*, 20, 961–81.

Binmore, K., and Voorhoeve, A. (2003). Defending Transitivity against Zeno's Paradox. *Philosophy and Public Affairs*, 31, 272–9.

Blyth, C. (1972). Some Probability Paradoxes in Choice from among Random Alternatives. *Journal of the American Statistical Association*, 67, 367–73.

Broome, J. (1991). *Weighing Goods*. Oxford: Basil Blackwell.

Cubitt, R. P., and Sugden, R. (2001). On Money Pumps. *Games and Economic Behavior*, 37, 121–60.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

Fishburn, P. C. (1982). Non-Transitive Measurable Utility. *Journal of Mathematical Psychology*, 26, 31–67.

—— (1984). Dominance in SSB Utility Theory. *Journal of Economic Theory*, 34, 130–48.

—— (1988). *Non-Linear Preference and Utility Theory*. Baltimore: Johns Hopkins University Press.

Gale, D., and Mas-Collel, A. (1975). An Equilibrium Existence Theorem for a General Model without Ordered Preferences. *Journal of Mathematical Economics*, 2, 9–15.

Gendin, S. (1996). Why Preference is Not Transitive. *The Philosophical Quarterly*, 46, 482–8.

Guala, F. (2000). The Logic of Normative Falsification. *Journal of Economic Methodology*, 7, 59–93.

Hammond, P. (1988). Consequentialist Foundations for Expected Utility. *Theory and Decision*, 25, 25–78.

Hansson, S. O. (2001). *The Structure of Values and Norms*. Cambridge: Cambridge University Press.

Hughes, R. I. G. (1980). Rationality and Intransitive Preferences. *Analysis*, 40, 132–4.

Kim, T., and Richter, M. K. (1986). Non-transitive Non-Total Consumer Theory. *Journal of Economic Theory*, 38, 324–68.

Kirchsteiger, G., and Puppe, C. (1996). Intransitive Choices Based on Transitive Preferences: The Case of Menu Dependent Information. *Theory and Decision*, 41, 37–58.

LaValle, I., and Fishburn, P. C. (1988). Context Dependent Choice with Nonlinear and Nontransitive Preferences. *Econometrica*, 56, 1221–39.

Levi, I. (1986). *Hard Choices: Decision-Making under Unresolved Conflict*. Cambridge: Cambridge University Press.

Lewis, D. (1986). *Counterfactuals*. Oxford: Basil Blackwell.

Loomes, G., and Sugden, R. (1982). Regret Theory. *Economic Journal*, 92, 805–24.

Machina, M. (1989). Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty. *Journal of Economic Literature*, 27, 1622–68.

Mandler, M. (2005). Incomplete Preferences and Rational Intransitivity of Choice. *Games and Economic Behavior*, 50, 255–77.

May, K. O. (1954). Intransitivity, Utility and the Aggregation of Preference Patterns. *Econometrica*, 22, 1–13.

McClennen, E. F. (1988). Sure-Thing Doubts. In P. Gärdenfors and N. -E. Sahlin (eds.), *Decision, Probability and Utility*, 166–82. Cambridge: Cambridge University Press.

—— (1990). *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

Mongin, P. (2000). Does Optimisation Imply Rationality? *Synthese*, 124, 73–111.

Packard, D. J. (1982). Cyclical Preference Logic. *Theory and Decision*, 14, 415–26.

Putnam, H. (1995). On the Rationality of Preferences. Paper given to conference at Santa Clara University, 4 March, mimeo.

Rabinowicz, W. (2000). Money Pump with Foresight. In M. Almeida (ed.), *Imperceptible Harms and Benefits*, 123–43. Dordrecht: Kluwer.

Rachels, S. (1998). Counterexamples to the Transitivity of Better Than. *Australian Journal of Philosophy*, 76, 71–83.
Rambo, E. H. (1999). Symbolic Interests and Meaningful Purposes. *Rationality and Society*, 11, 317–42.
Schick, F. (1986). Dutch Books and Money Pumps. *Journal of Philosophy*, 83, 112–19.
Sen, A. K. (1970). *Collective Choice and Social Welfare*. San Francisco: Holden-Day.
—— (1985). Rationality and Uncertainty. *Theory and Decision*, 18, 109–27.
—— (1997). Maximisation and the Act of Choice. *Econometrica*, 65, 745–79.
Shafer, W. J. (1976). Equilibrium in Economics without Ordered Preferences or Free Disposal. *Journal of Mathematical Economics*, 3, 135–7.
Stalnaker, R. (1968). A Theory of Conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, 98–112. Oxford: Blackwell.
Strotz, R. H. (1956). Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies*, 23, 165–80.
Sugden, R. (1985). Why be Consistent? *Economica*, 52, 167–84.
—— (2003). The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences. *American Economic Review*, 94/4, 1014–33.
Tempkin, L. (1996). A Continuum Argument for Intransitivity. *Philosophy and Public Affairs*, 25, 175–210.
Tullock, G. (1964). The Irrationality of Intransitivity. *Oxford Economic Papers*, 16, 401–6.
von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.
von Wright, G. (1963). *The Logic of Preference*. Edinburgh: Edinburgh University Press.
Walsh, V. (1996). *Rationality Allocation and Reproduction*. Oxford: Oxford University Press.

CHAPTER 7

DUTCH BOOK ARGUMENTS

ALAN HÁJEK

7.1 Introduction

Beliefs come in varying degrees. I am more confident that this coin will land heads when tossed than I am that it will rain in Canberra tomorrow, and I am more confident still that $2 + 2 = 4$. It is natural to represent my *degrees of belief*, or *credences*, with numerical values. *Dutch Book arguments* purport to show that there are rational constraints on such values. They provide the most famous justification for the Bayesian thesis that degrees of belief should obey the probability calculus. They are also offered in support of various further principles that putatively govern rational subjective probabilities.

Dutch Book arguments assume that your credences match your betting prices: you assign probability p to X if and only if you regard pS as the value of a bet that pays S if X, and nothing otherwise (where S is a positive stake). Here we assume that your highest buying price equals your lowest selling price, with your being indifferent between buying and selling at that price; we will later relax this assumption. For example, my credence in heads is $\frac{1}{2}$, corresponding to my valuing

I thank especially Brad Armendt, Jens-Christian Bjerring, Darren Bradley, Rachael Briggs, Andy Egan, Branden Fitelson, Carrie Jenkins, Stephan Leuenberger, Isaac Levi, Aidan Lyon, Patrick Maher, John Matthewson, Peter Menzies, Ralph Miles, Daniel Nolan, Darrell Rowbottom, Wolfgang Schwarz, Teddy Seidenfeld, Michael Smithson, Katie Steele, Michael Titelbaum, Susan Vineberg, and Weng Hong Tang, whose helpful comments led to improvements in this article.

a $1 bet on heads at 50 cents. A *Dutch Book* is a set of bets bought or sold at such prices as to guarantee a net loss. An agent is susceptible to a Dutch Book, and her credences are said to be "incoherent" if there exists such a set of bets bought or sold at prices that she deems acceptable (by the lights of her credences).

There is little agreement on the origins of the term. Some say that Dutch merchants and actuaries in the seventeenth century had a reputation for being canny businessmen; but this provides a rather speculative etymology. By the time Keynes wrote in 1920, the proprietary sense of the term "book" was apparently familiar to his readership: "In fact underwriters themselves distinguish between risks which are properly insurable, either because their probability can be estimated between narrow numerical limits or because it is possible to make a 'book' which covers all possibilities" (1920, p. 21). Ramsey's ground-breaking paper "Truth and Probability" (written in 1926 but first published in 1931), which inaugurates the Dutch Book argument,[1] speaks of "a book being made against you" (1980, p. 44; 1990, p. 79). Lehman (1955, p. 251) writes: "If a bettor is quite foolish in his choice of the rates at which he will bet, an opponent can win money from him no matter what happens. . . . Such a losing book is called by [bookmakers] a 'dutch book'." So certainly "Dutch Books" appear in the literature under that name by 1955. Note that Dutch Book arguments typically take the "bookie" to be the clever person who is assured of winning money off some irrational agent who has posted vulnerable odds, whereas at the racetrack it is the "bookie" who posts the odds in the first place.

The closely related notion of "arbitrage", or a risk-free profit, has long been known to economists—for example, when there is a price differential between two or more markets (currency, bonds, stocks, etc.). An arbitrage opportunity is provided by an agent with intransitive preferences, someone who for some goods A, B, and C, prefers A to B, B to C, and C to A. This agent can apparently be turned into a "money pump" by being offered one of the goods and then sequentially offered chances to trade up to a preferred good for a comparatively small fee; after a cycle of such transactions, she will return to her original position, having lost the sum of the fees she has paid, and this pattern can be repeated indefinitely. Money-pump arguments, like Dutch Book arguments, are sometimes adduced to support the rational requirement of some property of preferences—in this case, transitivity. (See Anand, Chapter 6 above, for further discussion of money-pump arguments, and for skepticism about their probative force that resonates with some of our subsequent criticisms of Dutch Book arguments.)

This chapter will concentrate on the many forms of Dutch Book argument, as found especially in the philosophical literature, canvassing their interpretation, their cogency, and their prospects for unification.

[1] Earman (1992) finds some anticipation of the argument in the work of Bayes (1764).

7.2 Classic Dutch Book Arguments for Probabilism

7.2.1 Probabilism

Philosophers use the term "probabilism" for the traditional Bayesian thesis that agents have degrees of belief that are rationally required to conform to the laws of probability. (This is silent on other issues that divide Bayesians, such as how such degrees of belief should be updated.) These laws are taken to be codified by Kolmogorov's (1933) axiomatization, and the best-known Dutch Book arguments aim to support probabilism, so understood. However, several aspects of that axiomatization are presupposed, rather than shown, by Dutch Book arguments. Kolmogorov begins with a finite set Ω, and an algebra $\mathcal{F}$ of subsets of Ω (closed under complementation and finite union); alternatively, we may begin with a finite set S of sentences in some language, closed under negation and disjunction. We then define a real-valued, bounded (unconditional) probability function P on $\mathcal{F}$, or on S. Dutch Book arguments cannot establish any of these basic framework assumptions, but rather take them as given.

The heart of probabilism, and of the Dutch Book arguments, is the numerical axioms governing P (here presented sententially):

1. *Non-negativity*: $P(X) \geq 0$ for all X in S.
2. *Normalization*: $P(T) = 1$ for any tautology T in S.
3. *Finite additivity*: $P(X \vee Y) = P(X) + P(Y)$ for all X, Y in S such that X is incompatible with Y.

7.2.2 Classic Dutch Book Arguments for the Numerical Axioms

We now have a mathematical characterization of the probability calculus. Probabilism involves the normative claim that if your degrees of belief violate it, you are irrational. The Dutch Book argument begins with a mathematical theorem:

Dutch Book Theorem. If a set of betting prices violate the probability calculus, then there is a Dutch Book consisting of bets at those prices.

The argument for probabilism involves the normative claim that if you are susceptible to a Dutch Book, then you are irrational. The sense of "rationality" at issue here is an ideal, suitable for logically omniscient agents rather than for humans; "you" are understood to be such an agent.

The gist of the proof of the theorem is as follows (all bets are assumed to have a stake of \$1):

Non-negativity. Suppose that your betting price for some proposition N is negative—that is, you value a bet that pays \$1 if N, 0 otherwise at some negative amount $\$ -n$, where $n > 0$. Then you are prepared to sell a bet on N for $\$ -n$—that is, you are prepared to pay someone $\$n$ to take the bet (which must pay at least \$0). You are thus guaranteed to lose at least $\$n$.

Normalization. Suppose that your betting price $\$t$ for some tautology T is less than \$1. Then you are prepared to sell a bet on T for $\$t$. Since this bet must win, you face a guaranteed net loss of $\$(1 - t) > 0$. If $\$t$ is greater than \$1, you are prepared to buy a bet on T for $\$t$, guaranteeing a net loss of $\$(t - 1) > 0$.

Finite additivity. Suppose that your betting prices on some incompatible P and Q are $\$p$ and $\$q$ respectively, and that your betting price on $P \vee Q$ is $\$r$, where $\$r > \$(p + q)$. Then you are prepared to sell separate bets on P (for $\$p$) and on Q (for $\$q$), and to buy a bet on $P \vee Q$ for $\$r$, assuring an initial loss of $\$(r-(p + q)) > 0$. But however the bets turn out, there will be no subsequent change in your fortune, as is easily checked.

Now suppose that $\$r < \$(p + q)$. Reversing "sell" and "buy" in the previous paragraph, you are guaranteed a net loss of $\$((p + q) - r) > 0$.

So much for the Dutch Book theorem; now, a first pass at the argument:

P1. Your credences match your betting prices.
P2. Dutch Book theorem: if a set of betting prices violate the probability calculus, then there is a Dutch Book consisting of bets at those prices.
P3. If there is a Dutch Book consisting of bets at your betting prices, then you are susceptible to losses, come what may, at the hands of a bookie.
P4. If you are so susceptible, then you are irrational.

∴ C. If your credences violate the probability calculus, then you are irrational.
∴ C′. If your credences violate the probability calculus, then you are epistemically irrational.

The bookie is usually assumed to seek cunningly to win your money, to know your betting prices, but to know no more than you do about contingent matters. None of these assumptions is necessary. Even if he is a bumbling idiot or a kindly benefactor, and even if he knows nothing about your betting prices, he *could* sell/buy you bets that ensure your loss, perhaps by accident; you are still *susceptible* to such loss. And even if he knows everything about the outcomes of the relevant bets, he cannot thereby expose you to losses *come what may*; rather, he can fleece you in the actual circumstances that he knows to obtain, but not in various possible circumstances in which things turn out differently.

The irrationality that is brought out by the Dutch Book argument is meant to be one *internal* to your degrees of belief, and in principle detectable by you by a priori reasoning alone. Much of our discussion will concern the exact nature of such "irrationality". Offhand, it appears to be *practical* irrationality—your openness to financial exploitation. Let us start with this interpretation; in Section 7.4 we will consider other interpretations.

7.2.3 Converse Dutch Book Theorem

There is a gaping loophole in this argument as it stands. For all it says, it may be the case that *everyone* is susceptible to such sure losses, and that obeying the probability calculus provides no inoculation. In that case, we have seen no reason so far to obey that calculus. This loophole is closed by the equally important, but often neglected

Converse Dutch Book Theorem. If a set of betting prices obey the probability calculus, then there does not exist a Dutch Book consisting of bets at those prices.

This theorem was proved independently by Kemeny (1955) and Lehman (1955). Ramsey seems to have been well aware of it (although we have no record of his proving it): "Having degrees of belief obeying the laws of probability implies a further measure of consistency, namely such a consistency between the odds acceptable on different propositions as shall prevent a book being made against you" (1980, p. 41; 1990, p. 79). A proper presentation of the Dutch Book argument should include this theorem as a further premise.

A word of caution. As we will see, there are many Dutch Book arguments of the form:

If you violate Φ, then you are susceptible to a Dutch Book

$\therefore$ You should obey Φ.

None of these arguments has any force without a converse premise. (If you violate Φ, then you will eventually die. A sobering thought, to be sure, but hardly a reason to join the ranks of the equally mortal Φers!) Ideally, the converse premise will have the form:

If you *obey* Φ, then you are *not* susceptible to a Dutch Book.

But a weaker premise may suffice:

If you *obey* Φ, then *possibly* you are *not* susceptible to a Dutch Book.[2]

If *all* those who violate Φ are susceptible, and at least *some* who obey Φ are not, you apparently have an incentive to obey Φ. If you don't, we know you are susceptible; if you do, at least there is some hope that you are not.

[2] Thanks here to Daniel Nolan.

7.2.4 Extensions

Kolmogorov goes on to extend his set-theoretic underpinnings to infinite sets, closed further under *countable* union; we may similarly extend our set of sentences S so that it is also closed under infinitary disjunction. There is a Dutch Book argument for the corresponding infinitary generalization of the finite additivity axiom:

3′. *Countable additivity*: If $A_1, A_2, \ldots$ is a sequence of pairwise incompatible sentences in S, then

$$P\left(\bigvee_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

Adams (1962) proves a Dutch Book theorem for countable additivity; Skyrms (1984) and Williamson (1999) give simplified versions of the corresponding argument.

Kolmogorov then analyzes the *conditional probability* of A given B by the ratio formula:

$$\text{(Conditional Probability)} \quad P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (P(B) > 0).$$

This too has a Dutch Book justification. Following de Finetti (1937), we may introduce the notion of a *conditional bet* on A, given B, which

— pays \$1 if A & B
— pays 0 if ¬A & B
— is called off if ¬B (i.e. the price you pay for the bet is refunded).

Identifying an agent's value for $P(A|B)$ with the value she attaches to this conditional bet, if she violates (Conditional Probability), she is susceptible to a Dutch Book consisting of bets involving A & B, ¬B, and a conditional bet on A given B.

7.3 Objections

We will not question here the Dutch Book theorem or its converse. But there are numerous objections to premises P1, P3, and P4.

There are various circumstances in which an agent's credence in X can come apart from her betting price for X: when X is unverifiable or unfalsifiable; when betting on X has other collateral benefits or costs; when the agent sees a correlation between X and any aspect of a bet on X (its price, its stake, or even its placement); and so on. More generally, the betting interpretation shares a number of problems with operational definitions of theoretical terms, and in particular behaviorism about mental states (see Eriksson and Hájek 2007). The interpretation also assumes

that an agent values money linearly—implausible for someone who needs $1 to catch a bus home, and who is prepared to gamble at otherwise unreasonable odds for a chance of getting it. Since in cases like this it seems reasonable for prices of bets with monetary prizes to be non-additive, if we identify credences with those prices, non-additivity of credences in turn seems reasonable. On the other hand, if we weaken the connection between credences and betting prices posited by P1, then we cannot infer probabilism from any results about rational betting prices—the latter may be required to obey the probability calculus; but what about *credences*? We could instead appeal to bets with prizes of *utilities* rather than monetary amounts. But the usual way of defining utilities is via a "representation theorem", again dating back to Ramsey's "Truth and Probability". Its upshot is that an agent whose preferences obey certain constraints (transitivity and so on) is representable as an expected utility maximizer according to some utility and probability function. This threatens to render the Dutch Book argument otiose—the representation theorem has already provided an argument for probabilism. Perhaps some independent, probability-neutral account of "utility" can be given; but in any case, a proponent of any Dutch Book argument should modify P1 appropriately.

All these problems carry over immediately to de Finetti's Dutch Book argument for (Conditional Probability), and further ones apparently arise for his identification of *conditional* credences with *conditional* betting odds. Here is an example adapted from one given by Howson (1995) (who in turn was inspired by a well-known counterexample, attributed to Richmond Thomason, to the so-called Ramsey test for the acceptability of a conditional). You may assign low conditional probability to your ever knowing that you are being spied on by the CIA, given that in fact you are—they are clever about hiding such surveillance. But you presumably place a high value on the corresponding conditional bet—once you find out that the condition of the bet has been met, you will be very confident that you know it!

It may seem curious how the Dutch Book argument—still understood literally—moves from a mathematical theorem concerning the existence of abstract bets with certain properties to a normative conclusion about rational credences via a premise about some bookie. Presumably the agent had better assign positive credence to the bookie's existence, his nefarious motives, and his readiness to take either side of the relevant bets as required to ensnare the agent in a Dutch Book—otherwise, the bare possibility of such a scenario ought to play no role in her deliberations. (Compare: if you go to Venice, you face the possibility of a painful death in Venice; if you do not go to Venice, you do not face this possibility. That is hardly a reason for you to avoid Venice; your appropriate course of action has to be more sensitive to your credences and utilities.) But probabilism should not legislate on what credences the agent has about such contingent matters. Still less should probabilism require this kind of paranoia when it is in fact unjustified—when she rightly takes her neighborhood to be free of such mercenary characters,

as most of us do. And even if such characters abound, she can simply turn down all offers of bets when she sees them coming. So violating the probability calculus may not be a practical liability after all. Objections of this kind cast doubt on an overly literal interpretation of the Dutch Book argument. (See Kyburg 1978; Kennedy and Chihara 1979; Christensen 1991; Hájek 2005.)

But even granting the ill effects, practically speaking, of violating the probability calculus, it is a further step to show that there is some *epistemic* irrationality in such violation. Yet it is this conclusion (C′) that presumably the probabilist really seeks. After all, as Christensen (1991) argues, if those who violated probability theory were tortured by the Bayesian Thought Police, that might show that violating probability theory is irrational in some sense—but surely not in the sense that matters to the probabilist.

P3 presupposes a so-called *package principle*—the value that you attach to a collection of bets is the sum of the values that you attach to the bets individually. Various authors have objected to this principle (e.g. Schick 1986; Maher 1993). Let us look at two kinds of concern. First, there may be interference effects between the *prizes* of the bets. Valuing money nonlinearly is a clear instance. Suppose that the payoff of each of two bets is not sufficient for your bus ticket, so taken individually they are of little value to you; but their combined payoff *is* sufficient, so the package of the two of them is worth a lot to you. (Here we are still interpreting Dutch Book arguments as taking literally all this talk of bets and monetary gains and losses.) Secondly, you may regard the *placement* of one bet in a package as correlated with the *outcome* of another bet in the package. I may be confident that Labour will win the next election, and that my wife is in a good mood; but knowing that she hates my betting on politics, my placing a bet on Labour's winning changes my confidence in her being in a good mood. This interference effect could not show up in the bets taken individually. We cannot salvage the argument merely by restricting "Dutch Books" to cases in which such interference effects are absent, for that would render false the Dutch Book theorem (so understood): your sole violations of the probability calculus might be over propositions for which such effects are present. Nor should the probabilist rest content with weakening the argument's conclusion accordingly; after all, *any* violation of the probability calculus is supposed to be irrational, even if it occurs solely in such problematic cases. The dilemma, then, is to make plausible the package principle without compromising the rest of the argument. This should be kept in mind when assessing any Dutch Book argument that involves multiple bets, as most do.

The package principle is especially problematic when the package is *infinite*, as it needs to be in the Dutch Book argument for countable additivity. Arntzenius, Elga, and Hawthorne (2004) offer a number of cases of infinite sets of transactions, each of which is favorable, but which are unfavorable in combination. Suppose, for example, that Satan has cut an apple into infinitely many pieces, labeled by the natural numbers, and that Eve can take as many pieces as she likes. If she takes only

finitely many, she suffers no penalty; if she takes infinitely many, she is expelled from the Garden. Her first priority is to stay in the Garden; her second priority is to eat as many pieces as she can. For each n (= 1, 2, 3, . . .), she is strictly better off choosing to eat piece #n. But the combination of all such choices is strictly worse than the status quo. Arntzenius, Elga, and Hawthorne consider similar problems with the agglomeration of infinitely many bets, concluding: "There simply need not be any tension between judging each of an infinite package of bets as favourable, and judging the whole package as unfavourable. So one can be perfectly rational even if one is vulnerable to an infinite Dutch Book" (p. 279).

P4 is also suspect unless more is said about the "sure" losses involved. For there is a good sense in which you may be susceptible to sure losses without any irrationality on your part. For example, it may be rational of you, and even rationally *required* of you, to be less than certain of various necessary a posteriori truths—that Hesperus is Phosphorus, that water is H_2O, and so on—and yet bets on the falsehood of these propositions are (metaphysically) guaranteed to lose. Some sure losses are not at all irrational; in Section 7.4 we will look more closely at which are putatively the irrational ones.

Moreover, for all we have seen, those who obey the probability calculus, while protecting themselves from sure monetary losses, may be guilty of *worse* lapses in rationality. After all, there are worse financial choices than sure monetary losses—for example, even greater *expected* monetary losses. (You would do better to choose the sure loss of a penny over a 0.999 chance of losing a million dollars.) And there are other ways to be irrational besides exposing yourself to monetary losses.

7.4 Intepretations and Variations

7.4.1 A Game-Theoretic Interpretation

A game-theoretic interpretation of the Dutch Book argument can be given. It is based on de Finetti's proposal of a game-theoretic basis for subjective expected utility theory. A simplified presentation is given in Seidenfeld (2001), although it is still far more general than we will need here. Inspired by this presentation, I will simplify again, as follows. Imagine a two-person, zero-sum game, between players whom for mnemonic purposes we will call the Agent and the Dutchman. The Agent is required to play first, revealing a set of real-valued numbers assigned to a finite partition of states—think of this as her probability assignment. The Dutchman sees this assignment, and chooses a finite set of weights over the partition—think of these as the stakes of corresponding bets, with the sign of each stake indicating whether the agent buys or sells that bet. The Agent wins the maximal total amount

that she can, given this system of bets—think of the actual outcome being the most favorable it could be, by her lights. The Dutchman wins the negative of that amount—that is, whatever the Agent wins, the Dutchman loses, and vice versa. Since the Dutchman may choose all the weights to be 0, he can ensure that the value of the game to the Agent is bounded above by 0. The upshot is that the Agent will suffer a sure loss from a clever choice of weights by the Dutchman if and only if her probability assignments violate the probability calculus.

This interpretation of the Dutch Book argument takes rather literally the story of a two-player interaction between an agent and a bookie that is usually associated with it. However, in light of some of the objections we saw in the last section, there are reasons for looking for an interpretation of the Dutch Book argument that moves beyond considerations of strategic conflict and maximizing one's gains.

7.4.2 The "Dramatizing Inconsistency" Interpretation

Ramsey's original paper offers such an interpretation. Here is the seminal passage:

> These are the laws of probability, which we have proved to be necessarily true of any consistent set of degrees of belief. Any definite set of degrees of belief which broke them would be inconsistent in the sense that it violated the laws of preference between options, such as that preferability is a transitive asymmetrical relation, and that if α is preferable to β, β for certain cannot be preferable to α if p, β if not-p. If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better and would then stand to lose in any event.
>
> We find, therefore, that a precise account of the nature of partial belief reveals that the laws of probability are laws of consistency, an extension to partial beliefs of formal logic, the logic of consistency . . .
>
> Having any definite degree of belief implies a certain measure of consistency, namely willingness to bet on a given proposition at the same odds for any stake, the stakes being measured in terms of ultimate values. Having degrees of belief obeying the laws of probability implies a further measure of consistency, namely such a consistency between the odds acceptable on different propositions as shall prevent a book being made against you.
>
> (1980, pp. 41–2; 1990, pp. 78–9)

This interpretation has been forcefully defended by Skyrms in a number of works (1980, 1984, 1987*a*); e.g. "Ramsey and de Finetti have provided a way in which the fundamental laws of probability can be viewed as pragmatic consistency conditions: conditions for the consistent evaluation of betting arrangements no matter how described" (1980, p. 120). Similarly, Armendt (1993, p. 4) writes of someone who violates the laws of probability: "I say it *is* a flaw of rationality to give, at the same time, two different choice-guiding evaluations to the same thing. Call this *divided-mind* inconsistency."

Notice an interesting difference between the quote by Ramsey and those of Skyrms and Armendt. Ramsey is apparently also making the considerably more controversial point that a violation of the *laws of preference*—not merely the laws of probability—is tantamount to inconsistency. This is more plausible for some of his laws of preference (e.g. transitivity, which he highlights) than for others (e.g. the Archimedean axiom or continuity, which are imposed more for the mathematical convenience of insuring that utilities are real-valued).

This version of the argument begins with P1 and P2 as before. But Skyrms and Armendt insist that the considerations of sure losses at the hands of a bookie are merely a *dramatization* of the real defect inherent in an agent's violating probability theory: an underlying inconsistency in the agent's evaluations. So their version of the argument focuses on that inconsistency instead. We may summarize it as follows:

P1. Your credences match your betting prices.
P2. Dutch Book theorem: if a set of betting prices violate the probability calculus, then there is a Dutch Book consisting of bets at those prices.
P3′. If there is a Dutch Book consisting of bets at your betting prices, then you give inconsistent evaluations of the same state of affairs (depending on how it is presented).
P4′. If you give inconsistent evaluations of the same state of affairs, then you are irrational.

∴C. If your credences violate the probability calculus, then you are irrational.
∴C′. If your credences violate the probability calculus, then you are *epistemically* irrational.

This talk of credences being "irrational" is implicit in Skyrms's presentation—he focuses more on the notion of inconsistency *per se*—but it is explicit in Armendt's.

This version of the argument raises new objections. "Inconsistency" is not a straightforward notion, even in logic. For starters, it is controversial just what counts *as* logic in this context. It would be glib to say that classical logic is automatically assumed. Apparently it is *not* assumed when we formulate the countable additivity axiom sententially—the logic had better be infinitary. In that case, is ω-inconsistency, the kind that might arise in countable Dutch Books, inconsistency of the troubling kind? (Consider an infinite set of sentences that has as members "Fn" for every natural number n, but also "$\neg(\forall x)Fn$".) Once we countenance nonclassical logics, which should guide our judgments of inconsistency? Weatherson (2003) argues that the outcomes of the bets appealed to in Dutch Book arguments must be *verified*, and thus that the appropriate logic is *intuitionistic*. Note that nothing in the Dutch Book arguments resolves these questions; yet the notion of *sure* losses looks rather different, depending on what we take to be logically "sure".

However we resolve such questions, the "inconsistency" at issue here is apparently something different again: a property of conflicting *evaluations*, and it is thus

essentially preference-based. Offhand, giving "two different choice-guiding evaluations", as Armendt puts it, seems to be a matter of not giving *identical* evaluations, a problem regarding the *number* of evaluations—two, rather than one. Understood this way, the alleged defect prima facie seems to be one of *inconstancy*, rather than *inconsistency*. To be sure, being "consistent" in ordinary English sometimes means repeating a particular task without noticeable variation, as when we say that Tiger Woods is a consistent golfer, or when we complain that the chef at a particular restaurant is inconsistent. But this is trading on a pun on the word, and it need not have anything to do with logic. Notice that it is surely this *non*-logical sense of the word that Ramsey has in mind when he speaks of "a certain measure of consistency, namely willingness to bet on a given proposition at the same odds for any stake", for it is hard to see how *logic* could legislate on that.

But arguably, the kind of inconstancy evinced by Dutch Book susceptibility *is* a kind of inconsistency. Crucial is Armendt's further rider, that of giving "two different choice-guiding evaluations *to the same thing*". The issue becomes one of how we individuate the "things", the objects of preference. Skyrms writes that the incoherent agent "will consider two different sets of odds as fair for an option depending on how that option is described; the equivalence of the descriptions following from the underlying Boolean logic" (1987*b*, p. 2). But even two logically equivalent sentences are not the same thing—they are two, rather than one. To be sure, they may correspond to a single profile of payoffs across all logically possible worlds (keeping in mind our previous concerns about rival logics). But is a failure to recognize this the sin of *inconsistency*, a sin of *commission*, or is it rather a failure of logical omniscience, a sin of *omission*? (In the end, it might not matter much either way if both are failures to meet the demands of epistemic rationality, at least in an ideal sense.) See Vineberg (2001) for skepticism of the viability of the "inconsistency" interpretation of the Dutch Book argument for the normalization axiom. This remains an area of lively debate.

That interpretation for the additivity axiom is controversial in a different way. Again, it may be irrational to give two different choice-guiding evaluations to the same thing. But those who reject the package principle deny that they are guilty of this kind of double-think. They insist that being willing to take bets individually does not rationally require being willing to take them in combination; recall the possibility of interference effects between the bets taken in combination. The interpretation is strained further for the *countable* additivity axiom; recall the problems that arose with the agglomeration of infinitely many transactions. In Section 7.6 we will canvass other Dutch Book arguments for which the interpretation seems quite implausible (not that Ramsey, Skyrms, or Armendt ever offered it for them).

Christensen (1996) is dubious of the inference from C to C′: while Dutch Books, so understood, may reveal an irrationality in one's *preferences*, that falls short of revealing some *epistemic* irrationality. Indeed, we may imagine an agent in whom the connection between preferences and epistemic states is sundered altogether.

(Cf. Eriksson and Hájek 2007.) As Christensen asks rhetorically, "How plausible is it, after all, that the intellectual defect exemplified by an agent's being more confident in P than in $(P \vee Q)$ is, at bottom, a defect in the agent's *preferences*?" (1996, p. 453).

7.4.3 "Depragmatized" Dutch Book Arguments

Such considerations lead Christensen to offer an alternative interpretation of Dutch Books (1996, 2001). First, he insists that the relationship of credences to preferences is normative: degrees of belief *sanction as fair* certain corresponding bets. Secondly, he restricts attention to what he calls "simple agents", ones who value only money, and do so linearly. He argues that if a simple agent's beliefs sanction as fair each of a set of betting odds, and that set allows construction of a set of bets whose payoffs are logically guaranteed to leave him monetarily worse off, then the agent's beliefs are rationally defective. He then generalizes this lesson to all rational agents.

Vineberg (1997) criticizes the notion of "sanctioning as fair" as vague and argues that various ways of precisifying it render the argument preference-based after all. Howson and Urbach (1993) present a somewhat similar argument to Christensen's—although without its notion of "simple agents"—cast in terms of a Dutch Bookable agent's inconsistent beliefs about subjectively fair odds. Vineberg levels similar criticisms against their argument. See also Maher (1997) for further objections to Christensen's argument, Christensen's (2004) revised version of it, and Maher's (2006) critique of that version.

7.5 Diachronic Dutch Book Arguments

The Dutch Book arguments that we have discussed are *synchronic*—all the bets are placed at the same time. *Diachronic* Dutch Book, or *Dutch strategy,* arguments are an important class in which the bets are spread across at least two times.

7.5.1 Conditionalization

Suppose that initially you have credences given by a probability function $P_{initial}$, and that you become certain of E (where E is the strongest such proposition). What should be your new probability function P_{new}? The favored updating rule among Bayesians is conditionalization; P_{new} is related to $P_{initial}$ as follows:

(Conditionalization) $\quad P_{new}(X) = P_{initial}(X|E) \quad (\text{provided} P_{initial}(E) > 0).$

The Dutch Book argument for conditionalization begins by assuming that you are committed to following a *policy* for updating—a function that takes as inputs your initial credence function, and the member of some partition of possible evidence propositions that you learn, and that outputs a new probability function. It is further assumed that this rule is known by the bookie (although even if it isn't, the bookie could presumably place the necessary bets in any case, perhaps by luck). The diachronic Dutch Book *theorem*, due to Lewis (1999), states that if your updating rule is anything other than conditionalization, you are susceptible to a diachronic Dutch Book. (Your updating policy is codified in the conditional bets that you take.) The *argument* continues that such susceptibility is irrational; thus, rationality requires you to update by conditionalizing. As usual, a converse theorem is needed to complete the argument; Skyrms (1987*b*) provides it.

7.5.2 Objections

Many of the objections to synchronic Dutch Book arguments reappear, some with extra force; and some new objections arise. Indeed, the conclusion of this argument is not as widely endorsed as is the conclusion of the classic synchronic Dutch Book argument (namely, probabilism). There are thus authors who think that the diachronic argument, unlike the synchronic argument, proves too much, citing various cases in which one is putatively *not* required to conditionalize. Arntzenius (2003), Bacchus, Kyburg, and Thalos (1990), and Bradley (2005) offer some.

Christensen (1996) argues that much as degrees of belief should be distinguished from corresponding betting prices (as we saw in Section 7.3), having a particular updating rule must be distinguished from corresponding conditional betting prices. The objection that "the agent will see the Dutch Book coming" has also been pursued with renewed vigor in the diachronic setting. Developing an argument by Levi (1987), Maher (1992) offers an analysis of the game tree that unfolds between the bettor and the bookie. Skyrms (1993) gives a rebuttal, showing how the bookie can ensure that the bettor loses nevertheless. Maher (1993, sect. 5.1.3) replies by distinguishing between accepting a sure loss and choosing a dominated act, and he argues that only the latter is irrational.

The package principle faces further pressure. Since there must be a time lag between a pair of the diachronic Dutch Book bets, the later one is placed in the context of a changed world and must be evaluated in that context. It is clearly permissible to revise your betting prices when you know that the world has changed since you initially posted those prices. The subsequent debate centers on just how much is built into the *commitments* you incur in virtue of having the belief revision policy that you do.

Then there are objections that have no analogue in the synchronic setting. Unlike the synchronic arguments, the diachronic argument for conditionalization makes a

specific assumption about how the agent interacts with the world, and that learning takes place by acquiring new *certainties*. But need evidence be so authoritative? Jeffrey (1965) generalizes conditionalizing to allow for less decisive learning experiences in which your probabilities across a partition $\{E_1, E_2, \ldots\}$ change to $\{P_{new}(E_1), P_{new}(E_2), \ldots\}$, where none of these values need be 0 or 1.

(Jeffrey conditionalization) $\quad P_{new}(X) = \sum_i P_{initial}(X|E_i)P_{new}(E_i).$

Jeffrey conditionalization is again supported by a Dutch Book and converse Dutch Book theorem (although some further assumptions are involved; see Armendt 1980; Skyrms 1987*b*). Lewis insists that the ideally rational agent's learning episodes *do* come in the form of new certainties; he regards Jeffrey conditionalization as a fallback rule for less-than-ideal agents. Rationality for Lewis thus involves more than just appropriately responding to evidence in the formation of one's beliefs; more tendentiously, it also involves the nature of that evidence itself. And it requires a commitment to some rule for belief revision. Van Fraassen (1989) disputes this. There is even controversy over what it is to follow a rule in the first place (Kripke 1982), which had no analogue in the synchronic argument. Note, however, that an agent who fails to conditionalize is surely susceptible to a Dutch Book *whether or not she follows some rival rule.* A bookie could diachronically Dutch Book her by accident, rather than by strategically exploiting her use of such a rule—even if the bookie merely stumbles upon the appropriate bets, they do still guarantee her loss.

How does the interpretation that Dutch Books dramatize evaluational inconsistencies fare in the diachronic setting? Christensen (1991) contends that there need be no irrationality in an agent's evaluations at different times being inconsistent with each other, much as there is no irrationality in a husband and wife having evaluations inconsistent with each other (thereby exposing them jointly to a Dutch Book). He offers a synchronic Dutch Book argument for conditionalization, appealing again to the idea that credences *sanction as fair* the relevant betting prices. See Vineberg (1997) for criticisms.

Van Fraassen (1984) gives a diachronic Dutch Book argument for the *Reflection Principle,* the constraint that an ideally rational agent's credences mesh with her expected future credences according to:

$$P_t(X|P_{t'}(X) = x) = x, \text{ for all } X \text{ and for all } x \text{ such that } P_t(P_{t'}(X) = x) > 0,$$

where P_t is the agent's probability function at time t, and $P_{t'}$ is her function at later time t'. Various authors (e.g. Christensen 1991; Howson and Urbach 1993) find conditionalization plausible, but the Reflection Principle implausible; and various authors find all the more that the argument for the Reflection Principle proves too much.

Suppose that you violate one of the axioms of probability—say, additivity. Then by the Dutch Book theorem, you are Dutch Bookable. Suppose, further, that you obey conditionalization. Then by the converse Dutch Book theorem for conditionalization, you are not Dutch Bookable. So you both are and are not Dutch Bookable—contradiction? Something has gone wrong. Presumably, these theorems need to have certain *ceteris paribus* clauses built in, although it is not obvious how they should be spelled out exactly.

More generally, the problem is that there are Dutch Book arguments for various norms—we have considered the norms of obeying the probability calculus, the Reflection Principle, updating by conditionalization, and updating by Jeffrey conditionalization. For a given norm *N*, the argument requires both a Dutch Book theorem:

> if you violate N, then you are susceptible to a Dutch Book

and a converse Dutch Book theorem:

> if you obey N, then you are immune to a Dutch Book.

But the latter theorem must have a *ceteris paribus* clause to the effect that you obey all the other norms. For if you violate, say, norm N', then by *its* Dutch Book theorem you are susceptible to a Dutch Book. So the converse Dutch Book theorem for *N* as it stands must be false: if you obey *N and violate* N' then you are susceptible to a Dutch Book after all. One might wonder how a theorem could ever render precise the required *ceteris paribus* clause in all its detail.

This problem only becomes more acute when we pile on still more Dutch Book arguments for still more norms. As we now will.

7.6 Some More Exotic Dutch Book Arguments, and Recent Developments

We have discussed several of the most important Dutch Book arguments, but they are just the tip of the iceberg. In this section we will survey briefly a series of such arguments for more specific or esoteric theses.

7.6.1 Semi-Dutch Book Argument for Strict Coherence

The first, due to Shimony (1955), is not strictly speaking a Dutch Book argument, but it is related closely enough to merit attention here. Call a *semi-Dutch Book* a set of bets that can at best break even, and that in at least one possible outcome has a net

loss. Call an agent *strictly coherent* if she obeys the probability calculus, and assigns $P(H|E) = 1$ only if E entails H. (These pieces of terminology are not Shimony's, but they have become standard more recently.) Simplifying his presentation, Shimony essentially shows that if you violate strict coherence, you are susceptible to a semi-Dutch Book. Such susceptibility, moreover, is thought to be irrational, since you risk a loss with no compensating prospect of a gain. Where Dutch Books militate against *strictly* dominated actions (betting according to Dutch Bookable credences), semi-Dutch Books militate against *weakly* dominated actions.

Semi-Dutch Book arguments raise new problems. Strict coherence cannot be straightforwardly added to the package of constraints supported by the previous Dutch Book arguments, since it is incompatible with updating by conditionalization. After all, an agent who conditionalizes on E becomes certain of E (given any possible condition), despite its not being a tautology. Earman (1992) takes this to reveal a serious internal problem with Bayesianism: a tension between its fondness for Dutch Book arguments, on the one hand, and conditionalization, on the other. But there is *no* sense, not even analogical, in which semi-Dutch Books dramatize inconsistencies. An agent who violates strict coherence can grant that the outcomes in which she would face a loss are *logically* possible, but she can *consistently* retort that this does not trouble her—after all, she is 100 percent confident that they will not obtain! Indeed, an omniscient God would be semi-Dutch Bookable, and none the worse for it.

7.6.2 Imprecise Probabilities

Few of our actual probability assignments are precise to infinitely many decimal places; and arguably, even ideally rational agents can have *imprecise* probability assignments. Such agents are sometimes modeled with *sets* of precise probability functions (Levi 1974; Jeffrey 1992), or with lower and upper probability functions (Walley 1991). There are natural extensions of the betting interpretation to accommodate imprecise probabilities. For example, we may say that your probability for X lies in the interval $[p, q]$ if and only if $\$p$ is the highest price at which you will buy, and $\$q$ is the lowest price at which you sell, a bet that pays \$1 if X, 0 otherwise. (Note that on this interpretation, maximal imprecision over the entire [0, 1] interval regarding everything would immunize you from all Dutch Books—you would never buy a bet with a stake of \$1 for more than \$0, and never sell it for less than \$1, so nobody could ever profit from your betting prices.) C. A. B. Smith (1961) shows that an agent can make lower and upper probability assignments that avoid sure loss but that nevertheless violate probability theory. Thus, the distinctive connection between probability incoherence and Dutch Bookability is cleaved for imprecise probabilities; probabilistic coherence is demoted to a sufficient but not necessary condition for the avoidance of sure loss. Walley

(1991) provides Dutch Book arguments for various constraints on upper and lower probabilities.

7.6.3 "Incompatibilism" about Chance and Determinism

Call the thesis that determinism is compatible with intermediate objective chances *compatibilism*, and call someone who holds this thesis a *compatibilist.* Schaffer (2007) argues that a compatibilist who knows that some event E is determined to occur, and yet who regards the chance of E at some time to be less than 1, is susceptible to a Dutch Book.

7.6.4 Popper's Axioms on Conditional Probability Functions

Unlike Kolmogorov, who axiomatized unconditional probability and then defined conditional probability thereafter, Popper (1959) axiomatized *conditional* probability directly. Stalnaker (1970) gives what can be understood as a Dutch Book argument for this axiomatization.

7.6.5 More Infinite Books

Suppose that your probability function is not concentrated at finitely many points—this implies that the range of that function is infinite (assuming an infinite state space). It is surely rational for you to have such a probability function; indeed, given the evidence at our disposal, it would surely be *irrational* for us to think that we can rule out, with probability 1, all but finitely many possible ways the world might be. Suppose, further, that your utility function is unbounded (although your utility for each possible outcome is finite). This too seems to be rationally permissible. McGee (1999) shows that you are susceptible to an infinite Dutch Book (involving a sequence of unconditional and conditional bets). He concludes: "in situations in which there can be infinitely many bets over an unbounded utility scale, no rational plan of action is available" (p. 257). McGee's argument is different from other Dutch Book arguments in two striking ways. First, it makes a rather strong and even controversial assumption about the agent's *utility* function. Secondly, McGee does not argue for some rationality constraint on a credence function; on the contrary, since the relevant constraint in this case (being concentrated on finitely many points) is implausible, he drives the argument in the opposite direction. The upshot is supposed to be that irrationality is unavoidable. One might argue, on the other hand, that this just shows that Dutch Bookability is not always a sign of irrationality.

The theme of seemingly being punished for one's rationality in situations involving infinitely many choices is pursued further in Barrett and Arntzenius (1999).

They imagine a rational agent repeatedly paying $1 in order to make a more profitable transaction; but after infinitely many such transactions, he has made no total profit on those transactions and has paid an infinite amount. He is better off at every stage acting in an apparently irrational way. For more on this theme, see Arntzenius, Elga, and Hawthorne (2004).

7.6.6 Group Dutch Books

If Jack assigns probability 0.3 to rain tomorrow and Jill assigns 0.4, then you can Dutch Book the pair of them: you buy a dollar bet on rain tomorrow from Jack for 30 cents and sell one to Jill for 40 cents, pocketing 10 cents. Hacking (1975) reports that the idea of guaranteeing a profit by judicious transactions with two agents with different betting odds can be found around the end of the ninth century AD, in the writings of the Indian mathematician Mahaviracarya. We have already mentioned Christensen's observation of the same point involving a husband and wife. And there are interesting Dutch Books involving a greater number of agents (in e.g. Bovens and Rabinowicz, forthcoming).

7.6.7 The Sleeping Beauty Problem

Most Dutch Book arguments are intended to support some general constraint on rational agents—*structural* features of their credence (or utility) profiles. We will end with an example of a Dutch Book argument for a very specific constraint: in a particular scenario, a rational agent is putatively required to assign a *particular* credence. The scenario is that of the Sleeping Beauty problem (Elga 2000). Someone is put to sleep, and then woken up either once or twice depending on the outcome of a fair coin toss (heads: once; tails: twice). But if she is to be woken up twice, her memory of the first awakening is erased. What probability should she give to heads at the first awakening? There are numerous arguments for answering $\frac{1}{2}$, and for answering $\frac{1}{3}$. Hitchcock (2004) gives a Dutch Book argument for the $\frac{1}{3}$ answer. Bradley and Leitgeb (2006) dispute this argument, offering further constraints on what a "Dutch Book" requires in order to reveal any irrationality in an agent.

7.7 Conclusion

We have seen a striking diversity of Dutch Book arguments. A challenge that remains is to give a *unified* account of them. Is there a single kind of fault that they all illustrate, or is there rather a diversity of faults as well? And if there is a single fault,

is it epistemic, or some other kind of fault? The interpretation according to which Dutch Books reveal an inconsistency in an agent's evaluations, for example, is more plausible for some of the Dutch Books than for others—it is surely implausible for McGee's Dutch Book and for some of the other infinitary books that we have seen. But in those cases, do we really want to say that the irrationality at issue literally concerns monetary losses at the hands of cunning bookies (which in any case is hardly an epistemic fault)?

Or perhaps irrationality comes in many varieties, and it is enough that a Dutch Book exposes it in *some* form or other. But if there are many different ways to be irrational, the *validity* of a Dutch Book argument for any particular principle is threatened. At best, it establishes that an agent who violates that principle is irrational *in one respect*. This falls far short of establishing that the agent is irrational all-things-considered; indeed, it leaves open the possibility that along all the other axes of rationality the agent is doing as well as possible, and even that overall there is nothing better that she could do. Moreover, it is worth emphasizing again that without a corresponding *converse* theorem that one can avoid a Dutch Book by obeying the principle, even the irrationality in that one respect has not been established—unless it is coherent that necessarily *all* agents are irrational in that respect. Dutch Books may reveal a pragmatic vulnerability of some kind, but it is a further step to claim that the vulnerability stems from irrationality.[3] Indeed, as some of the infinitary Dutch Books seem to teach us, some Dutch Books apparently do not evince any irrationality whatsoever. Sometimes your circumstances can be unforgiving through no fault of your own: you are damned whatever you do.

References

Adams, Ernest (1962). On Rational Betting Systems. *Archiv für Mathematische Logik und Grundlagenforschung*, 6, 7–29, 112–28.

Armendt, Brad (1980). Is There a Dutch Book Argument for Probability Kinematics?. *Philosophy of Science*, 474, 583–8.

—— (1993). Dutch Books, Additivity and Utility Theory. *Philosophical Topics*, 211, 1–20.

Arntzenius, Frank (2003). Some Problems for Conditionalization and Reflection. *Journal of Philosophy*, 100/7, 356–71.

—— Elga, Adam, and Hawthorne, John (2004). Bayesianism, Infinite Decisions, and Binding. *Mind*, 113, 251–83.

Bacchus, F., Kyburg, H. E., and Thalos, M. (1990). Against Conditionalization. *Synthese*, 85, 475–506.

Barrett, Jeff, and Arntzenius, Frank (1999). An Infinite Decision Puzzle. *Theory and Decision*, 46, 101–3.

[3] I thank an anonymous reviewer for putting the point this way.

Bayes, Thomas (1764). An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.

Bovens, Luc, and Rabinowicz, Wlodek (forthcoming). Dutch Books, Group Decision-Making, the Tragedy of the Commons and Strategic Jury Voting. *Synthese*.

Bradley, Darren, and Leitgeb, Hannes (2006). When Betting Odds and Credences Come Apart: More Worries for Dutch Book Arguments. *Analysis*, 66/2, 119–27.

Bradley, Richard (2005). Radical Probabilism and Bayesian Conditioning. *Philosophy of Science*, 72/2, 334–64.

Christensen, David (1991). Clever Bookies and Coherent Beliefs. *The Philosophical Review* C, no. 2, 229–47.

—— (1996). Dutch-Book Arguments De-pragmatized: Epistemic Consistency for Partial Believers. *Journal of Philosophy*, 93, 450–79.

—— (2001). Preference-Based Arguments for Probabilism. *Philosophy of Science*, 68/3, 356–76.

—— (2004). *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford: Oxford University Press.

de Finetti, B. (1937). La Prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7, 1–68; trans. as "Foresight: Its Logical Laws, its Subjective Sources", in H. E. Kyburg, Jr., and H. E. Smokler (eds.), *Studies in Subjective Probability*, 93–159. New York: Wiley, 1980.

Earman, John (1992). *Bayes or Bust?* Cambridge, MA: MIT Press.

Elga, Adam (2000). Self-Locating Belief and the Sleeping Beauty Problem. *Analysis*, 60/2, 143–7.

Eriksson, Lina, and Hájek, Alan (2007). What Are Degrees of Belief?. *Studia Logica*, 86, Special Issue on Formal Epistemology I, ed. Branden Fitelson, 183–213.

Hacking, Ian, (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.

Hájek, Alan (2005). Scotching Dutch Books?. *Philosophical Perspectives*, 19 (issue on Epistemology, ed. John Hawthorne), 139–51.

Hitchcock, Christopher (2004). Beauty and the Bets. *Synthese*, 139/3, 405–20.

Howson, Colin (1995). Theories of Probability. *British Journal for the Philosophy of Science*, 46, 1–32.

—— and Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Jeffrey, Richard C., (1965). *The Logic of Decision*. Chicago: University of Chicago Press.

—— (1992). Bayesianism with a Human Face. In *Probability and the Art of Judgment*, 77–107. Cambridge: Cambridge University Press.

Kemeny, J. (1995). Fair Bets and Inductive Probabilities. *Journal of Symbolic Logic*, 20, 263–73.

Kennedy, Ralph, and Chihara, Charles (1979). The Dutch Book Argument: Its Logical Flaws, its Subjective Sources. *Philosophical Studies*, 36, 19–33.

Keynes, John Maynard (1920). *A Treatise on Probability*. London: Macmillan and Co., Limited.

Kolmogorov, Andrei. N. (1993). *Grundbegriffe der Wahrscheinlichkeitsrechnung, Ergebnisse der Mathematik*. Berlin: Springer. Trans. as *Foundations of Probability*. New York: Chelsea Publishing Company, 1956.

Kripke, Saul (1982). *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Cambridge, MA: Harvard University Press.

Kyburg, Henry (1978). Subjective Probability: Criticisms, Reflections and Problems. *Journal of Philosophical Logic*, 7, 157–80.

Lehman, R. Sherman (1955). On Confirmation and Rational Betting. *Journal of Symbolic Logic*, 20/3, 251–62.

Levi, Isaac (1974). On Indeterminate Probabilities. *Journal of Philosophy*, 71, 391–418.

—— (1987). The Demons of Decision. *The Monist*, 70, 193–211.

Lewis, David (1999). *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Maher, Patrick (1992). Diachronic Rationality. *Philosophy of Science*, 59, 120–41.

—— (1993). *Betting on Theories*. Cambridge: Cambridge University Press.

—— (1997). Depragmatized Dutch Book Arguments. *Philosophy of Science*, 64, 291–305.

—— (2006). Review of David Christensen's *Putting Logic in its Place*. *Notre Dame Journal of Formal Logic*, 47, 133–49.

McGee, Vann (1999). An Airtight Dutch Book. *Analysis*, 59/4, 257–65.

Popper, Karl (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Ramsey, F. P. (1931). Truth and Probability. In *Foundations of Mathematics and Other Essays*, ed. R. B. Braithwaite, 156–98. London: Routledge & Kegan Paul. Repr. in H. E. Kyburg, Jr., and H. E. Smokler (eds.), *Studies in Subjective Probability*, 2nd edn., 23–52. New York: Wiley, 1980. Also repr. in *F. P. Ramsey: Philosophical Papers*, ed. D. H. Mellor, 52–94. Cambridge: Cambridge University Press, 1990.

Schaffer, Jonathan (2007). Deterministic Chance?. *British Journal of the Philosophy of Science*, 58, 113–40.

Schick, Frederic (1986). Dutch Bookies and Money Pumps. *Journal of Philosophy*, 83, 112–19.

Seidenfeld, T. (2001). Game Theory and its Relation to Bayesian Theory. In N. J. Smelser and P. B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, 5868–73. Oxford: Elsevier.

Shimony, A. (1955). Coherence and the Axioms of Confirmation. *Journal of Symbolic Logic*, 20, 1–28.

Skyrms, Brian (1980). Higher Order Degrees of Belief. In D. H. Mellor (ed.), *Prospects for Pragmatism*, 109–37. Cambridge: Cambridge University Press.

—— (1984). *Pragmatics and Empiricism*. New Haven: Yale University Press.

—— (1987*a*). Coherence. In N. Rescher (ed.), *Scientific Inquiry in Philosophical Perspective*, 225–42. Pittsburgh: University of Pittsburgh Press.

—— (1987*b*). Dynamic Coherence and Probability Kinematics. *Philosophy of Science*, 54/1, 1–20.

—— (1993). A Mistake in Dynamic Coherence Arguments?. *Philosophy of Science*, 60, 320–8.

Smith, Cedric A. B. (1961). Consistency in Statistical Inference and Decision. *Journal of the Royal Statistical Society* B, 23, 1–25.

Stalnaker, Robert (1970). Probability and Conditionals. *Philosophy of Science*, 37, 64–80.

van Fraassen, Bas (1984). Belief and the Will. *Journal of Philosophy*, 81, 235–56.

—— (1989). *Laws and Symmetry*. Oxford: Clarendon Press.

Vineberg, Susan (1997). Dutch Books, Dutch Strategies and What they Show about Rationality. *Philosophical Studies*, 86/2, 185–201.

—— (2001). The Notion of Consistency for Partial Belief. *Philosophical Studies*, 102, 281–96.

Walley, Peter (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Weatherson, Brian (2003). Classical and Intuitionistic Probability. *Notre Dame Journal of Formal Logic*, 44, 111–23.

Williamson, Jon (1999). Countable Additivity and Subjective Probability. *British Journal for the Philosophy of Science*, 50/3, 401–16.

CHAPTER 8

EXPERIMENTAL TESTS OF RATIONALITY

DANIEL READ

8.1 Introduction

Rational choice theory is an instrumental theory. It assumes that agents have a set of basic preferences and values which they undertake to satisfy, and it then specifies the optimal way to achieve those values. Rational choice theory is sometimes proposed as a purely normative theory, with no bearing on the descriptive question of what people actually do. It is also widely used as a working hypothesis about real human behavior, with many taking the view that behavior, when properly understood, follows the tenets of the theory, and an equally large number taking the opposite view. The majority of experimental studies in the fields of behavioral economics and judgment and decision-making are either explicit tests of rational choice theory or developments of theories that started out that way. The focus of this chapter is on the success of rational choice theory as a behavioral hypothesis.

The chapter is organized as follows. Section 8.2 is an informal account of the "conditions" which preferences must meet if they are to be judged rational. This

I am grateful to Paul Anand, Ken Binmore, and Jon Leland for useful discussion and comments, and especially to Stephen Humphrey for a very useful review. Special gratitude is also due to John Conlisk for providing me with his unpublished data.

Table 8.1. Decision table showing states, options, and consequences

Option	State of the world			
	S_1	S_2	…	S_m
O_1	C_{11}	C_{21}	…	C_{m1}
O_2	C_{12}	C_{22}	…	C_{m2}
…	…	…	…	…
O_n	C_{1n}	C_{2n}	…	C_{mn}

section draws primarily on the influential work of Savage (1954), who will throughout act as our guide to what is rational. Sections 8.3–8.6 consider some empirical challenges to the view that behavior conforms to these rationality conditions. The challenges involve showing that the preference ordering over options depends significantly on differences between choice circumstances that, according to the rationality conditions, are normatively irrelevant. In Section 8.7 I will return to Savage's own discussion of the descriptive applicability of his conditions.

8.2 Introduction to Conditions of Rational Choice

In this section I provide a sketch of rational choice theory that will form the backdrop to the subsequent discussion. I adopt a highly simplified version of the framework used by Savage (1954) in his *Foundations of Statistics*. Savage's views are of particular interest to empirical researchers because he was concerned both with providing a normative system, and with the problem of whether this system was descriptively accurate.[1] On the latter point he was more optimistic than many subsequent researchers.

In Savage's framework the decision-maker is seen as choosing between a range of options, which have consequences that are uncertain because they depend on unknown states of the world. The general situation can be described with the aid of Table 8.1. We can illustrate this with the decision of whether to take an umbrella to the office on a day that threatens rain (Table 8.2). Stated this way the options are {Take umbrella; Don't take umbrella}, the states are {Rain; No rain}, and the consequences are the anticipated outcome from choosing each option, conditional

[1] A *normative* theory concerns what a rational agent *should* do, and a *descriptive* theory what he *actually* does.

Table 8.2. Decision of whether to take an umbrella

Option	State of the world	
	Rain	No rain
Take umbrella	Dry	Dry
Don't take umbrella	Wet	Dry

on whether or not it rains. In the table the consequences are simply "Dry" or "Wet". Both options and states are mutually exclusive and exhaustive, meaning that no other options and states are possible, although it is always possible to describe them at different levels of detail. The option "Don't take umbrella", for example, can be broken down into "Take a raincoat but no umbrella", and "Take neither a raincoat nor an umbrella".

Rational choice theory specifies certain minimal conditions which *should* be met for the choosing agent to be rational.[2] There are two kinds of justification for maintaining that these are conditions of rationality. First, the rationality conditions are such that people will, on reflection, *want* to conform to them. Although someone might occasionally express preferences contrary to the conditions, when the contradiction is pointed out to her, she will want to change the expressed preferences and bring them in line. Or, as Robert Strotz put it, "it would be a strange man indeed who would persist in violating these precepts once he understood clearly in what way he was violating them" (1953, p. 393).

Second, they are necessary for preferences to be consistent. Someone who does not comply with the conditions, it is argued, runs the risk of having a Dutch Book taken against them, or being turned into a "money pump". To illustrate a money pump, imagine someone who prefers an orange to a lemon, a lemon to an apple, and an apple to an orange. This is, as will be indicated in the next paragraph, an intransitive preference. Someone with this preference (and, it must be acknowledged, no memory or common sense) might be induced to give a lemon plus a small amount of money for an orange, to give an orange plus a small amount for an apple, to give an apple plus a small amount for a lemon, and so on, forever.

The rationality conditions are often presented as "axioms" or "postulates". In this chapter they are presented informally, with many technical details being dropped. Hence I use the term "condition". The first two conditions are:

[2] It should be emphasized that Savage's view (and the closely related one of von Neumann and Morgenstern 1947) are mainstream accounts of rational choice, and alternatives are available which drop or replace the rationality conditions described here. Whether these alternatives are indeed theories of "rational" choice is a matter for debate (see Sugden 1991).

Completeness: For any pair of options O_1 and O_2, either O_1 is preferred to O_2, O_2 is preferred to O_1, or the decision-maker is indifferent between them.

Transitivity: For any triple of options, O_1, O_2, and O_3, if O_1 is preferred to O_2 and O_2 is preferred to O_3, then O_1 is preferred to O_3.

Combined, these produce a *preference ordering* over options, so that all options can be rank-ordered and *equivalence classes* can be formed, which are sets of options between which the decision-maker is indifferent.[3] Consequently, these conditions are all that is necessary for what are called "riskless" choices, meaning choices between options when only one state can actually occur (Marschak 1964). Such a riskless choice is made when you must decide whether to take an umbrella given that it is already raining.

When risk and uncertainty enter the picture—meaning that the decision problem requires at least two states to be fully specified—further rationality conditions are required. I will describe two, which are enough to get our discussion started. The first is:

Independence: Given a pair of options O_1 and O_2 that have the same consequences under some states of the world, then these common consequences will not influence preference between the options. (This is also called the *sure-thing principle*.)

This principle appears innocuous. To illustrate with the umbrella example above, it merely says that because the consequences for you are the same regardless of what you choose if it doesn't rain tomorrow, these consequences should not influence your ultimate decision. Some readers may be tempted to object that the consequences described in the umbrella example are incomplete—it is annoying to carry an umbrella, so we should fill in the cells with consequences like "Dry, but carrying umbrella"—but for the moment we will assume that the consequences given are correct. The reader's hypothetical objection will soon be given its voice.

Closely related to the independence condition is dominance:

Dominance: If under at least one state of the world the consequence of O_1 is preferred to that of O_2, and if under no state of the world is the consequence of O_2 preferred to that of O_1, then O_1 is preferred to O_2.

One option dominates the other if, no matter what occurs, it leads to a better outcome. In the umbrella example, taking an umbrella dominates not taking one,

[3] Both conditions have been challenged as necessary foundations for rationality. Briefly, why should an agent have a preference between all options, even those they have never encountered before or will never have to choose between? (e.g. Sugden 1991; Binmore 2007). Likewise, cannot an agent have choice-set-dependent values that make the consequences of (say) getting a lemon from the set {lemon, orange} be different from those of getting a lemon from {lemon, apple}? (e.g. Anand, Ch. 6 above; Sugden 1991). The discussion of regret theory below hints at this issue, which goes well beyond the scope of this chapter.

because if it rains you stay dry if you take an umbrella, but if it doesn't rain it makes no difference.

This dominance condition is sometimes called *state dominance*, which can be distinguished from *stochastic dominance*. We can illustrate the difference with a simple gamble based on the toss of a coin. Imagine I toss a *single* fair coin and give you a prize based on the outcome. You have two options, which pay off as follows:

	Heads	**Tails**
O_1	$50	0
O_2	$40	0

If you choose O_1 I will pay $50 if heads comes up and nothing if tails comes up, while if you choose O_2 I will pay $40 if heads comes up and nothing if tails comes up. *State* dominance dictates that you should choose O_1 because it might make you better off than O_2 and will certainly not make you worse off. Now imagine that I offer you a similar gamble, but this time I toss two fair coins (A and B), and O_1 and O_2 represent the choice of coin. Now the decision table is as follows:

	A: Heads B: Heads	**A: Heads B: Tails**	**A: Tails B: Heads**	**A: Tails B: Tails**
O_1 (Coin A)	50	50	0	0
O_2 (Coin B)	40	0	40	0

In probability terms, O_1 dominates O_2 (a 50 percent chance of $50, versus a 50 percent chance of $40), but it is nonetheless possible for the O_2 coin to come up heads while the O_1 coin comes up tails. *Stochastic* dominance refers to dominance in these probability terms. More formally, if O_1 can be transformed from O_2 by iteratively increasing the probability of a preferred consequence over a less preferred one, or by improving a consequence while holding its probability constant, then O_1 stochastically dominates O_2.

In the next four sections I describe experimental findings that appear to demonstrate how changing the circumstances of choice can make preferences vary in ways that systematically depart from the rationality conditions. The effect of changing circumstances can be considered a kind of "spotlight" effect, analogous to the concept of attentional spotlight from cognitive psychology (e.g. Cave and Bichot 1999). The fundamental idea is one that is more or less the default among psychologists who study choice behavior. It starts with the notion that people, and indeed all organisms, have a severe limitation in their capacity to process information. Moreover, and indeed as a consequence of this limitation, they do not have well-defined preferences over all options. Rather, when making a choice, the valuation is done at the moment, based on information recruited by the specific choice circumstances. These circumstances direct attention to certain problem features—features that can be described using the terminology of states, options, and consequences—and deflect attention from others. The circumstances also act as memory cues and

triggers for the construction of scenarios. Our preferences are a function, always partly and sometimes decisively, of what the spotlight reveals.

This spotlight effect, under a variety of names and descriptions,[4] is the central principle underlying the experimental investigation of preference. We can even summarize the core research strategy of behavioral economics and behavioral decision-making as having three steps. The researcher

1. Infers, based on introspection or other evidence, that certain option characteristics will influence preference in a specific way.
2. Demonstrates logically how the influence from (1) can appear contrary to one or more of the rationality conditions.
3. Develops a method for shining a spotlight on those characteristics from (1) by varying some element of the decision context.

To give an example that will reappear below, a researcher might notice that he is willing to pay $200 for a high-priced car stereo when it is an option on a new car but not when it is an after-market purchase. From this he hypothesizes a general tendency to evaluate the magnitude of costs relative to "irrelevant" background costs (step 1); he then shows that this tendency can lead to preferences inconsistent with the incompleteness axiom (step 2), and then develops a method for showing this in the laboratory by asking people to decide how much effort they would exert to save $X when their attention is drawn to a large or small background cost (step 3). The next sections discuss many results from applying this research strategy, with each section's having a heading indicating how step 3 is accomplished.

8.3 Change the Consequences, but in a Way that is Irrelevant to the Rationality Conditions

In many demonstrations of apparently irrational preference reversals, agents are placed in two situations in which options *appear* to differ objectively,[5] but the rationality conditions do not admit the relevance of these differences. Among these are two of the most famous, and historically the earliest, examples of

[4] Variations of this view are found in, among many other places, Payne, Bettman, and Johnson (1993); Read, Loewenstein, and Rabin (1999); Simon (1978); Trope and Liberman (2003); and Tversky and Kahneman (1981).

[5] The reason for using the term "appear" becomes evident in Sect. 8.7.

Table 8.3. Allais paradox in state form

Option	State of the world		
	1%	10%	89%
O_1	1 million	1 million	1 million
O_2	0	5 million	1 million
O_1^*	1 million	1 million	0
O_2^*	0	5 million	0

systematic violations of rational choice conditions—the Allais and Ellsberg paradoxes (Allais 1953; Ellsberg 1960). Both demonstrate violations of the independence condition, suggesting that states that share the same consequence for all acts will influence the choice among acts. I will focus on the Allais paradox, often considered to pose the greatest challenge to the rationality conditions (e.g. Shafer 1986).

The classic Allais paradox concerns comparison between two choices, shown in Table 8.3 as a choice between O_1 and O_2 and between O_1^* and O_2^*. The table shows the two choices in a form (due to Savage 1954) that clearly brings out how they appear to violate independence.

The independence condition specifies that if you prefer O_1 to O_2, then you will also prefer O_1^* to O_2^*: As with the "No rain" column of the umbrella example, the consequences in the "89%" column should not influence your choice, because they are the same for both options. In fact, however, the 89% column does influence choice, with many more people choosing O_1, when the 1 million payment is certain, than O_1^*, when any payment at all is unlikely.[6]

As already mentioned, the Allais questions are not usually presented in the form given here. In fact, paradoxical choices are more likely to occur if the choice is given in a less transparent form. This was done by Allais (1953) himself, and by Conlisk (1989) in his first study: O_1 is a "certainty of 1 Million"; while O_2 is described as "an 89% chance of 1 million, a 10% chance of 5 Million, and a 1% chance of Nothing". By drawing attention to the fact that one option offers certainty and the other does not, the disproportionate weight put on the small probability of losing is magnified along with the attractiveness of the sure thing (e.g. Kahneman and Tversky 1979).

[6] Savage (1954) reports that he made the common and inconsistent $O_1O_2^*$ choice initially, but then on reconsideration changed his mind and decided that he wanted the consistent $O_1O_1^*$. His discussion of this process (on pp. 101–2) is essential reading.

Table 8.4. Allais questions reformulated as a lottery over lotteries

Option	State of the world		
	1%	10%	89%
O_1	Lottery I	Lottery I	1 million
O_2	Lottery II	Lottery II	1 million
O_1^*	Lottery I	Lottery I	0
O_2^*	Lottery II	Lottery II	0

This effect of problem description on the Allais task is profound. Conlisk (1989) showed that the paradox practically "goes away" if the problem is described as a "lottery over lotteries". He first informed respondents of two available lotteries:

Lottery I: Certainty of 1 million
Lottery II: 1/11 chance of nothing; 10/11 chance of 5 million

He then asked them to make the standard Allais choices, now described as a lottery over these lotteries. These choices are depicted, in state form, in Table 8.4. There were no longer any *systematic* preference reversals, and the same proportion of people chose O_1 over O_2 as O_1^* over O_2^*.

If the Allais paradox can be eliminated this way, it might seem that the problem has disappeared. Perhaps it arises only when we choose the wrong way to formulate the problem, and when we find the right way—one that does not bamboozle the respondent with irrelevant detail, or induce them to misrepresent the problem in their mind—their true preferences are revealed. Some variant on this viewpoint has been proposed by many researchers, and the debate continues (e.g. Binmore 2007). I will point out two objections to this defense of the rationality conditions.

First, merely because preferences can change, and even become consistent through choice redescription, does not mean that these now consistent preferences are the agent's true preferences, and that the inconsistent preferences were counterfeit.[7] Using the spotlight metaphor helps. Each form of the Allais question is a spotlight that highlights some aspects of the question and produces corresponding preferences. Two spotlights can produce the same preferences because they highlight the same aspects of the question, or because, while they highlight different aspects, they both produce the same preference ordering. From a theoretical perspective, there

[7] A third issue, deliberately avoided in this chapter, is that merely because choices are consistent does not mean that they are optimal from an instrumental point of view. An agent can consistently choose options that are worse for him over those that are better (some discussion: Sugden 1991; Read 2007).

is an inferential asymmetry—while inconsistency in one context demonstrates the operation of the spotlight, consistency in another context does not demonstrate that it is not operating.

The second objection to the defense invokes a theme that will be present throughout the remainder of this chapter. The fact that preferences vary as a function of choice description might be a greater challenge to the rationality conditions than any local violation of those conditions. If I simultaneously prefer O_1 to O_2, and O_2^* to O_1^* it is bad enough, but if I also sometimes prefer O_1^* to O_2^* and/or O_2 to O_1 I am doing more than just violating the independence condition, and I am prime material for money pumps and Dutch Books. This point has led many researchers to propose a further rationality condition that is perhaps more fundamental than the original conditions themselves:

Invariance: Preferences do not depend upon *irrelevant* aspects of the context, of the option descriptions, or of the procedures used to elicit those preferences.

An alternative way of putting this is that the rationality conditions are defined over the consequences of options and the states under which they occur, and not over how those options are described or how preferences are obtained. The key word in the definition is *irrelevant*, and this word could bring an uncomfortably subjective element to the rationality conditions. Since every preference reversal is a consequence of *some* variation in the options, we must then decide whether a difference is relevant or not.[8]

To illustrate the problem, while most might agree that Conlisk's two versions of the Allais paradox should not have yielded different preferences, it is not obvious that the differences between the two original choice pairs are irrelevant in this way. Loomes and Sugden (1982; see also Bell 1982) proposed that one way the two pairs might differ is in the degree of regret the respondent anticipates if he does badly. In the first pair, choosing O_2 and ending up with nothing could produce great regret because the agent will know with certainty that she would have had $1 million if she had chosen differently. To avoid this regret, she might choose O_1. In the second pair, however, choosing O_2 and ending up with nothing does not lead to much regret because you doubt that you would have got anything anyway. Perhaps this is a good explanation of the Allais paradox, and it does "rescue" the rationality conditions. But similar justifications may not be available in every case. It is a major challenge to determine whether a given difference between options does or does not justify the resultant variation in preference. If invariance is not obviously followed, and if we permit some at-first-glance irrelevant variations in the task to be rationalized using a method like the one just described, then deciding whether the rationality conditions are being met becomes a matter for judgement. We will

[8] Anand's discussion of intransitivity in Ch. 6 above and elsewhere (Anand 1993) involves a close examination of what differences are relevant in the context of intransitive choice.

return, inevitably, to this question once we have seen more apparent violations of the principle of invariance.

The Allais paradox is noteworthy because, for many people at least, the rationality of their "paradoxical" choices is compelling even after reflection. It is difficult to persuade people that the differences between the two choice tasks are in fact irrelevant (Slovic and Tversky 1974), a fact that will be well known to anyone who uses them in teaching. The preference reversals discussed next are more like those observed in comparisons between Conlisk's different ways of describing the same problem. The fact that there is a problem is obvious, and most people will feel that something about the preference pattern needs to be explained.

8.4 Change the Method by which Options are Evaluated

If preferences for options that we agree are *exactly the same* change systematically in response to apparently irrelevant changes in circumstances, then we have a prima facie case for violations of invariance and decidability. Of the many ways to elicit such preference changes, perhaps the most venerable and reliable is through changing the valuation procedure. For instance, we can ask people to choose between pairs of options, to reject one option from a pair, to rank several options, to assign a value to them, or to otherwise specify an option that is equivalent to another option. It turns out that preferences are systematically influenced by all these procedures, and others as well. Simply stating your preferences in one way versus another seems to direct the spotlight on different reasons for having those preferences, and thus change the preferences altogether.

The term "preference reversal" was coined by Lichtenstein and Slovic (1971) to describe a systematic discrepancy between choice and pricing. They offered respondents a choice between gambles, and later asked them to price each gamble separately. One gamble in each pair offered a high probability of winning a small prize (the so-called *P*-bet, e.g. a 95 percent chance of winning $2.50 and a 5 percent chance of losing $0.75), the other offered a moderate probability of winning a larger prize (the *$*-bet, e.g. a 40 percent chance of winning $8.50, a 60 percent chance of losing $1.50). There were many preference reversals, with the *P*-bet being chosen more often than the *$*-bet, but the *$*-bet being priced higher. This was a very strong effect: in their first experiment, 73 percent of respondents always put a lower price on a *P*-bet they had chosen than on a *$*-bet they had not chosen. Lichtenstein and Slovic (1973) replicated this with gamblers in Las Vegas, and, in a later paper, Grether and Plott (1979) imposed all the controls available to experimental

economics (such as greater task transparency and the use of real monetary incentives) and replicated Lichtenstein and Slovic's results almost exactly.[9]

An appealing explanation for preference reversals is that they arise from what Tversky, Sattath, and Slovic (1988) called *compatibility* effects. When values are expressed on a quantitative dimension (such as price), they are disproportionately influenced by information in the same or similar currency to that dimension. When pricing a gamble, for instance, the price put on it will be highly influenced by the potential payoffs. Hence, the *$*-bet gets a higher price than the *P*-bet. Compatibility effects can be viewed as one instantiation of the anchoring-and-adjustment heuristic, by which quantities are estimated by starting with an available estimate (even an arbitrary one), and then incompletely adjusting that estimate in the appropriate direction based on other information. When pricing a gamble, the logical starting point is the prize to be won, which is then adjusted downward based on the probability of that prize. The insufficient adjustment leads the *$*-gamble to be overpriced.

These classic preference reversals are a special case of a systematic difference in preferences elicited under *choice* and *matching*. Matching is the procedure, like pricing, in which one option is given (e.g. a gamble) and the respondent specifies another option equivalent in value to the first (e.g. a price). Tversky, Sattath, and Slovic (1988) showed this using a series of questions such as the following (p. 373):

> About 600 people are killed each year in Israel in traffic accidents. The ministry of transportation investigates various programs to reduce the number of casualties. Consider the following two programs, described in terms of yearly costs (in millions of dollars) and the number of casualties per year that is expected following the implementation of each program:

	Expected casualties	**Cost**
Program X	500	$55 million
Program Y	570	$12 million

(Notice that in this problem the future states are not given explicitly, as in the Savage table, but rather given as a summary of consequences multiplied by their probability and summed over all possible states.) One group chose between the two programs, while another group (the *Implicit choice* group) saw the matrix with one value missing, which they then completed to make the two programs equal in

[9] Grether and Plott (1979) also provided evidence about whether people will spontaneously recognize inconsistencies. Their respondents first priced several items, then immediately made a set of choices, and then priced the remaining items. The likelihood of a preference reversal was not influenced by whether choice preceded pricing or pricing preceded choice, suggesting that the choice inconsistency was not noticed. Lichtenstein and Slovic had deliberately separated their two tasks by some time and with a lot of filler tasks, presumably because they were concerned that their respondents would strive to be consistent. Grether and Plott's research suggests that inconsistencies must be very obvious for them to be noticed.

value. To see how the responses were compared, imagine someone who has chosen Program X, implying they would be willing to pay $43 million to save seventy lives. Now, suppose she is asked to complete the matrix below:

Option	Expected casualties	Cost
Program X	500	??
Program Y	570	$12 million

For choice and implicit choice to be consistent, the value specified should be no less than $55 million, otherwise she would prefer Program Y. For each blank, there is a similar range of values consistent with a choice of X over Y (e.g. casualties in Program X can be no more than 570, and so on), and values outside of this range indicate an inconsistency. In fact, for this particular question Tversky *et al.* (1988) found that 67 percent of the direct choices were for Program X, as compared with only 4 percent of the implicit choices.

In general, Tversky *et al.* (1988) observed many more direct choices of the option superior on the *primary dimension,* meaning the dimension that people view as most important. In the Israeli traffic example, the primary dimension is the number of lives that can be saved. Although choice can be made using lexicographic procedures, such as choosing the option superior on the most important dimension, matching cannot. Matching demands an explicit tradeoff between option features. For instance, in the above example it is possible to choose Program X using the lexicographic rule that "life has no value". But in matching, this rule would imply an infinite cost for Program X.

A further example of preference reversals resulting from differing valuation procedures is the comparison between choice and ranking, reported by Bateman, Day, Loomes, and Sugden (2006). They conducted a study testing Allais-type preferences, like those above but with some modifications.[10] Respondents either chose between several pairs of gambles, or else rank-ordered all the gambles contained in the pairs. They found that Allais-type preferences were common in choice (replicating earlier studies), but not in ranking. Ranking did not eliminate all apparent irrationalities, however, and actually increased the rate at which another rationality condition, betweenness (a derivation of those already given), was violated. Bateman *et al.* give a theoretical explanation for their finding, but for our purposes the most important result is that two theoretically interchangeable measurement procedures yield strikingly different results.

One important difference between binary choice and ranking is in the number of options on the table. In binary choice there are two, while in ranking there are three or more, so that even if the agent focuses exclusively on the preference order

[10] They tested the *common ratio effect*, an extension of the original Allais task. The finding is that if a person is indifferent between two lotteries, one riskier than the other, then they will prefer the riskier of the two if all the probabilities in the two lotteries are scaled down through multiplication by a common positive constant < 1.

between O_1 and O_2, there is at least an "O_3" to which both options can be compared. It turns out that another source of violations of the rationality conditions is that preference order depends even on options that are never chosen.

8.5 Include More Options in the Consideration Set

A seemingly trivial implication of the rationality conditions is that if an additional option is taken into consideration, it will not influence the relative value of those already there. That is, if O_1 is preferred to O_2, and O_3 is then made available, the only permissible rankings of the three options are those that maintain the preference of O_2 over O_3. For example, to take the example from Table 8.2, suppose you are planning not to take your umbrella today, but then somebody reminds you that you could take a raincoat instead; this should not make you decide to take your umbrella. This follows both from the Decidability and Transitivity conditions, and is sometimes called the *independence of irrelevant alternatives*.

Bateman *et al.*'s (2006) comparison between ranking and choice, however, suggests that this might not always be true. For ranking and choice to differ, it must be that sometimes O_1 is preferred to O_2 in binary choice, while O_2 is preferred to O_1 when additional options are on the table. It will not surprise the reader, therefore, to learn that it is easy to construct scenarios in which many people choose O_1 from the set $\{O_1, O_2\}$ and then switch to O_2 when given the set $\{O_1, O_2, O_3\}$.

One robust set of such context effects will be explained with the aid of Figure 8.1. It depicts factors that might influence the choice between O_1 and O_2, two options that vary on two dimensions, with higher values on the dimensions corresponding to improvement. The dimensions could be payoff and probability, price and quality, or anything. In addition to O_1 and O_2, two further options are depicted, labeled D and E for *dominated* and *extreme*. D is dominated by O_1 but not by O_2; E is an extreme item that is better than B on one dimension, and worse than B on the other. Examples of such items are gambles like the following (e.g. Wedell 1991; Herne 1999):

O_1: 50 percent chance of \$100
O_2: 20 percent chance of \$500
$D(O_3)$: 19 percent chance of \$490
$E(O_4)$: 5 percent chance of \$1500

There are two effects of adding the third item, either D or E. The first is the usual substitution effect, in which some people choose the added option rather than one of the original pair. Of course, this does not often happen with D, since it is

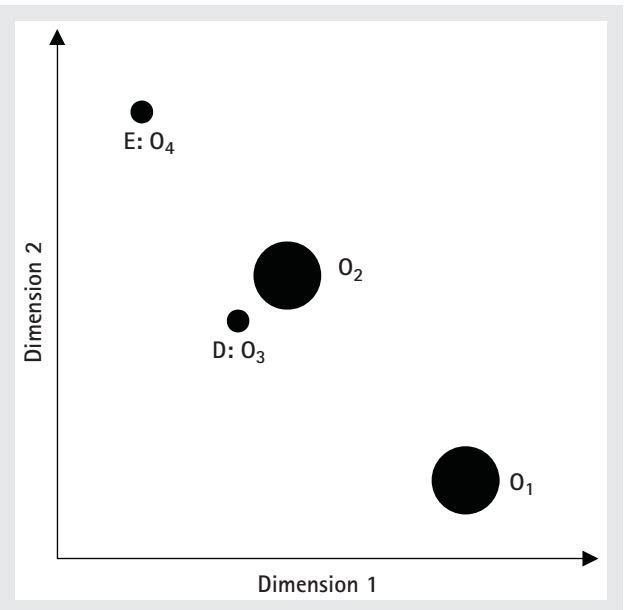

Fig. 8.1. Adding additional items to a choice set.

dominated by O_2. The second effect is that people's preferences among the original choice pair changes, with choices of O_1 switching to choices of O_2. The effect of adding D is called the *asymmetric dominance effect* (and sometimes the *attraction effect*), and the effect of adding E is called the *compromise effect* (sometimes *extremeness aversion*).[11]

The labels D and E come from the proposed locus of the spotlight effect. The idea is that one way people determine what they prefer is by thinking of reasons for having the preference. The reasons are "justifications" elicited by the context, and different contexts elicit different justifications (e.g. Simonson 1989; Shafir, Simonson, and Tversky 1993). For these options, adding a new item gives an additional reason for choosing O_2 over O_1 which can overturn the preference expressed without the new item. The new reason is that O_2 is clearly better than D, or that it is a compromise between the two extremes O_1 and E. The role of reasons can be imagined by thinking of what someone would have to do to justify their choice to someone—they could say, for instance, that "Gamble O_2 is a compromise between the boring O_1 and the too-risky E". This, however, is a contingent argument. If the added option had been different, one that instead made O_1 the compromise candidate, matters would be quite different.

The number of options has also been shown to matter in investigations of *joint* versus *separate* evaluation (Hsee 1996). In joint evaluation, the menu consists of

[11] The original reports of these effects were in studies of consumer behavior. Huber, Payne, and Puto (1982) first described the asymmetric dominance effect, and Simonson (1989) the compromise effect.

several items, so that people are able to make direct comparisons during evaluation. In separate evaluation, however, they are evaluated one at a time. The original preference reversal studies of Lichtenstein and Slovic are one example of such a situation, but in those studies the evaluation procedure (choice versus pricing) was confounded with whether there were one or two gambles (corresponding to separate and joint evaluation) being evaluated. It turns out, however, that even when the valuation procedure is held constant, joint and separate evaluation can produce different preferences. A compelling example, because it involved "real-world" transactions and real money, was described by List (2002), who replicated the method and manipulations of Hsee (1998) with dealers and collectors at a sports card show. List's participants were invited to bid on one or both of two sets of sports cards. The *inferior* set contained ten high-quality (near-mint) cards, while the *superior* set contained thirteen cards, including the ten high-quality cards plus three low-quality cards. The estimated "true" market price of the superior set was about $3 higher than that of the inferior set. There were three groups: one group bid separately on the inferior set, another bid on the superior set, and a third bid on both. The result was that during separate evaluation, both dealers and collectors (but especially collectors) bid more for the inferior set of cards, but during joint evaluation they both bid more for the superior set.[12]

Hsee's (1998) argument, reflecting a recurring theme, is that option features vary in the degree to which they can be easily evaluated in isolation, and those features that are difficult to evaluate will show a disproportionate increase in importance during joint evaluation. The traders put too little weight on the *number* of cards when looking at only one set,[13] while the poor quality of some of those cards is very salient when looking at the superior set. List's study shows just how easy it is to find features that are difficult to evaluate in isolation.

8.6 Describe the Options in a Different Way

In the previous sections, there were some differences between situations that produced preference reversals. Either the evaluation method or the context was altered.

[12] The fact the asymmetry was smaller among dealers points to the importance of learning and experience in the application of rational choice theory. Binmore (e.g. 1997, pp. 19–20) argues that rational choice theory is applicable only to situations in which "the conditions are favorable", and specifically that "there is a good deal of evidence that adequately motivated people sometimes can and do learn to choose rationally if they are put in similar situations repeatedly and provided with meaningful feedback on the outcome of their previous decisions".

[13] A robust finding is that people are quite insensitive to *scope* of a good when placing a value on it (Frederick and Fischhoff 1998).

No such differences are actually required to produce preference reversals. Even seemingly modest redescriptions of options can accomplish the task, as demonstrated by the widely cited "Asian disease" problem of Tversky and Kahneman (1981). Their respondents chose between two medical responses to an epidemic that, in the absence of some response, would certainly kill 600 people. In one version (they called this the *loss frame*) respondents chose between "400 deaths for sure" and "a one third chance of 600 deaths, and a two thirds chance of no deaths". In the second version (the *gain frame*) they chose between "saving 200 lives for sure" and "a one third chance of saving everybody, and a two thirds chance of saving nobody". A majority prefer to save 200 lives for sure in the gain frame, but to take a chance of saving no one in the hope of saving all in the loss frame. Although when presented together, the two choices are transparently the same, the different descriptions highlight different features of the options.

Kahneman and Tversky attributed this result to a general pattern of risk seeking for losses and risk avoidance for gains. A further demonstration of this produced one of the most striking examples of how relatively modest variations in problem description can lead to changes in preference. Kahneman and Tversky (1979) asked separate groups to make the following choices:[14]

Group 1 The following two gambles will be enacted simultaneously. Choose one from (a), and one from (b)

(a) Choose between:
  1A) Win $240
  1B) 25% chance to win $1000;
      75% chance to win nothing

(b) Choose between:
  1C) Lose $750
  1D) 75% chance to lose $1000;
      25% chance to lose nothing

Group 2 Choose between:

  2A) 25% chance to win $240;
      75% chance to lose $760
  2B) 25% chance to win $250
      75% chance to lose $750.

The pair most frequently chosen by Group 1 was {1A, 1D}—about 65 percent of people chose this combination—while for Group 2 everyone chose 2B. The latter is not surprising, since 2A stochastically dominates 2B. However, if 1A and 1D are combined, it can be seen that they *are* option 2A, while combining 1B and 1C is the

[14] It doesn't matter whether the same group or different groups do the task—the results are identical. I have used this problem in classes for many years, and these within-respondent results are invariably indistinguishable from those reported by Kahneman and Tversky (1979).

dominant 2B. Kahneman and Tversky explain this using the same analysis as used for the Asian disease problem. The Group 1(a) question is in the domain of gains and leads to risk avoidance, while the Group 1(b) question is in the domain of losses and leads to risk seeking.

There are three important features of responses to these questions. First, they demonstrate how we evaluate outcomes from the perspective of a reference point (i.e. as losses or gains) and that the reference point is highly malleable. Second, we see a violation of (stochastic) dominance, so that unlike most systematic variations in preference, including those described earlier, not only do people have different preferences under different circumstances, but it is obvious when they are making a mistake.

But the most important lesson from these questions is that it demonstrates the central feature of the attentional spotlight. The choice of {1A, 1D} shows that respondents failed to integrate their chosen options into a portfolio, but rather treated them as separate options, as if each choice they were making was the only one they would ever have to make. And they did this despite the fact that the two gambles were explicitly described as being "enacted simultaneously" and were given one after the other. Very similar results, often involving quite astonishing degrees of isolation, have been described by Redelmeier and Tversky (1992) and others (e.g. McCafferey and Baron 2006; for a review and more examples see Read, Loewenstein, and Rabin 1999).

Although I earlier described the Allais paradox as a case in which the consequences were changed, but in a way that is irrelevant to the rationality conditions, the above result might lead us to reconsider this conclusion. Perhaps the difference in consequences between the two Allais questions is itself an illusion due to the operation of the spotlight. Consider the following variant of the Allais problem, itself adapted from one of Conlisk's (1989) variations. This shows the second pair of Allais options (O_1^* and O_2^*) transformed into a two-stage game.

The decision-maker has a ticket numbered between 1 and 100 but does not know which. If the number is between 11 and 100 he gets nothing, but if it is between 1 and 10 he has a chance of a prize, with the chance depending on whether he chooses O_1^* or O_2^*. Consider two occasions on which he can make this choice, either at CN_1 (choice node 1) or at CN_2. If he chooses at CN_1 he will probably choose O_1^*, as 202/236 people did in Conlisk's study. On the other hand, if he chooses at CN_2, he is now much more likely to choose O_2^*, as 95/212 did in Conlisk's study.[15] The rate of systematic preference change will be very high.

It is difficult to rationalize this. Anyone faced with a decision like that at CN_2 will have come to that point by way of a series of choices and accidents, like those at CN_1, that were not *guaranteed* to achieve that choice point. This uncertainty is represented by a single chance node in Figure 8.2, but in reality it would be a

[15] John Conlisk kindly provided these data which were not included in his 1989 paper.

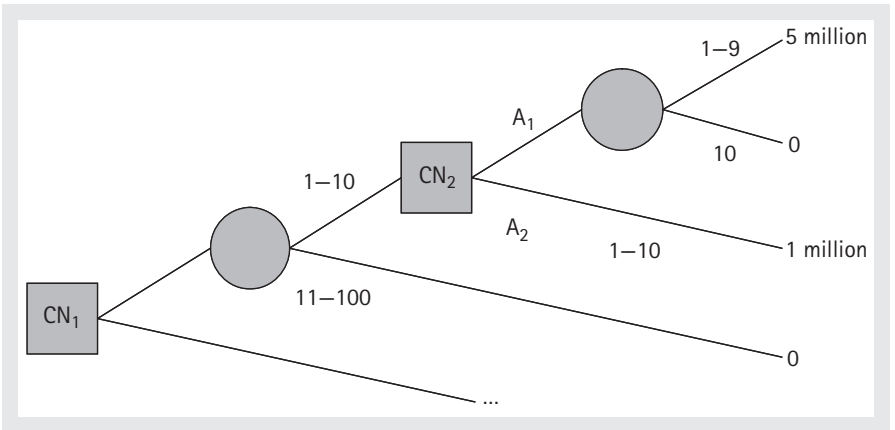

Fig. 8.2. Sequential Allais decisions.

multiplicity of chance occurrences—for instance, after years of struggle an entrepreneur has the opportunity to sell out for a big killing, or to take a chance by continuing with the chance of an even bigger killing or failing altogether. If the entrepreneur was required to decide before he knew he would get to that stage (at the equivalent of CN_1) he would very likely make different and more risk-seeking decisions than if he decided afterwards (CN_2).

This is a very general pattern. Whenever we are faced with a choice, the fact that we have that choice at all is contingent on events that might not have occurred. Consequently, if we had decided in advance, based on the compounded expectations in which many consequences along with their probabilities or uncertainties were combined, we would have made different decisions than if we decide based on the local returns.

The framing effects just described work because people do not consider information or consequences that are not explicitly included in the frame they have been given. They either put too little weight on unspoken consequences (as in the lives saved or lost in the Asian disease problem), or do not combine sequences of outcomes and uncertainties into composite or "portfolio" prospects. The effects of framing can result from even more subtle manipulations. One such manipulation is simply to divide states into increasing numbers of categories (e.g. Fischhoff, Slovic, and Lichtenstein 1978; Fox and Clemen 2005). I will describe one example of this, which Starmer and Sugden (1993) call "event-splitting".

Starmer and Sugden offered people a choice between lotteries. The respondent could draw a lottery ticket from a pool, and would win a prize contingent on the number that came up (the state of the world) and the option they had chosen. The choices were described using a *strip display*, a way of indicating separately, for each ticket number, what the outcomes will be for each option. This allowed the researchers either to combine or to divide ranges of states that yielded the

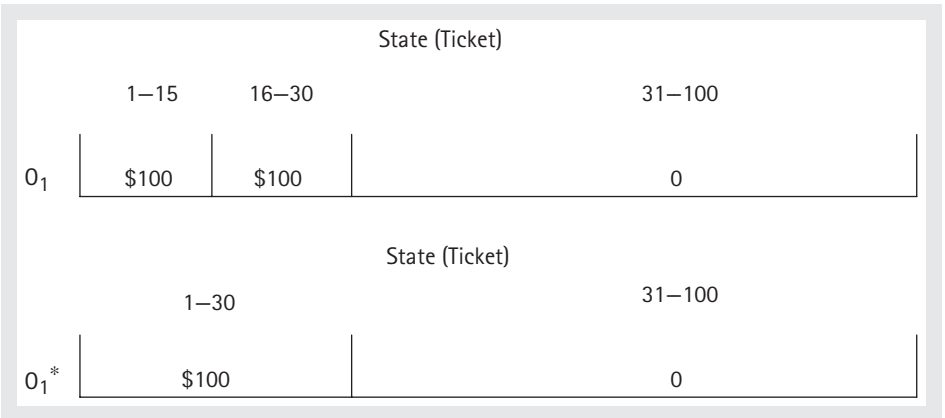

Fig. 8.3. Two strip displays like those used in the event-splitting study of Starmer and Sugden (1993).

same outcome. An example of two strip displays for the "same" option is given in Figure 8.3. Observe that O_1 and O_2 are identical, except that in O_1 the numbers 1–30 are divided into two states.

Starmer and Sugden obtained choices between lotteries like O_1 and a standard lottery, and lotteries like O_1^* and the same standard lottery. The standard lotteries were always described in the same way. Since O_1 and O_1^* are formally identical, the same proportion of people should have preferred them to the standard, otherwise this would imply that one of two identical options is preferred to the other. In fact, the option, like O_1, for which the state space for the most attractive outcome was subdivided, was typically preferred to the one in which that space was undivided.

Starmer and Sugden's result, along with an abundance of related findings, show that the weight we put on information is highly dependent on how closely we scrutinize it. We can return to our example of the decision about whether to take an umbrella when it rains. What this result suggests is that if we divided the state "Rain" into the two states "Some rain but no more than 2 centimeters", and "More than 2 centimeters" we would then be more likely to take that umbrella.[16]

[16] This point is made briefly, but the central problem is related to one raised by Luce and Raiffa (1957). Suppose we have complete ignorance about which state of the world obtains. The "principle of insufficient reason" then dictates that we should treat each state as equally likely. But given that the states about which we are ignorant can be divided and combined in various ways, each equally sensible, it follows that the principle of insufficient reason will dictate different probability distributions over the same states. It turns out that people are prone to the bias that might be expected from such a naïve application of this principle (Fox and Clemen 2005).

8.7 Small Worlds and the Attentional Spotlight

In this chapter I have reviewed, or rather sampled, findings from the experimental study of preference. Taken in their entirety, these studies—along with literally hundreds of others[17]—challenge the view that the rationality conditions described in Section 8.2 are a good model of human behavior. In this section I consider the nature of this challenge in more detail, continuing to allow Savage to stand in for the prototypical rational choice theorist.

When Savage discussed his rationality conditions, he did so by describing what he called a *small world*. His example was the decision of whether, when making an omelette, to break a sixth egg, which might be rotten, into a bowl containing five good eggs, or to be cautious and first break it into a saucer. The states of the world were whether the egg was bad or good, and the only consequences considered were the quality of the omelette and the quantity of dishwashing. The example (from Section 8.2) of whether or not to carry an umbrella is another small world. The worlds are *small* because the states, the options, and the consequences are not described in the finest possible level of detail.

Savage also envisioned an idealized *grand world*, which differed from the small world in that the states in the columns were descriptions of all possible distinctions between present and future states that might occur, the rows were all possible options one had for living one's life (starting from the present), and the consequences were likewise comprehensive descriptions of what choosing a given life strategy would be given those possible future states. In Savage's words (1954, p. 83): "To make an extreme idealization . . . a person has only one decision to make in his whole life. He must, namely, decide how to live and this he might in principle do once and for all."

A small world, therefore, differs from the grand world in that many distinctions are ignored, and the focus is on only those distinctions that are relevant to the decision. By collapsing the grand world in different ways, we can create an infinite number of small worlds. As the quotation also makes clear, a grand world is an idealization, never to be achieved in actuality. Savage himself said that "the task implied in making [a grand world] decision is not even remotely resembled by human possibility" (p. 16). All that it is possible to achieve are small worlds at different levels of detail. Savage understood that the rationality conditions, if they were to have any significance whatsoever, had to be applicable to small worlds.

[17] Several journals continue to publish new anomalies relative to consistency conditions like those of Savage in almost every issue. These include *Journal of Behavioral Decision Making*, *Organizational Behavior and Human Decision Processes*, and *Journal of Consumer Research*. Increasingly, these anomalies are finding their way into mainstream economics journals as well.

The collapsing of grand worlds into small worlds has obvious parallels with the process of shining a spotlight on certain aspects of the world and ignoring others. When deciding whether to take an umbrella to work, for instance, it is natural to focus on states relevant to the decision (e.g. "Rain" versus "No rain"), and on consequences relevant to the use of umbrellas ("Will I get wet or not?"), but to ignore everything else. This can be described as the formulation of a small world, or as the deployment of the attentional spotlight.

But if the grand world can be collapsed into different (and not necessarily even nested) small worlds, this raises the question of *which* small worlds the rationality conditions should apply to. One possible view is that as long as the conditions hold over any small world, they are satisfied. For instance, if the small world of Table 8.2 accurately describes how a decision-maker thinks of the problem at the moment of choice, then so long as he takes his umbrella, the conditions are satisfied. If changes in the evaluation circumstances change the small world in *any* way, the rationality conditions should then be reapplied to this new situation. From this perspective, the rationality conditions are not definitely violated by any of the studies described above, and perhaps they never can be. It is unlikely that many knowingly and deliberately choose to violate these conditions when their applicability is obvious, and if their applicability is not obvious, then no violation is possible. We can illustrate this point by referring back to Kahneman and Tversky's two-stage question from Section 8.6. If the small world to which the rationality conditions are intended to apply is at the level of the question (i.e. the Group 1(a) and Group 2 questions taken separately), then no individual choice can violate those conditions, and no choice *pattern* is even relevant, since it is a pattern formed by combining separate and independent small worlds. Likewise, the apparent inconsistency between the choices by Groups 1 and 2 is not relevant to the rationality conditions because it also involves comparisons between different small worlds. Finally, when, as in Group 2, the small world involves a clear opportunity to violate a rationality condition (in this case stochastic dominance), nobody does so.

Many will feel that so stripping the rationality conditions of their content is an unsatisfactory way to shield them from refutation. Savage did not propose this, and in fact provided a more restricted description of the small worlds to which the rationality conditions should apply. He focused on what he called *real microcosms*, which are small worlds that locally satisfy the rationality conditions (as in, the decision-maker takes the umbrella given Table 8.2) but, in addition: (a) the probabilities of states are the same in the small world as in the grand world, and (b) the utility of options is the same in the small world and the grand world (see Savage 1954, p. 86). A small world that meets the rationality conditions internally but does not satisfy these additional requirements is, in Savage's own words, "a pseudo-microcosm that is not a real microcosm". He was, as the following passage shows, optimistic that the rationality conditions were applicable to circumstances beyond that of the small world: "I feel, if I may be allowed to say so, that the possibility of

being taken in by a pseudo-microcosm that is not a real microcosm is remote, but the difficulty I find in defining an operationally applicable criterion is, to say the least, ground for caution" (Savage 1954, p. 90).

The reason for Savage's optimism is the *psychological* premise, which he shares with many researchers who come to behavioral studies from the discipline of economics, that while someone might initially be misled by a small world, they will eventually settle on a real microcosm when they have time to reflect on their preferences. Savage illustrates this with the following example: "A man buying a car for $2134.56 is tempted to order it with a radio installed, which will bring the total price to $2228.41, feeling that the difference is trifling. But when he reflects that, if he already had the car, he certainly would not spend $93.85 for a radio for it, he realizes that he has made an error" (p. 103).

Savage's intuition that an expensive radio might be chosen when its price is added to the sticker price of a car, but not when it is seen in isolation, is undoubtedly correct.[18] But Savage presumes that the man changes his mind because "he has made an error" which is now corrected. But there is nothing in the rationality conditions themselves that shows that either option is the correct one (cf. Shafer 1986, for a related discussion of the same passage from Savage).

The difference between the small world and the spotlight viewpoints revolves around the question of whether or not reframing a decision moves the decision-maker towards a real microcosm. According to the small-world view, this is what happens. We can see this reflected not only in the above passage, but more forcefully yet in Savage's reevaluation of his own preferences following his mistake when answering Allais's question (see n. 6). According to the spotlight view, however, while rethinking highlights different aspects of choice, the changed preferences are not necessarily "correct" or even better than those they have replaced. Rather, they are determined by the path taken by the process of thinking and rethinking. Moreover, since most people are like Savage, in that they do wish to maintain consistency in their preferences, even local conformity to the rationality conditions does not indicate that we have located a real microcosm, since a different path of thinking could have led to different preferences that conformed equally well to those conditions.

To illustrate this, consider a case of intransitive preference. It is rare to find situations in which people openly state that they simultaneously prefer A to B, B to C, and C to A. Indeed, experimentalists traditionally separate questions with filler items so that respondents will not realize they are making a mistake. According to the small-world view, if we give an agent the three pairwise choices, he or she will eventually work out the *correct* preference ordering for the three options. According to the spotlight view, on the other hand, the final preferences of this

[18] This passage, like many of Savage's examples, inspired an extensive and ongoing research program, starting with Tversky and Kahneman (1981).

agent will depend on the order in which the choices are presented. If we start with the choice between C and A, then we will end up with the preference order C–A–B, while if we start with the choice between A and B, we will end up with the order A–B–C.[19]

This phenomenon was characterized as *coherent arbitrariness* by Ariely, Loewenstein, and Prelec (2003, 2006), who investigated it with a series of clever experiments. In one study, people first stated how much they were willing to pay for an item, such as an average bottle of wine. The situation was engineered so that some would give higher prices than others, by first asking respondents whether they would pay a dollar amount equal to the last two digits of their social security number, and then to state their maximum willingness to pay. This naturally resulted in a strong anchoring effect, with people willing to pay *much* more when their social security number ended with a high number than a low one. The prices they gave to other goods were also consistent with these initial prices. Everyone, for example, was willing to pay more for a fancy bottle of wine than for an average one. But because the average bottle was so overpriced in the "high number" condition, the expensive bottle was correspondingly overpriced. The preferences were consistent, but based on an arbitrary starting point.

The same authors, indeed, later showed that even whether an activity was positive, so that one should pay to do it, or negative, so that one should pay not to do it, depends on apparently trivial features of the decision context, and that once this is decided, people will be consistent with that judgment. In their study, people asked how much they would pay to hear Ariely read from Walt Whitman's *Leaves of Grass* would pay more to enjoy a longer reading, while a matched group asked how much they would have to be paid to hear him required more if they had to *endure* a longer reading (Ariely, Loewenstein, and Prelec 2006).

Research like Ariely *et al.*'s provides a good place to conclude this survey, because it both demonstrates the ultimate implications of the spotlight view and raises other issues about rationality to which I will only allude. In this chapter I have deliberately restricted my focus to the question of whether preferences do, in any meaningful way, conform to rationality conditions like Savage's. The experimental evidence reviewed gave scant support to the view that the rationality conditions are adhered to. Instead, preferences shift dramatically as we change the circumstances in which they are to be expressed. I likened this to the operation of a spotlight, where each new circumstance highlights a restricted subset of task features and leads to the creation of a correspondingly restricted (and distorted) representation

[19] This is the primary reason why we do not expect individuals to be transformed into money pumps by repeating the same choices. Even if a person can be found to have a cycle of preferences, they will refuse to engage in a transparent cycle of transactions. But as Rabin and Thaler (2001) have observed, this does not mean that we cannot create money pumps across people. If we can identify situations in which the majority of people have a cyclical preference and different endowments, we can be sure to earn money by offering a single trade to each person.

of the world. This representation is not, except accidentally, coordinated with other representations that we would have adopted if the spotlight had been different. In Savage's words, the spotlight directs us to a "pseudo-microcosm that is not a real microcosm". I have not unconsidered the problem of whether rationality can and should be identified with some degree of *local* conformity to the rationality conditions. That is, given that in the broad sense we are inevitably inconsistent, is it nonetheless better to be consistent as far as we can be? No evidence is available to answer this question. Ariely *et al.* show some ways in which local consistency can lead to error, but they have not shown (or tried to show) that inconsistency would lead to *less* error. To answer this question, we need different criteria for rationality than the conditions of Savage or von Neumann and Morgenstern, criteria which can rank choices as good or bad, so that we can know whether we are more or less likely to choose the good options when we are consistent or when we are inconsistent.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axioms de l'école Américaine. *Econometrica*, 21, 503–46.

Anand P. (1993). The Philosophy of Intransitive Preference. *Economic Journal*, 103, 337–46.

Ariely, D., Loewenstein, G., and Prelec, D. (2003). Coherent Arbitrariness: Stable Demand Curves without Stable Preferences. *Quarterly Journal of Economics*, 118, 73–106.

——— ——— ——— (2006). Tom Sawyer and the Construction of Value. *Journal of Economic Behavior and Organization*, 60, 1–10.

Bateman, A., Day, B., Loomes, G., and Sugden, R. (2006). Ranking versus Choice in the Elicitation of Preferences. Working Paper, University of East Anglia.

Bell, D. (1982). Regret in Decision Making under Uncertainty. *Operations Research*, 30, 961–81.

Binmore, K. (2007). *Rational Decisions*. Princeton: Princeton University Press.

Cave, Kyle R., and Bichot, Narcisse P. (1999). Visuo-spatial Attention: Beyond a Spotlight Model. *Psychonomic Bulletin and Review*, 6/2, 204–23.

Conlisk, J. (1989). Three Variants on the Allais Example. *American Economic Review*, 79, 392–407.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643–69.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978). Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 330–4.

Fox, C., and Clemen, R. T. (2005). Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior. *Management Science*, 51, 1417–32.

Frederick, S., and Fischhoff, B. (1998). Scope (In)Sensitivity in Elicited Valuations. *Risk, Decision, and Policy*, 3, 109–23.

Grether, D. M., and Plott, C. R. (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. *American Economic Review*, 69, 623–38.

Herne, K. (1999). The Effect of Decoy Gambles on Individual Choices. *Experimental Economics*, 2, 31–40.

Hsee, C. (1996). The Evaluability Hypothesis: An Explanation of Preference Reversals between Joint and Separate Evaluations of Alternatives. *Organizational Behavior and Human Decision Processes*, 46, 247–57.

—— (1998). Less is Better: When Low-Value Options are Judged More Highly than High-Value Options. *Journal of Behavioral Decision Making*, 11, 107–21.

Huber, J., Payne, J. W., and Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research*, 9, 90–8.

Kahnemann, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263–91.

Lichtenstein, S., and Slovic, P. (1971). Reversals of Preference between Bids and Choices in Gambling Decisions. *Journal of Experimental Psychology*, 89, 46–55.

—— —— (1973). Response-Induced Reversals of Preference in Gambling: An Extended Replication in Las Vegas. *Journal of Experimental Psychology*, 101, 16–20.

List, J. A. (2002). Preference Reversals of a Different Kind: The "More is Less" Phenomenon. *American Economic Review*, 92, 1636–43.

Loomes, G., and Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. *Economic Journal*, 92, 805–24.

Luce, R. D., and Raiffa H. (1957). *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

Marschak, J. (1964). Actual versus Consistent Decision Behavior. *Behavioral Science*, 9, 103–10.

McCaffery, E. J., and Baron, J. (2006). Isolation Effects and the Neglect of Indirect Effects of Fiscal Policies. *Journal of Behavioral Decision Making*, 19, 289–302.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The Adaptive Decision Maker*. New York: Cambridge University Press.

Rabin, M., and Thaler, R. H. (2001). Risk Aversion. *Journal of Economic Perspectives*, 15/1, 219–32.

Read, D. (2007). Time and the Marketplace. *Marketing Theory*, 7, 59–74.

—— Loewenstein, G., and Rabin, M. (1999). Choice Bracketing. *Journal of Risk and Uncertainty*, 19, 171–97.

Redelmeier, D. A., and Tversky, A. (1992). On the Framing of Multiple Prospects. *Psychological Science*, 3, 191–3.

Savage, Leonard J. (1954). *The Foundations of Statistics*. New York: Wiley (2nd edn., New York: Dover, 1972).

Shafer, G. (1986). Savage Revisited. *Statistical Science*, 1, 463–85.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-Based Choice. *Cognition*, 49, 11–36.

Simon, H. (1978). Rationality as Process and as a Product of Thought. *American Economic Review*, 68, 1–16.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, 16, 158–74.

Slovic, P., and Tversky, A. (1974). Who Accepts Savage's Axioms? *Behavioral Science*, 19, 368–73.

Starmer, C., and Sugden, R. (1993). Testing for Juxtaposition and Event-Splitting Effects. *Journal of Risk and Uncertainty*, 6, 235–54.

Strotz, R. H. (1953). Cardinal Utility. *American Economic Review*, 43, 384–97.

Sugden, R. (1991). Rational Choice: A Survey of Contributions from Economics and Psychology. *The Economic Journal*, 101, 751–85.

Trope, Y., and Liberman, N. (2003). Temporal Construal. *Psychological Review*, 110, 403–21.

Tversky, A., and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211, 453–8.

—— Sattath, S., and Slovic, P. (1988). Contingent Weighting in Judgment and Choice. *Psychological Review*, 95, 371–84.

von Neumann, J., and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd edn. Princeton: Princeton University Press.

Wedell, D. H. (1991). Distinguishing among Models of Contextually Induced Preference Reversals. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17, 767–78.

CHAPTER 9

STATE-DEPENDENT UTILITY

EDI KARNI

9.1 Introduction

In January 1971, Robert Aumann wrote a letter to Leonard Savage in which he raised conceptual difficulties with Savage's notion of subjective probabilities. In his letter, Aumann describes a man who loves his wife very much and without whom his life is "less 'worth living' ". The wife falls ill; if she is to survive, she must undergo a routine yet dangerous operation. The husband is offered a choice between betting $100 on his wife's survival or on the outcome of a coin flip. Even if the husband believes that his wife has an even chance of surviving the operation, he may still rather bet on her survival, because winning $100 if she does not survive is "somehow worthless". If he bets on the outcome of a coin flip, he might win but not be able to enjoy his winnings because his wife dies. In this situation, Aumann argues, Savage's notion of states (that is, whether, following the surgery, the wife is dead or alive) and consequences are confounded to the point that there is nothing that one may call a consequence (that is, something whose value is state-independent).

In his response, Savage admits that the difficulties Aumann identifies are indeed serious. In his defense, Savage writes:

> The theory of personal probability and utility is, as I see it, a sort of framework into which I hope to fit a large class of decision problems. In this process, a certain amount of pushing, pulling, and departure from common sense may be acceptable and even advisable. . . . To some—perhaps to you—it will seem grotesque if I say that I should not mind being hung

> so long as it be done without damage to my health or reputation, but I think it desirable to adopt such language so that the danger of being hung can be contemplated in this framework. (Cited in Drèze 1987, p. 78)

To the specific example cited by Aumann, Savage responds: "In particular, I can contemplate the possibility that the lady dies medically and yet is restored in good health to her husband" (Drèze 1987, p. 80).

The need to invoke scenarios such as those depicted by Savage may be useful if the issue at hand is purely formal. To presume that real people engage in such bizarre mental exercises in order to *clarify* in their own minds the relevant considerations involved in their decisions is utterly farfetched.

Consider a decision-maker who must decide whether or not to purchase disability insurance. The uncertainties he faces include whether and when, during the period covered by the insurance policy, he might become disabled and, if he does, the nature of the disability and its associated financial consequences. Against this backdrop, the decision-maker must evaluate alternative disability insurance policies (their premiums and corresponding indemnities). It is natural to suppose that the evaluation process entails an assessment of the likelihood of staying healthy, the likelihoods of various potential disabilities, and the valuation of the corresponding financial needs. It is also natural to think that the decision-maker's financial needs depend on his state of health.

A second example is the choice of life insurance policy. In this instance, the focus is on the uncertainty regarding the decision-maker's longevity, his lifetime consumption, and the size of his estate at the time of his death. A life insurance policy is an instrument that allows the decision-maker to assure his beneficiaries a certain estate value in case of untimely death. It is quite conceivable that deciding the merits of alternative life insurance policies involves, in addition to the assessment of the specific mortality risks, the evaluation of the consumption–bequest plans whose utilities are intrinsically dependent on the time of the decision-maker's death.

A third example concerns unemployment risk. Presumably, the welfare implications of alternative unemployment insurance programs depend on the impact of loss of income due to unemployment on the individual affected. The assessment of this impact depends, in turn, on the unemployed "utility of income" in the unemployment and employment states.

Even if the practice of quantifying the subjective valuations of goods by a utility function is applicable, the examples of disability, life, and unemployment insurance suggest that the utility of consumption in general, and wealth or income in particular, depend on the decision-maker's physical and mental health, on whether he is dead or alive, and on his employment status. These are some examples of a class of problems in which the dependence of the decision-maker's preferences on the underlying state constitutes an indispensable feature of the decision problem.

Other problems of the same general nature include the choice of health insurance coverage (see Arrow 1974; Karni 1985), the choice of flight insurance coverage (see Eisner and Strotz 1961), and the provision of collective protection (see Cook and Graham 1977). These problems do not fit naturally into Savage's framework and require a different treatment.

This chapter reviews the analytical issues raised by the state-dependent evaluation of consequences and the attempts to address them. An earlier review of these issues is provided in Drèze and Rustichini (2004).

9.2 Subjective Expected Utility Theory

During the last half-century, subjective expected utility theory became the dominant model of rational decision-making under uncertainty and the main analytical tool used in the economic analysis of institutions and schemes designed to improve the allocation of risk bearing. In this section I present a critical review of subjective expected utility theory, arguing that, contrary to common practice, state-independent preferences do not entail state-independent utility functions.

9.2.1 The Analytical Framework

Decision-making under uncertainty involves choosing among alternative courses of action whose consequences are not unique. To formalize the idea of uncertainty, Savage (1954) introduced an analytical framework consisting of a set S, whose elements are *states of nature* (or states, for brevity); an arbitrary set C, of *consequences*; and the set F, of *acts* (functions from the set of states to the set of consequences). Acts represent courses of action, consequences describe anything that may happen to the person, and states are resolutions of uncertainty, that is, "a description of the world so complete that, if true and known, the consequences of every action would be known" (Arrow 1971, p. 45). Implicit in this definition is the notion that there is a unique true state. Subsets of S are referred to as *events*. An event is said to obtain if the true state belongs to it.

Decision-makers are characterized by binary relations on F, referred to as preference relations. Let $\succcurlyeq$ denote a preference relation, then $f \succcurlyeq f'$ means that the course of action f is at least as desirable as the course of action f'. The strict preference relation $\succ$ is defined as $f \succ f'$ if $f \succcurlyeq f'$ and not $f' \succcurlyeq f$. It has the following interpretation: given a choice between the courses of action f and f', the decision-maker will choose the course of action f. The indifference relation $\sim$ is defined by $f \sim f'$ if $f \succcurlyeq f'$ and $f' \succcurlyeq f$ and has the usual interpretation.

9.2.2 The Preference Structure

A course of action is evaluated by its potential consequences and their corresponding likelihoods. Subjective expected utility theory postulates a preference structure that permits the expression of the decision-maker's valuation of the consequences by a utility function, his assessment of their likelihoods by a subjective probability measure on the set of events, and the evaluation of acts by the mathematical expectations of the utility with respect to the subjective probability. The theory admits distinct formulations based on alternative specifications of the sets of states and consequences. In the original formulation of Savage (1954), the set of consequences is arbitrary, and the set of states infinite. In the model of Anscombe and Aumann (1963), the consequences are lotteries, and the set of states is finite. Because of the transparency it affords, I present next the model of Anscombe and Aumann.

Let X be a finite set of arbitrary prizes, or outcomes, and denote by $\mathcal{L}(X)$ the set of consequences that, in this model, are probability distributions on X. Elements of $\mathcal{L}(X)$ are referred to as *roulette lotteries*. Acts, in this formulation, are functions from a finite set of states, S, to $\mathcal{L}(X)$. Let F denote the set of all acts. For all f, $f' \in F$, and $\alpha \in [0, 1]$, define the act $\alpha f + (1 - \alpha) f'$ by:

$$\left(\alpha f + (1 - \alpha)\, f'\right)(s) = \alpha f(s) + (1 - \alpha)\, f'(s) \text{ for all } s \in S, \tag{1}$$

where $f_s \in \mathcal{L}(X)$. This definition embodies Anscombe and Aumann's "reversal of order in compound lotteries" axiom. Specifically, the term of the left-hand side of Eq. 1 has the interpretation of a compound lottery in which, prior to discovering the true state, the decision-maker can flip a coin to determine his choice between two acts, and the expression on the right-hand side of Eq. 1 has the interpretation that, after having observed the true state, the decision-maker must flip a coin to determine whether his payoff is the consequence assigned to that state by the first or the second act. The identification of these two terms implies that "if the prize you receive is to be determined by both a horse race and the spin of a roulette wheel, then it is immaterial whether the wheel is spun before or after the race" (Anscombe and Aumann 1963, p. 201).

The preference structure is depicted axiomatically. The first three axioms are those of expected utility theory:

(A1) (Weak order) $\succcurlyeq$ is transitive (i.e. for all f, f', $f'' \in F$, if $f \succcurlyeq f'$ and $f' \succcurlyeq f''$, then $f \succcurlyeq f''$) and complete (i.e. for all f, $f' \in F$, either $f \succcurlyeq f'$ or $f' \succcurlyeq f$).

(A2) (Archimedean) For all f, f', $f'' \in F$, if $f \succ f' \succ f''$, then $\alpha f + (1 - \alpha) f'' \succ f' \succ \beta f + (1 - \beta) f''$ for some $\alpha, \beta \in (0, 1)$.

(A3) (Independence) For all f, f', $f'' \in F$ and $\alpha \in [0, 1]$, if $f \succcurlyeq f'$, then $\alpha f + (1 - \alpha) f'' \succcurlyeq \alpha f' + (1 - \alpha) f''$.

Intuitively speaking, a state is null if the decision-maker believes that it is impossible to obtain. This is supposedly manifested by his indifference among all acts whose consequences are the same in every other state. To define null states formally, it is convenient to use the following notations: Denote by $f_{-s}\,p$ the act whose sth coordinate is the lottery p and $(f_{-s}\,p)(s') = f(s')$ for all $s' \in S - \{s\}$. Then a state s is said to be *null* if $f_{-s}\,p \sim f_{-s}q$, for all $f \in F$ and $p, q \in \mathcal{L}(X)$, otherwise it is *nonnull*.

By the von Neumann–Morgenstern (1947) theorem, a preference relation $\succcurlyeq$ on F satisfies weak order, Archimedean, and independence if and only if there exists a real-valued function w on $X \times S$ such that, for all $f, f' \in F$,

$$f \succcurlyeq f' \;\Leftrightarrow\; \sum_{s\in S}\sum_{x\in X} w(x,s)\, f(x,s) \geq \sum_{s\in S}\sum_{x\in X} w(x,s)\, f'(x,s). \qquad (2)$$

Moreover, the function w is unique up to cardinal unit-comparable positive transformation (i.e. $\hat{w}$ is another real-valued function on $X \times S$ representing $\succcurlyeq$ in the sense of (2) if and only if $\hat{w}(x,s) = bw(x,s) + a_s$, $b > 0$), and it is constant on X if and only if s is null.

The dependence of the value of w on the states reflects, at least in part, the decision-maker's beliefs regarding the likelihood of the states. It may also indicate that the valuation of the prize x is not independent of the state in which it is received. To separate the decision-maker's beliefs from his valuation of the prizes (i.e., to decompose, in a unique way, w into a product $w(x,s) = u(x)\pi(s)$, where u is a real-valued function on X and π is a probability measure on S) Anscombe and Aumann (1963) require that the preference relation be nontrivial and state-independent. Formally,

(A4) (Nontriviality) There exist $f, f' \in F$ such that $f \succ f'$.

For every $s \in S$, let $\succcurlyeq_s$ be the induced preference relation on $\mathcal{L}(X)$ defined by $p \succcurlyeq_s q$ if $f_{-s}\,p \succcurlyeq f_{-s}q$, for all $f \in F$ and $p, q \in \mathcal{L}(X)$.

(A5) (State independence) For all nonnull $s, s' \in S$, $\succcurlyeq_s = \succcurlyeq_{s'}$.

9.2.3 Representation

A theorem due to Anscombe and Aumann (1963) gives necessary and sufficient conditions for the existence of subjective expected utility representation of $\succcurlyeq$.

Theorem 1. A preference relation $\succcurlyeq$ on F satisfies axioms (A1)–(A5) if and only if there exists a nonconstant, real-valued function u on X and a probability measure π on S, such that, for all $f, f' \in F$,

$$f \succcurlyeq f' \;\Leftrightarrow\; \sum_{s\in S}\pi(s)\sum_{x\in X} u(x)\, f(x,s) \geq \sum_{s\in S}\pi(s)\sum_{x\in X} u(x)\, f'(x,s). \qquad (3)$$

Moreover, the function u is unique up to a positive linear transformation, π is unique, and $\pi(s) = 0$ if and only if s is null.

As in the theory of Savage (1954), in Theorem 1 the utilities assigned to consequences are independent of the underlying states, and the probabilities assigned to events are independent of acts. These assignments, however, are not implied by the axiomatic structure. This observation is essential and merits further elaboration.

9.2.4 Discussion

State independence requires that the decision-maker's ordinal ranking of lotteries be the same across all nonnull states. Moreover, because $\mathcal{L}(X)$ includes all the degenerate lotteries δ_x, $x \in X$, where δ_x assigns the unit probability mass to x, (A5) also implies that the ordinal ranking of prizes is independent of the underlying events. If the prizes are monetary, (A5) implies *state-independent risk attitudes*. It does not, however, rule out that the states affect the decision-maker's well-being. In other words, *state-independent preferences do not imply state-independent utility function*. The utility and probability that figure in Theorem 1 are jointly unique—that is, the probability is unique given the utility, and the utility is unique (up to a positive affine transformation) given the probability. It is possible, therefore, to define other probability measures and state-dependent utility functions to obtain an equivalent subjective expected utility representations. To see this, let γ be a strictly positive, real-valued function on S and $\Gamma = \sum_{s \in S} \gamma(s)\pi(s)$. Define $\hat{u}(x, s) = u(x)/\gamma(s)$, for all $x \in X$ and $s \in S$, and $\hat{\pi}(s) = \gamma(s)\pi(s)/\Gamma$, for all $s \in S$. Then, by the representation (3) and the uniqueness of u in Theorem 1, for all f and f' in F,

$$f \succcurlyeq f' \;\Leftrightarrow\; \sum_{s \in S} \hat{\pi}(s) \sum_{x \in X} \hat{u}(x, s)\, f(x, s) \geq \sum_{s \in S} \hat{\pi}(s) \sum_{x \in X} \hat{u}(x, s)\, f'(x, s). \quad (4)$$

Thus the utility–probability pair $(\hat{u}, \hat{\pi})$ induces a subjective expected utility representation of $\succcurlyeq$ that is equivalent to the one induced by the pair (u, π). This shows that the uniqueness of the probability in Theorem 1 is predicated on the convention that the utility function be state-independent, that is, that constant acts be constant utility acts. This convention is not implied by the axioms and, consequently, lacks choice-behavioral implications. Put differently, the state-independent utility function that figures in Theorem 1 is observationally equivalent to the state-dependent utility function in Eq. 4. Hence the validity of the state-independent utility convention is not subject to refutation within the framework of the methodology of revealed preference.

Similarly unsatisfactory is a related aspect of the model: namely, the interpretation of null states. Ideally, a state should be designated as null and be ascribed zero probability if and only if the decision-maker believes it to be impossible to obtain.

By definition, however, a state is null if the decision-maker displays indifference among all acts that assign the same lotteries to all the other states. This definition does not distinguish states that the decision-maker perceives as unobtainable from states in which all lotteries are equally desirable. Consider, for example, a passenger who is indifferent about the size of his estate in the event that he dies. For such a passenger, a plane crash is a null event and is assigned zero probability, even though he believes that a plane crash is a possibility. This problem renders the representation of beliefs by subjective probabilities dependent on the implicit, unverifiable assumption that in every event some outcomes are strictly more desirable than others. If this assumption is not warranted, the procedure may result in a misrepresentations of beliefs.

Imposing state-independent preferences makes sense in some situations. Its imposition as a general characteristic of preferences, however, is rather uncompelling, and may only be justified by the desire to obtain unique subjective probabilities. Unfortunately, as the preceding discussion shows, state independence is not sufficient to ensure that the subjective probabilities thus defined represent the decision-maker's beliefs. Furthermore, it precludes using the model to analyze important decision problems involving life, health, and unemployment risks, to mention but a few examples. It is conceivable that the states that figure in this sort of problem alter the decision-maker's ordinal ranking of the outcomes, his risk attitudes, or both. The loss of a leg, for example, may well reverse a decision-maker's preferences between spending his vacation hiking or listening to music. It may also make him more risk-averse. A decision theory that admits state-dependent preferences is clearly needed.

9.3 Subjective Expected Utility with State-Dependent Utility Functions

To disentangle utility from subjective probabilities, or tastes from beliefs, in a meaningful way, it is necessary to observe the decision-maker's response to a shift in the state probabilities. This type of observation is precluded in Savage's framework. To overcome this difficulty, the literature pursued two distinct approaches to modeling state-dependent preferences and state-dependent utility functions. The first entails abandoning the revealed-preference methodology by considering verbal expressions of preferences over hypothetical alternatives, and is described next. The second presumes that costly actions are available by which decision-makers may affect the likelihoods of events.

9.3.1 Hypothetical Preferences and Subjective Expected Utility Representations of State-Dependent Preferences

A preference relation $\succcurlyeq$ on F is said to display *state dependence* if $\succcurlyeq_s \neq \succcurlyeq_{s'}$ for some nonnull states s and s'. To overcome the problem of the indeterminacy of the subjective probabilities and utilities associated with state-dependent preferences, Karni and Schmeidler (1981) departed from the revealed-preference methodology, postulating instead the existence of a preference relation on hypothetical lotteries whose prizes are outcome–state pairs. Because the hypothetical lotteries imply distinct, hence incompatible, marginal distributions on the state space, preferences among such lotteries are introspective and may be expressed only verbally. For example, a decision-maker who has to choose between watching a football game in an open stadium or staying at home and watching the game on TV is supposed to be able to say how he would have chosen if the weather forecast were 80 percent chance of showers, or 35 percent chance of showers, during the game. Karni and Schmeidler assume that decision-makers are able to conceive of such hypothetical situations and evaluate them by invoking the same mental processes that govern their actual decisions, and the verbal expression of preferences provides information that can be used to determine the probabilities and utilities. Specifically, suppose that the preference relation, $\hat{\succcurlyeq}$, on hypothetical lotteries satisfies the axioms of expected utility and is consistent with the actual preference relation on acts. Then the expected utility representation of the hypothetical preferences yields state-dependent utility functions, $u(\cdot, s)$, and consistency implies that the same utility functions are implicit in the additive representation of the actual preferences. This allows the identification of the subjective probabilities implicit in the valuation functions, $w(\cdot, s)$, in Eq. 2, that is, $\pi(s) = w(x, s)/u(x, s)$, $s \in S$. Moreover, the actual preference relation has the expected utility representation, $\sum_{s\in S} \pi(s) \sum_{x\in X} u(x, s) f(x, s)$, and the preference relation $\hat{\succcurlyeq}$ is represented by $\sum_{s\in S} \sum_{x\in X} u(x, s)\ell(x, s)$, where ℓ denotes a hypothetical outcome–state lottery. The function u is unique up to cardinal unit-comparable transformation, and the probability π restricted to the event of all essential states is unique satisfying $\pi(s) = 0$ if s is obviously null, and $\pi(s) > 0$ if s is obviously nonnull.

Karni and Mongin (2000) observe that if the decision-maker's beliefs constitute an independent cognitive phenomenon, quantifiable by a probability measure, then the subjective probability that figures in the representation above is the numerical expression of these beliefs. Grant and Karni (2004) provide a probabilistically sophisticated version of this approach. (Probabilistic sophistication is a term introduced by Machina and Schmeidler (1992) to describe decision-makers who behave as if they assess the likelihoods of events by subjective probabilities, but whose preferences are not necessarily linear in these probabilities.)

A weaker version of this approach, based on restricting consistency to a subset of hypothetical lotteries that have the same marginal distribution on S, due to Karni, Schmeidler, and Vind (1983), yields a subjective expected utility representation with state-dependent preferences. However, the subjective probabilities in this representation are arbitrary, and the utility functions, while capturing the decision-maker's state-dependent risk attitudes, do not necessarily represent his evaluation of the consequences in the different states. Wakker (1987) extends the theory of Karni, Schmeidler, and Vind to include the case in which the set of consequences is a connected topological space.

Other theories yielding subjective expected utility representations with state-dependent utility functions invoke preferences on conditional acts (i.e. preference relations over the set of acts conditional on events). Fishburn (1973), Drèze and Rustichini (1999), and Karni (2007) advance such theories. Skiadas (1997) proposes a model, based on hypothetical preferences, that yields a representation with state-dependent preferences. In this model, acts and states are primitive concepts, and preferences are defined on act–event pairs. For any such pair, the consequences (utilities) represent the decision-maker's expression of his holistic valuation of the act. The decision-maker is not supposed to know whether the given event occurred; hence his evaluation of the act reflects, in part, his anticipated feelings, such as disappointment aversion.

9.3.2 Subjective Expected Utility with Moral Hazard and State-Dependent Preferences

A different, choice-based approach to modeling expected utility with state-dependent utility functions presumes that decision-makers believe that they possess the means to affect the likelihood of the states. This idea was originally proposed by Drèze (1961, 1987). Departing from Anscombe and Aumann's (1963) "reversal of order in compound lotteries" axiom, Drèze assumes that a decision-maker who strictly prefers that the uncertainty of the lottery be resolved before that of the acts does so because the information allows him to affect the likely realization of the outcome of the underlying states (the outcome of a horse race, for example). The means by which the decision-maker may affect the likelihoods of the events are not an explicit aspect of the model. Drèze's axiomatic structure implies a unique separation of state-dependent utilities from a set of probability distributions over the set of states of nature. Choice is represented as expected utility-maximizing behavior in which the expected utility associated with any given act is itself the maximal expected utility with respect to the probabilities in the set.

Karni (2006*b*) pursues the idea that observing the choices over actions and bets of decision-makers who believe they can affect the likelihood of events by their

actions provides information that reveals their beliefs. Unlike Drèze, Karni treats the actions by which a decision-maker may influence the likelihood of the states as an explicit ingredient of the model. Because Savage's notion of states requires that this likelihood be outside the decision-maker's control, to avoid confusion, Karni uses the term *effects* instead of states to designate phenomena on which decision-makers can place bets and whose realization, they believe, can be influenced by their actions. Like states, effects resolve the uncertainty of bets; unlike states, their likelihood is affected by the decision-maker's choice of action.

Let Θ be a finite set of effects, and denote by A an abstract set whose elements are referred to as *actions*. Actions correspond to initiatives a decision-maker may undertake that he believes affect the likely realization of alternative effects. Let $Z(\theta)$ be a finite set of prizes that are feasible if the effect θ obtains; denote by $L(Z(\theta))$ the set of lotteries on $Z(\theta)$. Bets are analogous to acts and represent effect-contingent lottery payoffs. Formally, a bet, b, is a function on Θ such that $b(\theta) \in L(Z(\theta))$. Denote by B the set of all bets, and suppose that it is a convex set, with a convex operation defined by $(\alpha b + (1-\alpha)b')(\theta) = \alpha b(\theta) + (1-\alpha)b'(\theta)$, for all $b, b' \in B$, $\alpha \in [0,1]$, and $\theta \in \Theta$. The choice set is the product set $\mathbb{C} := A \times B$ whose generic element, (a, b), is an action–bet pair. Action–bet pairs represent conceivable alternatives among which decision-makers may have to choose. The set of *consequences* C consists of prize–effect pairs; that is, $C = \{(z, \theta) \mid z \in Z(\theta), \theta \in \Theta\}$.

Decision-makers are supposed to be able to choose among action–bet pairs—presumably taking into account their beliefs regarding the influence of their choice of actions on the likelihood of alternative effects—and, consequently, on the desirability of the corresponding bets and the intrinsic desirability of the actions. For instance, a decision-maker simultaneously chooses a health insurance policy and an exercise and diet regimen. The insurance policy is a bet on the effects that correspond to the decision-maker's states of health; adopting an exercise and diet regimen is an action intended to increase the likelihood of good states of health. A decision-maker is characterized by a preference relation $\succcurlyeq$ on $\mathbb{C}$.

Bets that, once accepted, render the decision-maker indifferent among all the actions are referred to as constant valuation bets. Such bets entail compensating variations in the decision-maker's well-being due to the direct impact of the actions and the impact of these actions on the likely realization of the different effects and the corresponding payoff of the bet. To formalize this idea, let $I(b; a) = \{b' \in B \mid (a, b') \sim (a, b)\}$ and $I(p; \theta, b, a) = \{q \in L(Z(\theta)) \mid (a, b_{-\theta}q) \sim (a, b_{-\theta}p)\}$. A bet $\bar{b} \in B$ is said to be a *constant valuation bet according to* $\succcurlyeq$ if $(a, \bar{b}) \sim (a', \bar{b})$ for all $a, a' \in \hat{A}$, and $b \in \cap_{a \in \hat{A}} I(\bar{b}; a)$ if and only if $b(\theta) \in I(\bar{b}(\theta); \theta, \bar{b}, a)$ for all $\theta \in \Theta$ and $a \in \hat{A}$. Let B^{cv} denote the subset of constant valuation bets. Given $p \in L(Z(\theta))$, I denote by $\overline{b_{-\theta}p}$ the constant valuation bet whose θth coordinate is p.

An effect $\theta \in \Theta$ is *null given the action* a if $(a, b_{-\theta}p) \sim (a, b_{-\theta}q)$ for all $p, q \in L(Z(\theta))$ and $b \in B$, otherwise it is *nonnull given the action* a. In general, an effect may be null under some actions and nonnull under others. Two effects, θ and θ', are said to be *elementarily linked* if there are actions $a, a' \in A$ such that $\theta, \theta' \in \Theta(a) \cap \Theta(a')$, where $\Theta(a)$ denotes, the subset of effects that are nonnull given a. Two effects are said to be *linked* if there exists a sequence of effects $\theta = \theta_0, \ldots, \theta_n = \theta'$ such that every θ_j is elementarily linked with θ_{j+1}.

The preference relation $\succcurlyeq$ on $\mathbb{C}$ is nontrivial if the induced strict preference relation, $\succ$, is nonempty. Henceforth, assume that the preference relation is nontrivial, every pair of effects is linked, and every action–bet pair has an equivalent constant valuation bet.

For every a, define the conditional preference relation $\succcurlyeq_a$ on B by: $b \succcurlyeq_a b'$ if and only if $(a, b) \succcurlyeq (a, b')$. The next axiom requires that, for every given effect, the ranking of lotteries be independent of the action. In other words, conditional on the effects, the risk attitude displayed by the decision-maker is independent of his actions. Formally,

(A6) (Action-independent risk attitudes) For all $a, a' \in A$, $b \in B$, $\theta \in \Theta(a) \cap \Theta(a')$ and $p, q \in L(Z(\theta))$, $b_{-\theta}p \succcurlyeq_a b_{-\theta}q$ if and only if $b_{-\theta}p \succcurlyeq_{a'} b_{-\theta}q$.

The next theorem, due to Karni (2006), gives necessary and sufficient conditions for the existence of representations of preference relations over the set of action–bet pairs with effect-dependent utility functions and action-dependent subjective probability measures on the set of effects.

Theorem 2. Let $\succcurlyeq$ be a preference relation on $\mathbb{C}$ that is nontrivial, every pair of effects is linked, and every action–bet pair has an equivalent constant valuation bet. Then $\{\succcurlyeq_a \mid a \in A\}$ are weak orders satisfying the Archimedean, independence, and action-independent risk attitudes axioms if and only if there exists a family of probability measures $\{\pi(\cdot; a) \mid a \in A\}$ on Θ; a family of effect-dependent, continuous, utility functions $\{u(\cdot; \theta) : Z(\theta) \to \mathbb{R} \mid \theta \in \Theta\}$; and a continuous function $f : \mathbb{R} \times A \to \mathbb{R}$, increasing in its first argument, such that, for all $(a, b), (a', b') \in \mathbb{C}$,

$$(a, b) \succcurlyeq (a', b')$$

if and only if

$$f\left(\sum_{\theta \in \Theta} \pi(\theta; a) \sum_{z \in Z(\theta)} u(z; \theta)\, b(z; \theta), a\right) \geq f\left(\sum_{\theta \in \Theta} \pi(\theta; a') \sum_{z \in Z(\theta)} u(z; \theta)\, b'(z; \theta), a'\right). \tag{5}$$

Moreover, $\{v(\cdot;\theta) : Z(\theta) \to \mathbb{R} \mid \theta \in \Theta\}$ is another family of utility functions, and g is another continuous function representing the preference relation in the sense of Eq. 5 if and only if, for all $\theta \in \Theta$, $v(\cdot,\theta) = \lambda u(\cdot,\theta) + \varsigma(\theta)$, $\lambda > 0$, and, for all $a \in A$, $g(\lambda x + \varsigma(a), a) = f(x,a)$, where $x \in \{\sum_{\theta\in\Theta} \pi(\theta;a) \sum_{x\in Z(\theta)} u(z;\theta)b(z;\theta) \mid b \in B\}$ and $\varsigma(a) = \sum_{\theta\in\Theta} \varsigma(\theta)\pi(\theta;a)$. The family of probability measures $\{\pi(\cdot;a) \mid a \in A\}$ on Θ is unique satisfying $\pi(\theta;a) = 0$ if and only if θ is null given a.

The function $f(\cdot,a)$ in Eq. 5 represents the direct impact of the action on the decision-maker's well-being. The indirect impact of the actions, due to variations they produce in the likelihood of effects, is captured by the probability measures $\{\pi(\cdot;a)\}_{a\in A}$. However, the uniqueness of utility functions in Eq. 5 is due to a normalization; it is therefore arbitrary in the same sense as the utility function in Theorem 1 is. To rid the model of this last vestige of arbitrariness, Karni (2008) shows that if a decision-maker is Bayesian in the sense that his posterior preference relation is induced by the application of Bayes's rule to the probabilities that figure in that representation of the prior preference relation, then the representation is unique, and the subjective probabilities represent the decision-maker's beliefs.

If a preference relation $\succcurlyeq$ on $\mathbb{C}$ satisfies conditional effect independence (i.e. if $\succcurlyeq_a$ satisfies a condition analogous to (A5), with effects instead of states), then the utility functions that figure in Theorem 2 represent the same risk attitudes and assume the functional form $u(z;\theta) = \sigma(\theta)u(z) + \kappa(\theta)$, $\sigma(\cdot) > 0$. In other words, effect independent risk attitudes do not imply effect-independent utility functions. The utility functions are effect-independent if and only if constant bets are constant utility bets.

9.4 Risk Aversion with State-Dependent Preferences

The *raison d'être* of many economic institutions and practices, such as insurance and financial markets, cost-plus procurement contracts, and labor contracts, is the need to improve the allocation of risk bearing among risk-averse decision-makers. The analysis of these institutions and practices was advanced with the introduction, by de Finetti (1952), Pratt (1964), and Arrow (1971), of measures of risk aversion. These measures were developed for state-independent utility functions, however, and are not readily applicable to the analysis of problems involving state-dependent utility functions such as optimal health or life insurance. Karni (1985) extends the theory of risk aversion to include state-dependent preferences.

9.4.1 The Reference Set and Interpersonal Comparison of Risk Aversion

A central concept in Karni's (1985) theory of risk aversion with state-dependent preferences is the reference set. To formalize this concept, let $\mathcal{B}$ denote the set of real-valued function on S, where $S = \{1, \ldots, n\}$ is a set of states. Elements of $\mathcal{B}$ are referred to as *gambles*. As in the case of state-independent preferences, a state-dependent preference relation on $\mathcal{B}$ is said to display risk aversion if the upper contour sets $\{b \in \mathcal{B} \mid b \succcurlyeq b'\}$, representing the acceptable gambles at b', $b' \in \mathcal{B}$, are convex. It displays risk proclivity if the lower contour set, $\{b \in \mathcal{B} \mid b' \succcurlyeq b\}$, representing the unacceptable gambles at b' are convex. It displays these attitudes in the strict sense if the corresponding sets are strictly convex.

For a given preference relation, the reference set consists of the most preferred gambles among gambles of equal mean. Formally, $\mathcal{B}(c) = \{b \in \mathcal{B} \mid \sum_{s\in S} b(s)\,p(s) = c\}$, and the *reference set corresponding to* $\succcurlyeq$ is defined by $RS = \{b^*(c) \mid c \geq 0\}$, where $b^*(c) \in \mathcal{B}(c)$ and $b^*(c) \succcurlyeq b$ for all $b \in \mathcal{B}(c)$. If $\succcurlyeq$ displays strict risk aversion, then the corresponding utility functions $\{u(\cdot, s)\}_{s\in S}$ are strictly concave, and the reference set RS is well-defined and is characterized by the equality of the marginal utility of money across states (i.e. $u'(b^*(s), s) = u'(b^*(s'), s')$ for all $s, s' \in S$). (Figure 9.1 depicts the reference set for strictly risk-averse preferences in the case $S = \{1, 2\}$.) For such preference relations, it is convenient to depict the reference set as follows: Define $f_s(w) = (u')^{-1}(u'(w, 1), s)$, $s \in S$, $w \in \mathbb{R}$. By definition, f_1 is the identity function, and by the concavity of the utility functions, $\{f_s\}_{s\in S}$ are increasing functions. The reference set is depicted by the function $F\colon \mathbb{R}_+ \to \mathbb{R}^n$ defined by $F(w) = (f_1(w), \ldots, f_n(w))$. If the utility functions are state-independent, the reference set coincides with the subset of constant gambles.

Given a preference relation $\succcurlyeq$ and a gamble b, the *reference equivalence* of b is the element, $b^*(b)$, of the reference set corresponding to $\succcurlyeq$ that is indifferent to b. Let $\bar{b} = \sum_{s\in S} b(s)\,p(s)$; the *risk premium associated with* b, $\rho(b)$, is defined by $\rho(b) = \sum_{s\in S}[\bar{b} - b^*(b)]\,p(s)$. Clearly, if a preference relation displays risk aversion, the risk premium is nonnegative (see Figure 9.1).

Broadly speaking, two preference relations $\succcurlyeq^u$ and $\succcurlyeq^v$ displaying strict risk aversion are comparable if they have the same beliefs and agree on the most preferred gamble among gambles of the same mean. Formally, let p be a probability distribution on S representing the beliefs embodied in the two preference relations. Then $\succcurlyeq^u$ and $\succcurlyeq^v$ are said to be *comparable* if $RS^u = RS^v$. Note that if the utility functions are state-independent, all risk-averse preference relations are comparable.

Let $\rho^u(b)$ and $\rho^v(b)$ denote the risk premiums associated with a preference relation $\succcurlyeq^u$ and $\succcurlyeq^v$, respectively, displaying strict risk aversion. Then $\succcurlyeq^u$ is said to

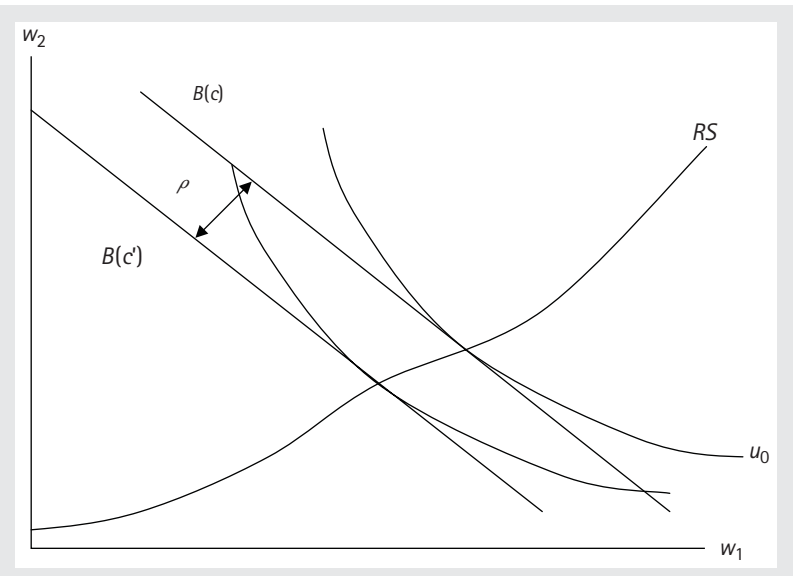

Fig. 9.1. The reference set and risk premium for state-dependent preferences.

display greater risk aversion than $\succcurlyeq^v$ if $\rho^u(b) \geq \rho^v(b)$ for all $b \in \mathcal{B}$. Given $h(\cdot, s)$, $h = u, v$, denote by h_1, h_{11} the first and second partial derivatives with respect to the first argument. The following theorem, due to Karni (1985), gives equivalent characterizations of interpersonal comparisons of risk aversion.

Theorem 3. Let $\succcurlyeq^u$ and $\succcurlyeq^v$ be comparable preference relations displaying strict risk aversion whose corresponding state-dependent utility functions are $\{u(\cdot, s)\}_{s \in s}$ and $\{v(\cdot, s)\}_{s \in s}$. Suppose that u and v are twice continuously differentiable with respect to their first argument. Then the following conditions are equivalent:

(i) $-\dfrac{u_{11}(w, s)}{u_1(w, s)} \geq -\dfrac{v_{11}(w, s)}{v_1(w, s)}$ for all $s \in S$ and $w \in \mathbb{R}$.

(ii) For every probability distribution p on S, there exists a strictly increasing concave function T_p such that $\sum_{s \in S} u(f_s(w), s)p(s) = T_p[\sum_{s \in S} v(f_s(w), s)p(s)]$, and T_p' is independent of p.

(iii) $\rho^u(b) \geq \rho^v(b)$ for all $b \in \mathcal{B}$.

In the case of state-independent preferences, the theory of interpersonal comparisons of risk aversion is readily applicable to the depiction of changing attitudes towards risk displayed by the same preference relation at different wealth levels. In the case of state-dependent preferences, the prerequisite of comparability must be imposed. In other words, the application of the theory of interpersonal comparisons is complicated by the requirement that the preference relations be

comparable. A preference relation, $\succcurlyeq$, displaying strict risk aversion is said to be *autocomparable* if, for any $b^{**}, b^* \in RS$, $N_\varepsilon(b^{**}) \cap RS = (b^{**} - b^*) + N_\varepsilon(b^*) \cap RS$, where $N_\varepsilon(b^{**})$ and $N_\varepsilon(b^*)$ are disjoint neighborhoods in $\mathbb{R}^n$. The reference sets of autocomparable preference relations are depicted by $F(w) = (a_s w)_{s\in S}$, where $a_s > 0$. All preference relations that have expected utility representation with state-independent utility function are obviously autocomparable.

Denote by $\mathbf{x}$ the constant function in $\mathbb{R}^n$ whose value is x. An autocomparable preference relation is said to display *decreasing* (*increasing, constant*) *absolute risk* aversion if $\rho(b) > (<, =)\rho(b + x)$ for every $x > 0$. For autocomparable preference relations with state-dependent utility functions $\{U(\cdot, s)\}_{s\in S}$, equivalent characterizations of decreasing risk aversion are analogous to those in Theorem 3, with $u(w, s) = U(w, s)$ and $v(w, s) = U(w + x, s)$.

9.4.2 Application: Disability Insurance

The following disability insurance scheme illustrates the applicability of the theory of risk aversion with state-dependent preferences. Let the elements of S correspond to potential states of disability (including the state of no disability). Suppose that an insurance company offers disability insurance policies (Π, I) according to the formula $\Pi(I) = \beta\bar{I}$, where I is a positive, real-valued function on S representing the indemnities corresponding to the different states of disability; $\bar{I}$ represents the actuarial value of the insurance policy; Π is the insurance premium corresponding to I; and $\beta \geq 1$ is the loading factor. The insurance scheme is actuarially fair if $\beta = 1$.

Let p be a probability measure on S representing the relative frequencies of the various disabilities in the population. Consider a risk-averse, expected-utility-maximizing decision-maker whose risk attitudes depend on his state of disability. Let $\widetilde{w} = \{w(s)\}_{s\in S}$ represent the decision-maker's initial wealth corresponding to the different states of disability. The decision-maker's problem may be stated as follows: Choose I^* so as to maximize $\sum_{s\in S} u(w(s) - I(s) - \Pi(I), s)p(s)$ subject to the constraints $\Pi(I) = \beta\bar{I}$ and $I(s) \geq 0$ for all s.

If the insurance is actuarially fair, the optimal distribution of wealth, $\widetilde{w}^* = \{w(s)\}_{s\in S}$ is the element of the reference set whose mean value is $\bar{w} = \sum_{s\in S} w(s)p(s)$. Consequently, the optimal insurance is given by $I^*(s) = w^*(s) - w(s)$, $s \in S$. Thus comparable individuals, and only comparable individuals, choose the same coverage under fair insurance for every given $\widetilde{w}$.

If the insurance is actuarially unfair (that is, $\beta > 1$), the optimal disability insurance requires that the indemnities be equal to the total loss above state-dependent minimum deductibles (see Arrow 1974). In other words, there is a subset T of disability states and $\lambda > 0$ such that $u'(\hat{w}(s), s) = \lambda$ for all $s \in T$ and $u'(w(s), s) < \lambda$ otherwise, and $I^*(s) = \hat{w}(s) - w(s)$ if $s \in T$ and $I^*(s) = 0$ otherwise. The values

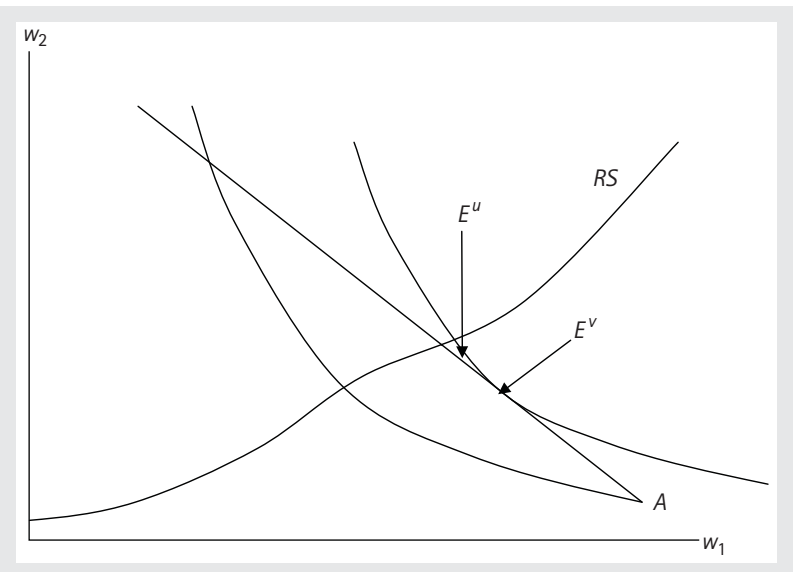

Fig. 9.2. Optimal disability insurance coverage with different degrees of risk aversion.

$\{\hat{w}(s)\}_{s\in T}$ are generalized deductibles. Karni (1985) shows that if $\succcurlyeq^u$ and $\succcurlyeq^v$ are comparable preference relations displaying strict risk aversion in the sense of Theorem 3, then, *ceteris paribus*, if $\succcurlyeq^u$ displays a greater degree of risk aversion than $\succcurlyeq^v$, then $\hat{w}^u(s) \geq \hat{w}^v(s)$ for all $s \in S$, where $\hat{w}^i(s)$, $i \in \{u, v\}$ are the optimal deductibles corresponding to $\succcurlyeq^i$. Thus, *ceteris paribus*, the more risk-averse decision-maker takes out a more comprehensive disability insurance. For the two-states case in which 1 is the state with no disability and 2 is the disability state, the situation is depicted in Figure 9.2. The point A indicates the initial (risky) endowment, and the points E^u and E^v indicate the equilibrium positions of decision-makers whose preference relations are $\succcurlyeq^u$ and $\succcurlyeq^v$, respectively. The preference relation $\succcurlyeq^u$ displays greater risk aversion than $\succcurlyeq^v$ and its equilibrium position, E^u, entails a more comprehensive coverage.

References

Anscombe, F. J., and Aumann, R. J. (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics*, 43, 199–205.

Arrow, K. J. (1971). *Essays in the Theory of Risk Bearing.* Chicago: Markham Publishing Co.

—— (1974). Optimal Insurance and Generalized Deductibles. *Scandinavian Actuarial Journal*, 1–42.

Cook, P. J., and Graham, D. A. (1977). The Demand for Insurance and Protection: The Case of Irreplaceable Commodities. *Quarterly Journal of Economics*, 91, 143–56.

DE FINETTI, B. (1952). Sulla preferibilità. *Giornale degli Economisti e Annali di Economia*, 11, 685–709.

DRÈZE, J. H. (1961). Les fondements logiques de l'utilité cardinale et de la probabilité subjective. La Décision. Colloques Internationaux de CNRS.

—— (1987). Decision Theory with Moral Hazard and State-Dependent Preferences. In *Essays on Economic Decisions under Uncertainty*, 23–89. Cambridge: Cambridge University Press.

—— and RUSTICHINI, A. (1999). Moral Hazard and Conditional Preferences. *Journal of Mathematical Economics*, 31, 159–81.

—— —— (2004). State-Dependent Utility and Decision Theory. In S. Barbera, P. Diamon, and C. Seidl (eds.), *Handbook of Utility Theory*, ii. 839–92. Dordrecht: Kluwer.

EISNER, R., and STROTZ, R. H. (1961). Flight Insurance and the Theory of Choice. *Journal of Political Economy*, 69, 355–68.

FISHBURN, P. C. (1973). A Mixture-Set Axiomatization of Conditional Subjective Expected Utility. *Econometrica*, 41, 1–25.

GRANT, S., and KARNI, E. (2004). A Theory of Quantifiable Beliefs. *Journal of Mathematical Economics*, 40, 515–46.

KARNI, E. (1985). *Decision Making under Uncertainty: The Case of State-Dependent Preferences.* Cambridge, MA: Harvard University Press.

—— (2006). Subjective Expected Utility Theory without States of the World. *Journal of Mathematical Economics*, 42, 325–42.

—— (2007). A Foundations of Bayesian Theory. *Journal of Economic Theory*, 132, 167–88.

—— (2008). A Theory of Bayesian Decision Making. Unpublished MS.

—— and MONGIN, P. (2000). On the Determination of Subjective Probability by Choice. *Management Science*, 46, 233–48.

—— and SCHMEIDLER, D. (1981). An Expected Utility Theory for State-Dependent Preferences. Working Paper 48–80, Foerder Institute for Economic Research, Tel Aviv University.

—— —— and VIND, K. (1983). On State-Dependent Preferences and Subjective Probabilities. *Econometrica*, 51, 1021–31.

MACHINA, M. J., and SCHMEIDLER, D. (1992). A More Robust Definition of Subjective Probability. *Econometrica*, 60, 745–80.

PRATT, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica*, 32, 122–36.

SAVAGE, L. J. (1954). *The Foundations of Statistics.* New York: John Wiley.

SKIADAS, C. (1997). Subjective Probability under Additive Aggregation of Conditional Preferences. *Journal of Economic Theory*, 76, 242–71.

VON NEUMANN, J., and MORGENSTERN, O. (1947). *Theory of Games and Economic Behavior*, 2nd edn. Princeton: Princeton University Press.

WAKKER, P. P. (1987). Subjective Probabilities for State-Dependent Continuous Utility. *Mathematical Social Sciences*, 14, 289–98.

CHAPTER 10

CHOICE OVER TIME

PAOLA MANZINI
MARCO MARIOTTI

10.1 Introduction

Many economic decisions have a time dimension—hence the need to describe how outcomes available at future dates are evaluated by individual agents. The history of the search for a "rational" model of preferences over (and choices between) dated outcomes bears some interesting resemblances and dissimilarities to the corresponding search in the field of risky outcomes. First, a standard and widely accepted model was settled upon. This is the exponential discounting model (EDM) (Samuelson 1937), for which the utility from a future prospect is equal to the present discounted value of the utility of the prospect. That is, an outcome x available at time t is evaluated now, at time $t = 0$, as $\delta^t u(x)$, with δ a constant discount factor and u an (undated) utility function on outcomes. So, according to the EDM, x at time t is preferred now to y at time s if

$$\delta^t u(x) > \delta^s u(y).$$

We wish to thank Steffen Andersen, Glenn Harrison, Michele Lombardi, Efe Ok, Andreas Ortmann, and Daniel Read for useful comments and guidance to the literature. We are also grateful to the ESRC for their financial support through grant n. RES000221636. Any error is our own.

Similarly, a sequence of timed outcomes $x_1, x_2, \ldots x_T$ is preferred to another sequence $y_1, y_2, \ldots, y_T$ if

$$\sum_{i=1}^{T} \delta^{i-1} u(x_i) > \sum_{i=1}^{T} \delta^{i-1} u(y_i).$$

Subsequently, an increasing number of systematic "anomalies" were demonstrated in experimental settings. This spurred the formulation of more descriptively adequate "non-exponential" models of time preferences.

This mirrors the events for the standard model of decision under risk, the expected utility model, in which case observed experimental anomalies led to the formulation of non-expected utility models. However, unlike the case of choice between risky outcomes, for choice over time no normative axioms of "rationality" were formulated which had the same force as, say, the von Neumann–Morgenstern independence axiom of utility theory. Perhaps for this reason, economists have been readier to accept one specific alternative model, that of hyperbolic discounting.

In this chapter we review both the theoretical modeling and the experimental evidence relating to choice over time. Most of the space is devoted to choices between outcome–date pairs, which have been better studied, especially experimentally, but in Section 10.4 we also discuss choices between time sequences of outcomes. In the next section we examine the axiomatic foundation for models based on discounting, exponential or otherwise. In Section 10.3 we review the "new breed" of models that has emerged as a response to experimental observations. Section 10.5 looks in more detail at the empirical evidence, while Section 10.6 is devoted to evaluating the explanatory power of the various theories. Section 10.7 concludes.

10.2 Axiomatics of Exponential Discounting for Outcome–Date Pairs

We begin by describing a basic axiomatization of exponential discounting for outcome–date pairs due to Fishburn and Rubinstein (1982). This will help us in giving a sense of the types of EDM violations that one may expect to observe in practice.

We should make clear at the outset that we follow the standard economic approach of taking *preferences* (*as revealed by binary choices*), as the primitives of the analysis. Any "utility" emerging from the analysis will simply describe the primitive preferences in a numerical form. We are not, therefore, considering "experience" utility (i.e. the psychological benefit one gets from experience) as a primitive, an approach which is more typical in the psychology literature. Also, we focus on time

preferences as if the agent can *commit* to them: this is in order to avoid a discussion of the thorny issue of time consistency,[1] which would deserve a treatment on its own.

Let $X \subseteq \mathcal{R}_+$, with $0 \in X$, represent the set of possible outcomes (interpreted as gains, with 0 representing the status quo), and denote by $T \subseteq \mathcal{R}_+$ the set of times at which an outcome can occur (with $t = 0 \in T$ representing the "present"). Unless specified, T can be either an interval or a discrete set of consecutive dates.

A time-dependent outcome is denoted as (x, t): this is a promise, with no risk attached, to receive outcome $x \in X$ at date $t \in T$. Let $\succcurlyeq$ be a preference ordering on $X \times T$. The interpretation is that $\succcurlyeq$ is the preference expressed by an agent who deliberates in the present about the promised receipts of certain benefits at certain future dates.

As usual, let $\succ$ and $\sim$ represent the symmetric and asymmetric components, respectively, of $\succcurlyeq$. Fishburn and Rubinstein's (1982) characterization uses the following axioms:[2]

Order: $\succcurlyeq$ is reflexive, complete, and transitive.

Monotonicity: If $x > y$, then $(x, t) \succ (y, t)$.

Continuity: $\{(x, t) : (x, t) \succcurlyeq (y, s)\}$ and $\{(x, t) : (y, s) \succcurlyeq (x, t)\}$ are closed sets.

Impatience: Let $s < t$. If $x > 0$, then $(x, s) \succ (x, t)$, and if $x = 0$ then $(x, s) \sim (x, t)$.

Stationarity: If $(x, t) \sim (y, t + t')$, then $(x, s) \sim (y, s + t')$, for all $s, t \in T$ and $t' \in \mathcal{R}$ such that $s + t', t + t' \in T$.

The first four axioms alone guarantee that preferences can be represented by a real-valued "utility" function u on $X \times T$ with the natural continuity and monotonicity property (i.e. u is increasing in x and decreasing in t, and it is continuous in both arguments when T is an interval). The addition of stationarity allows the following restrictions:

Theorem 1 (Fishburn and Rubinstein 1982). If Order, Monotonicity, Continuity, Impatience, and Stationarity hold, then, given any $\delta \in (0, 1)$, there exists a continuous and increasing real-valued function u on X such that

$$(x, t) \succcurlyeq (y, s) \Leftrightarrow \delta^t u(x) \geq \delta^s u(t)$$

In addition, $u(0) = 0$, and if X is an interval, then u is unique (for a given δ) up to multiplication by a positive constant.

[1] Initiated by Strotz (1956).

[2] Fishburn and Rubinstein (1982) consider the general case where the outcome can involve a loss as well as a gain, i.e. $x < 0$, and they do not require that $0 \in X$. Here we focus on the special case only to simplify the exposition.

The representation coincides formally with exponential discounting, but note well the wording of the statement. One may fix the "discount factor" δ arbitrarily to represent a given preference relation that satisfies the axioms, provided the "utility function" u is calibrated accordingly. In other words, for any two discount factors δ and δ', there exist two utility functions u and v such that (u, δ) preferences in the representation of Theorem 1 are identical to (v, δ') preferences in the same type of representation. In order to interpret δ as a uniquely determined parameter expressing "impatience", one would need an *external* method to fix u. This is an important observation, often neglected in applications, which naturally raises the question about what then exactly is *impatience* here. Benoit and Ok (2007) deal with this question by proposing a natural method to compare the delay aversions of time preferences, analogous to methods to compare the risk aversion of preferences over lotteries. As they show, in the EDM it is possible that the delay aversion of a preference represented by (u, δ) is greater than that represented by (v, δ') even though $\delta > \delta'$.

Moreover, given the uniqueness of u only up to multiplication by constants, and the positivity of u for positive outcomes, an additive representation (at least for strictly positive outcomes) is as good as the exponential discounting representation. That is, taking logs and rescaling utilities by dividing by $-\log \delta$, one could write instead

$$(x, t) \succcurlyeq (y, s) \Leftrightarrow u(x) - t \geq u(y) - s.$$

Coming back to the axioms, Continuity is a standard technical axiom. Order is a rationality property deeply rooted in the economic theory of choice. Cyclical preferences, for example, are traditionally banned from economic models. Monotonicity and impatience are also universally assumed in economic models, which are populated by agents for whom more of a good thing is better, and especially for whom a good thing is better if it comes sooner: certainly these are reasonable assumptions in several contexts, though, as we shall see, not in others.

Stationarity, however, does not appear to have a very strong justification, either from the normative or from the positive viewpoint. So it should not be too surprising to observe violations of this axiom in practice, and in fact, as we shall see later in some detail, plenty of them have been recorded. What is surprising, rather, is the willingness of economists to have relied unquestionably for so many years on a model, the EDM, which takes stationarity for granted. Indeed Fishburn and Rubinstein themselves explicitly state that "we know of no persuasive argument for stationarity as a psychologically viable assumption" (1982, p. 681). This led them to consider alternative separable representations that do not rely on stationarity. One assumption (which is popular in the theory of measurement) is the following:

Thomsen separability: If $(x, t) \sim (y, s)$ and $(y, r) \sim (z, t)$, then $(x, r) \sim (z, s)$.

This allows a different representation result:

Theorem 2 (Fishburn and Rubinstein 1982). If Order, Monotonicity, Continuity, Impatience, and Thomsen separability hold, and X is an interval, then there are continuous real-valued functions u on X and δ on T such that

$$(x, t) \succcurlyeq (y, s) \Leftrightarrow \delta(t)u(x) \geq \delta(s)u(y).$$

In addition, $u(0) = 0$ and u is increasing, while δ is decreasing and positive.

This is therefore an axiomatization of a discounting model, in which the discount factor is not constant. However, while Thomsen's separability is logically much weaker than stationarity and it is useful to gauge the additional strength needed to obtain a constant discount factor, one may wonder how intuitive or reasonable a condition it is itself. One might not implausibly argue, for example, that if exactly $(y - x)$ is needed to compensate for the delay of $(s - t)$ in receiving x, and if exactly $(z - y)$ is needed to compensate for the delay of $(t - r)$ in receiving y, then exactly $(z - x)$ is needed to compensate for the delay of $(s - r)$ in receiving x. This argument does not seem to us introspectively much more cogent than stationarity,[3] though it permits the elegant and flexible representation of Theorem 2.

It should be clear from the above results and discussion that the EDM for outcome–date pairs is best justified on the basis of its simplicity and usefulness in applications. Violations especially of the stationarity aspect of it are to be expected, and while they have captured most of the attention, it is perhaps violations of other properties, such as Order, which would appear to be more intriguing, striking as they do more directly at the core of traditional thinking about economic rationality.

10.3 Recent Models for Outcome–Date Pairs

10.3.1 Hyperbolic Discounting

As we mentioned already, over the past twenty years or so a body of empirical evidence has emerged documenting that actual behavior consistently and systematically contradicts the predictions of the standard model. As we discuss more fully in Section 10.5, various exponential discounting "anomalies" have been identified.[4] As we explain further in Section 10.6, in a sense some of these are not anomalies

[3] Fishburn and Rubinstein (1982) also provide a different argument for Thomsen separability, based on an independence condition when the domain of outcomes is enriched to include gambles.

[4] For a survey of these violations, see Loewenstein and Thaler (1989) or Loewenstein and Prelec (1992); for a thorough treatement of issues concerning choice over time, see Elster and Loewenstein (1992).

at all: they do not violate any of the axioms in the theorems above, but only make specific demands on the shape of the utility function. Among those that do violate the axioms in the representations, one particular effect has captured the limelight: preferences are rarely stationary, and people often exhibit a strict preference for "immediacy". Decision-makers may be indifferent between some immediate outcome and a delayed one, but in case they are both brought forward in time, the formerly immediate outcome loses completely its attractiveness. More formally, if x and y are two possible outcomes, situations of the type described above can be summarized as

$$(x, 0) \succ (y, t) \quad \text{and} \quad (y, t+\tau) \succ (x, \tau)$$

Note that this violates jointly four of the five axioms in the characterization of Theorem 1, with the exception of Impatience. Let $x' \neq x$ be such that $(x', 0) \sim (y, t)$ (such an x' exists by Continuity). It must be that $x' < x$ (for otherwise if $x' > x$, then by Monotonicity $(x', 0) \succ (x, 0) \succ (y, t)$, and by Order $(x', 0) \succ (y, t)$). By Stationarity $(x', \tau) \sim (y, t+\tau)$. Then by Monotonicity again $(x, \tau) \succ (y, t+\tau)$, a contradiction with the observed preference.

However, this is commonly interpreted as a straight violation of Stationarity, since the latter is sometimes defined in terms of strict preference as well as indifference. It is, however, compatible with the weaker requirement of Thomsen separability.

As a matter of fact, many researchers observing these phenomena do not pay attention to any axiomatic system at all, preferring rather to concentrate directly on the EDM representation itself (sometimes implicitly assuming a linear utility). In the EDM representation the displayed preferences are written as

$$u(x) > \delta^t u(y) \quad \text{and} \quad \delta^\tau u(x) < \delta^{t+\tau} u(y),$$

which is impossible for any utility function u and fixed δ.

This present time bias (immediacy effect) is a special case of what is known as *preference reversal* (or sometimes "common ratio effect" in analogy with expected utility anomalies in the theory of choice under risk), expressed by the pattern:

$$(x, t) \succ (y, s) \quad \text{and} \quad (y, t+\tau) \succ (x, s+\tau).$$

Strictly speaking, as the agent is expressing preferences at one point in time (the present), nothing is really "reversed": the agent simply expresses preferences over different objects, and these preferences happen not to be constrained by the property of stationarity. The reason for the "reversal" terminology betrays the fact that often, especially in the evaluation of empirical evidence, it is implicitly assumed that there is a coincidence between the *current* preferences over future receipts (so far denoted $\succcurlyeq$) and the *future* preferences over the same receipts to be obtained at the same dates. In other words, now dating preferences explicitly, $(x, t) \succcurlyeq_0 (y, s)$ is assumed to be equivalent to $(x, t) \succcurlyeq_\tau (y, s)$, where $\succcurlyeq_\tau$ with $\tau \leq s, t$ is the

preference at date τ. If today you prefer one apple in one year to two apples in one year and one day, in one year you also prefer one apple immediately to two apples the day after. It is far from clear that this is a good assumption. In this way, the displayed observed pattern can be taken as a "reversal" of preferences during the passage of time from now to date τ. Whether this is a justified interpretation or not, the displayed pattern does contradict the EDM. But this is a somewhat "soft" anomaly, in the sense that it does not contradict basic tenets of economic theory, and it can be addressed simply by changes in the functional form of the objective function which agents are supposed to maximize. Notably, it can be explained by the now popular model of *hyperbolic discounting* (HDM)[5] (as well as by other models). In the HDM it is assumed that the discount factor is a hyperbolic function of time.[6] In its general form, $\delta : T \to \mathcal{R}$ is given as

$$\delta(t) = (1 + at)^{-\frac{b}{a}} \quad \text{with} \quad a, b > 0.$$

In the continuous time case, in the limit as a approaches zero, the model approaches the EDM, that is

$$\lim_{a \to 0} (1 + at)^{-\frac{b}{a}} = e^{-bt}.$$

For any given b (which can be interpreted as the discount rate), a determines the departure of the discounting function from constant discounting and is inversely proportional to the curvature of the hyperbolic function.

Hyperbolic discount functions imply that discount rates decrease over time. The hyperbolic functional form captures in an analytically convenient way the idea that the rate of time preference between alternatives is not constant but varies, and in particular decreases as delay increases. So people are assumed to be more impatient for tradeoffs (between money and delay) near the present than for the same tradeoffs pushed further away in time. It can account for preference reversals.

This model fits in the representation of Theorem 2 in Section 10.2. Preference reversal can easily be reconciled within an extension of the EDM, in which the requirement of stationarity has been weakened to Thomsen separability.

The present time bias can be captured even more simply in the most widely used form of declining discount model, the *quasi-hyperbolic* model or "(β, δ) model". In it, the rate of time preference between a present alternative and one available in the next period is $\beta\delta$, whereas the rate of time preference between two consecutive future alternatives is δ. Therefore (x, t) is evaluated now as $u(x)$ if $t = 0$ and as

[5] e.g. Phelps and Pollack (1968); Loewenstein and Prelec (1992); Laibson (1997); and Frederick, Loewenstein, and O'Donoghue (2002).

[6] For documentation of behavior compatible with this functional form, see e.g. Ainslie (1975); Benzion, Rapoport, and Yagil (1989); Laibson (1997); Loewenstein and Prelec (1992); and Thaler (1981). It is important to stress that Harrison and Lau (2005) have argued against the reliability of the elicitation methods used to obtain this empirical evidence. They argue that this evidence is a direct product of the lack of control for credibility in experimental settings with delayed payment.

$\beta\delta^t u(x)$ if $t > 0$, where $\beta \in (0, 1]$ (the case of $\beta = 1$ corresponds to exponential discounting). So we may have

$$u(x) > \beta\delta u(y) \quad \text{and} \quad \beta\delta^{t+\tau} u(y) > \beta\delta^{\tau} u(x),$$

"rationalizing" the present time bias. As we expand, further below, this same approach can be applied in the case of sequences of outcomes (see Section 10.4).

10.3.2 Relative Discounting

Ok and Masatlioglu (2007) have recently proposed an interesting and challenging axiomatic model which, though retaining a certain notion of discounting, dispenses with the usual idea of evaluating future outcomes in terms of their present value. In their "relative" discounting model (RDM), in other words, it is not possible in general to attribute a certain value to outcome–date pairs (x, t) and state that the outcome–date pair with the higher value is preferred. More precisely, their representation (axiomatized for the case where the set of outcomes X is an open interval) is of the following type: there exists a positive, real-valued, and increasing utility function u on outcomes and a "relative discount" function $\delta : T \times T \to \mathcal{R}$ defined on date *pairs* such that

$$(x, t) \succsim (y, s) \Leftrightarrow u(x) \geq \delta(s, t)u(y).$$

The relative discount function δ is positive, continuous, and decreasing in its first argument for any fixed value of the second argument (with $\delta(\infty, t) = 0$), and $\delta(s, t) = 1/\delta(t, s)$. The model is axiomatized in terms of a set of axioms which includes some weak (but rather involved) separability conditions.

The authors' own interpretation of the preference $(x, t) \succsim (y, s)$ is that "the worth at time t of the utility of y that is to be obtained at time s is strictly less than the worth at time t of the utility of x that is to be obtained at time t". They argue that one of the the main novelties of the RDM is that the comparison between the values of (x, t) and (y, s) is not made in the present but at time t or s. However, it seems hard to tell when a comparison between atemporal utilities is made. When comparing *outcome–date pairs*, and not utilities, it is certainly at time 0 that the agent is making the comparison. So one could as naturally say that the comparison between the utilities $u(x)$ and $u(y)$ is also made at time 0, but instead of discounting the utility of the later outcome by the entire delay with which it is to be received, it is discounted only by a measure of its delay relative to the earlier outcome, whose utility is not discounted at all (psychologically, this corresponds to "projecting" the future into the present, which seems reasonable). While this might appear a little like splitting hairs, the issue might become important if the present agent were allowed to disagree with his later selves on the atemporal evaluation of outcomes—that is, on the function u to be used (in the existing model this disagreement

between current and future selves cannot happen, by an explicit assumption made on preferences). A final, and in our opinion appealing, interpretation of the model is as a threshold model with an additive time-dependent threshold in which the term $\delta(s,t)$ is seen not as a multiplicative relative discount factor but just as a "utility fee" to be incurred for an additional delay. In fact, just as we did for the EDM representation in Section 10.2, here, too, we can apply a logarithmic transformation to obtain a representation of the type

$$(x,t) \succsim (y,s) \Leftrightarrow u(y) \geq u(x) + \delta(s,t).$$

Whatever the interpretation, one virtue of the RDM is that it can explain some "hard" anomalies: notably, particular types of preference intransitivities (although no cycle within a given time t is allowed—contrast this with the "vague time preference" model discussed below). The relative discounting representation includes as special cases both exponential and hyperbolic discounting. Therefore, beside intransitivities, it can also account for every soft anomaly for which the HDM can account. In this sense the model is successful. On the flip side, one might argue that it is almost too general, and many other special cases are also included in it. For example, the subadditive discounting or similarity ideas discussed in the next section can also be formulated in this framework.

A similar model has been studied independently by Scholten and Read (2006), who call it the "discounting by interval" model. Their interpretation, motivation and analysis is quite different, however, from that of Ok and Masatlioglu (2007). In their model, the discount function is defined on intervals of time, which is equivalent to defining it on pairs of dates, as for the RDM. But the authors argue for comparisons between alternatives to be made by means of usual *present values*, for which the later outcome is first discounted to the date of the earlier outcome (using the discount factor which is appropriate for the relevant interval) and then discounted again to the present (using the discount factor which is appropriate for this different interval). So, formally: for $s > t$,

$$\begin{aligned}(x,t) \succsim (y,s) \Leftrightarrow \delta(0,s)u(y) &\geq \delta(0,s)\delta(s,t)u(x) \Leftrightarrow u(y) \\ &\geq \delta(s,t)u(x).\end{aligned}$$

Scholten and Read do not axiomatize their model, but focus on interesting experimental evidence suggesting some possible restrictions of the discounting function.

10.3.3 Similarities and Subadditivity

While not proposing fully-fledged models, contributions by Read (2001) and Rubinstein (2001, 2003) put forth some analytical ideas regarding how to interpret certain types of anomalies. We consider the contributions by these two authors in turn.

10.3.3.1 *Subadditivity*

Read (2001) suggest that a model of *subadditive discounting* might apply. This means that the average discount rate for a period of time might be lower than the rate resulting from compounding the average rates of different sub-periods. Furthermore, he suggests that the finer the partition into sub-periods, the more pronounced this effect should be. Formally, $[0, T]$ is a time period divided into the intervals $[t_0, t_1], \ldots, [t_{k-1}, T]$. Let $\delta_T = \exp^{-r_T T}$ be the average discount factor for the period $[0, T]$ (where r_T is the discount rate for that period), and $\delta_i = \exp^{-r_i T}$ the average discount factor that applies to the sub-period beginning at i (where r_i is the discount rate for that period). Then, if there is subadditivity, for any amount x available at time t_k, and letting u denote an atemporal utility function, we have that

$$u(x)\delta_T > u(x)\delta_0\delta_1 \cdot \ldots \cdot \delta_{k-1}.$$

More abstractly, this general idea could even be defined independently of the existence of an atemporal utility function. Given preferences $\succcurlyeq$ on outcome–date pairs, if

$$(x, t_k) \sim (x_{k-1}, t_{k-1}) \sim \ldots \sim (x_0, 0)$$
$$\text{and} \quad (x, t_k) \sim (x_0', 0),$$

subadditivity could be taken as implying that

$$x_0' > x_0.$$

It is important to note, though, that in the absence of further assumptions on preferences the existence of a separable discount function is not guaranteed. The RDM discussed in the previous section characterizes subadditive discounting by $\delta(t, r) > \delta(t, s)\delta(s, r)$.

This is reminiscent of some empirical evidence for decisions under risk, according to which the total compound subjective probability of an event is higher the higher the number of sub-events into which the event is partitioned (e.g. Tversky and Koehler 1994). Preferences for which discounting is subadditive may not be compatible with hyperbolic discounting; that is, discount rates may be constant or increasing in time, contradicting the HDM, while implying subadditivity. This is precisely the evidence found by Read (2001).

10.3.3.2 *Similarity*

Rubinstein (2001, 2003) argues that similarity judgments may play an important role when making choices over time (or under risk). He also shifts attention to the procedural aspects of decision-making. He suggests that a decision procedure

he originally defined for choices under risk (in Rubinstein 1988) can be adapted to model choices over time, too. Let $\approx_{time}$ and $\approx_{outcome}$ be similarity relations (reflexive and symmetric binary relations) on times and outcomes respectively. So $s \approx_{time} t$ reads "date s is similar to date t" and $x \approx_{outcome} y$ reads "outcome x is similar to outcome y". Rubinstein examines the following procedure to compare any outcome–date pairs (x, t) and (y, s):

Step 1 If $x \geq y$ and $t \leq s$, with at least one strict inequality, then $(x, t) \succ (y, s)$. Otherwise, move to step 2.

Step 2 If $t \approx_{time} s$, $not(x \approx_{outcome} y)$ and $x > y$, then $(x, t) \succ (y, s)$. If $x \approx_{outcome} y$, $not(t \approx_{outcome} s)$ and $t < s$, then $(x, t) \succ (y, s)$.

If neither the premise in step 1 nor the premise in step 2 applies, the procedure is left unspecified. Rubinstein used this idea to show how it serves well to explain some anomalies, some of which run counter to the HDM as well as to the EDM.

Of course, once the broad idea has been accepted, many variations of this procedure seem also plausible. For example Tversky (1969) had suggested a "lexicographic semiorder" procedure according to which agents rely on their ranking of the attributes of an alternative in a lexicographic way when choosing between different alternatives. The first attribute of each alternative is compared. If, and only if, the difference exceeds some fixed threshold value is a choice then made accordingly. Otherwise, the agent compares the second attribute of each alternative, and so on. Yet another procedure reminiscent of Tversky's lexicographic semi-order is described in the next section.[7]

Finally, Rubinstein's (2001) experiments show that precisely the same type of decision situations that create a difficulty for the EDM may also be problematic for the HDM, while they may be easily and convincingly accounted for by similarity-based reasoning. He argues that, in this sense, the change to hyperbolic discounting is not radical enough.

10.3.4 Vague Time Preferences

Manzini and Mariotti (2006) introduce the notion of "vague time preferences" as an application of their general two-stage model of decision-making.[8] The starting consideration is that the perception of events distant in time is in general "blurred". Even when a decision-maker is able to choose between, say, an amount x of money now and an amount y of money at time t, it may be more difficult to compare the

[7] Kahneman and Tversky (1979), too, discuss the intransitivities possibly resulting from the "editing" phase of prospect theory, in which small differences between gambles may be ignored.

[8] See Manzini and Mariotti (2007).

same type of alternatives once these are both distant in time. This difficulty in comparing alternatives available in the future may blur the differences between them in the decision-maker's perception. In other words, the passage of time weakens not only the perception of the alternatives (which are perceived, in Pigou's famous phrase,[9] "on a diminished scale" because of the defectiveness of our "telescopic faculty"), but the very *ability to compare* alternatives with one another.

In the "vague" time preferences model, the central point is that the evaluation of a time-dependent alternative is made up of two main components: the pure time preference (it is better for an alternative to be available sooner rather than later, and there exists a limited ability to trade off outcome for time), and *vagueness*: when comparing different alternatives, the further away they are in time, the more difficult it is to distinguish between them.

For (x, t) to be preferred to (y, s) on the basis of a time–outcome tradeoff, the utility of x may exceed the utility of y *by an amount which is large enough so that the individual can tell the two utilities* apart. The amount by which utilities must differ in order for the decision-maker to perceive the two alternatives as distinct is measured by a the positive vagueness function σ, a real-valued function on outcomes. When the utilities differ by more than σ, then we say that the decision-maker prefers the alternative yielding the larger utility *by the primary criterion*. Formally the primary criterion consists of a possibly incomplete preference relation on outcome–date pairs, represented by an interval order as follows:

$$(x, t) \succ (y, s) \Leftrightarrow u(x, t) > u(y, s) + \sigma(y, s),$$

where u is monotonic, increasing in outcomes and decreasing in time. When neither alternative yields a sufficiently high utility, the decision-maker is assumed to resort to some additional heuristic in order to make his choice (*secondary criterion*). Since each alternative has a time and an outcome component, two natural heuristics are distinguished. In the "outcome prominence" version, the decision-maker will first try to base his choice on which of the two available ones is the greater outcome; and only if this comparison is not decisive will he resolve his choice by selecting the earlier alternative. On the contrary, in the "time prominence" version of the model, the decision-maker first compares the two alternatives by the time dimension. If one comes earlier, then that is his choice; otherwise he looks at the other dimension, the outcome, and selects on the basis of which is higher.

Formally, let $\succ$ be defined as in the display above, and let $a \sim b$ if and only if neither $a \succ b$ nor $b \succ a$. Assume that P and I are the asymmetric and symmetric parts, respectively, of a complete order on the set of pure outcomes X. Finally, let $\succcurlyeq^*$ (with $\succ^*$ and $\sim^*$ the corresponding symmetric and asymmetric parts, respectively) denote a complete preference relation (not necessarily transitive) on the set

[9] See Pigou (1920), p. 25.

of alternatives (i.e. outcome–date pairs) $X \times T$, and let $i = (x_i, t_i) \in X \times T$ for $i \in \{a, b\}$. Then the two alternative models are as follows:

Outcome Prominence Model (OPM):

1. $a \succ^* b \Leftrightarrow$
   (a) $a \succ b$ (primary criterion), or
   (b) $(a \sim b, x_a P x_b)$ or $(a \sim b, x_a I x_b, t_a < t_b)$ (secondary criterion)
2. $a \sim^* b \Leftrightarrow (a \sim b, x_a I x_b, t_a = t_b)$.

Time Prominence Model (TPM):

1′. $a \succ^* b \Leftrightarrow$
   (a) $a \succ b$ (primary criterion), or
   (b) $(a \sim b, t_a < t_b)$ or $(a \sim b, t_a = t_b, x_a P x_b)$ (secondary criterion)
2′. $a \sim^* b \Leftrightarrow (a \sim b, x_a I x_b, t_a = t_b)$

In its simplest specification, the (σ, δ) model, there are just two parameters, with δ taken as the individual's discount factor (which embodies the "pure time preference" component of preference), σ a positive constant measuring the individual's vagueness, and u assumed linear in outcome.

10.4 Preferences over Sequences of Outcomes

When it comes to sequences of outcomes available at given times, the standard exponential discounting model still widely used is that introduced by Samuelson (1937), whereby sequence $((x_1, t_1), (x_2, t_2), \ldots, (x_T, T))$ is preferred to sequence $((y_1, t_1), (y_2, t_2), \ldots, (y_T, T))$ whenever the present discounted utility of the former is greater than the present discounted utility of the latter:

$$\sum_{t=1}^{T} u(x_t)\delta^{t-1} > \sum_{t=1}^{T} u(y_t)\delta^{t-1}.$$

As in the case of outcome–date pairs, Loewenstein and Prelec (1992) highlighted that there exist a number of anomalies which cannot be accommodated within the standard framework. We will discuss these anomalies in greater detail in Section 10.5, while here we limit ourselves to presenting the functional form that Loewenstein and Prelec (1992) introduce to account for these phenomena. They propose that the utility of some sequence $x = ((x_1, t_1), (x_2, t_2), \ldots, (x_T, T))$ should

be represented by

$$U(x) = \sum_{i=1}^{T} v(x_i)\delta(t_i),$$

where δ is a discount function assumed to be a generalized hyperbola, $\delta(t) = \frac{1}{(1+at)^{\frac{b}{a}}}$, as in the general case of hyperbolic discounting we saw earlier, and v is a value function on which the following restrictions are imposed:

V1: the value function is steeper in the loss than in the gain domain:

$$v(x) < -v(-x).$$

V2: the value function is more elastic for losses than for gains:

$$\epsilon_v(x) < \epsilon_v(-x) \quad \text{for} \quad x > 0, \quad \text{where} \quad \epsilon_v \equiv \frac{x\partial v(x)}{v(x)}.$$

V3: the value function is more elastic for outcomes that are larger in absolute magnitude:

$$\epsilon_v(x) < \epsilon_v(y) \quad \text{for} \quad 0 < x < y \quad \text{or} \quad y < x < 0.$$

Manzini, Mariotti and Mittone (2006) pursue a different approach, in which, building on Manzini and Mariotti (2006), they postulate a theoretical model which extends the one for outcome–date pairs to sequences. In order to rank monetary reward sequences, the decision-maker looks first at the standard exponential discounting criterion. However, preferences are incomplete, so sequences are only partially ordered by the criterion. Here too they are completed by relying on a secondary criterion. Sequence x is preferred over another sequence y if the discounted utility of x exceeds the discounted utility of y by at least $\sigma(y)$. When sequences cannot be compared by means of discounted utilities, the decision-maker is assumed to focus on one prominent attribute of the sequences. This prominent attribute ranks (maybe partially) the sequences and allows a specific choice to be made. This latter aspect of the model is in the spirit of Tversky, Sattath, and Slovic's (1988) *prominence hypothesis*. The attribute may be context-dependent, so that, for instance, in the outcome–date pairs case, as we saw above, each alternative has two obvious attributes that may become prominent: the date and the outcome.

We stress that, at a fundamental level, the only departure from the standard choice-theoretic approach is that the decision-maker's behavior is described by combining sequentially two possibly incomplete preference orderings, instead of using directly a complete preference ordering. In the case of monetary sequences we use the following representation for preferences. Let $\succ^*$ denote the strict binary preference relation on the set of alternatives (sequences) A, where a typical sequence has the form $i = (i_1, i_2, \ldots, i_T)$. For given u, σ, δ with the usual meaning,

and secondary criterion P_2, then for all $a, b \in A$, we have $a \succ^* b$ if and only if either

1. $\sum_{t=1}^{T} u(a_t)\delta^{t-1} > \sum_{t=1}^{T} u(b_t)\delta^{t-1} + \sigma(b)$,

or

2. $\sum_{t=1}^{T} u(a_t)\delta^{t-1} \leq \sum_{t=1}^{T} u(b_t)\delta^{t-1} + \sigma(b)$,
   $\sum_{t=1}^{T} u(b_t)\delta^{t-1} \leq \sum_{t=1}^{T} u(a_t)\delta^{t-1} + \sigma(a)$, and $a P_2 b$.

The above obviously begs the question of which secondary criterion one should use. This can be suggested by the empirical evidence available, so we postpone examining this issue further, to explore suggestions from data (see Sections 10.5 and 10.6).

We should note, finally, that although positive discounting of some form or other is deeply ingrained in much economic thinking and in virtually all economic policy, the issue of whether this is a justified assumption is open. Fishburn and Edwards (1997) axiomatize, in a discrete time framework, a "discount-neutral" model of preferences over sequences that differ at a finite number of periods. Their general representation takes the following form:

$$a \succcurlyeq b \Leftrightarrow \sum_{\{t:a_t \neq b_t\}} u_t(a_t) \geq \sum_{\{t:a_t \neq b_t\}} u_t(b_t),$$

where the u_t are real-valued functions on an outcome set X_t that may possibly vary with the date. The axioms they use for this model express conditions of order, continuity sensitivity (every period can affect preference), and of course (given the additive form) independence across periods. When it is also assumed that the outcome sets X_t are the same, further separability assumptions of a measure-theoretic nature allow the following specialization of the model:

$$a \succcurlyeq b \Leftrightarrow \sum_{\{t:a_t \neq b_t\}} \delta(t)u(a_t) \geq \sum_{\{t:a_t \neq b_t\}} \delta(t)u(b_t),$$

where $\delta(t)$ is a positive number for any period t. It is not required to be included in the interval $(0, 1)$, and therefore it is consistent with "negative discount rates". Finally, a form of stationarity yields a constant, but possibly negative, discount rate model:

$$a \succcurlyeq b \Leftrightarrow \sum_{\{t:a_t \neq b_t\}} \delta^{t-1}u(a_t) \geq \sum_{\{t:a_t \neq b_t\}} \delta^{t-1}u(b_t),$$

where δ is a uniquely defined positive number.

10.5 Assessing Empirical Evidence

Our starting point has been to underline how some observed patterns of choice are irreconcilable with the standard theoretical model. So far, in assessing the theories, we have taken the empirical evidence at face value. However, a more rigorous assessment of the reliability of the empirical evidence itself is called for.

Indeed, assessing time preferences is a nontrivial matter. A common theme emerging from the huge literature is that their reliable elicitation poses several methodological problems and results in vastly different ranges for discount factor estimates.[10] Although a plethora of studies exist which elicit time preferences, these have hardly proceeded in a highly standardized way. Many confounding factors occur from one study to another, which hamper systematic comparisons to determine to what extent these differences depend on the elicitation methods themselves, as opposed to other differences in experimental design. Moreover, as we shall explain, some recent empirical advances even put into serious question certain results of the "traditional" evidence.

10.5.1 Psychological Effects

To begin with, there are two families of possible psychological effects which act as confounding factors in the evaluations of time preferences: "hypothetical bias" and "affective response". The first term refers to the fact that a substantial proportion of experimental subjects make different choices when answering hypothetical questions as compared with situations where the answer determines the reward of the responder. For instance, it is one thing to ask a subject how much he is prepared to pay for a cleaner environment in the abstract, and quite another to ask the same question as part of a policy document that is going to determine the amount of taxation.[11] Because of this, it would seem reasonable to want to rely on experimental evidence arising from designs which are *incentive-compatible*—that is, such that the respondent's reward for participation depends on the answer he or she has given.

By "affective response" we refer to the emotive states that might be evoked when experimental subjects have to evaluate the delayed receipt of a good or a service, as compared to money. For instance, Loewenstein and Prelec (1993) explain by a "preference for improving sequences" the behavior of a consistent

[10] See e.g. Frederick, Loewenstein and O'Donoghue (2002), table 1.

[11] The literature on whether or not the payment of experimental subjects has an effect on response is huge. See e.g. Plott and Zeiler (2005); Read (2005); Hertwig and Ortmann (2001); Ortmann and Hertwig (2005), to cite just a few. Cummings, Harrison and Rutström (1995) have examined this in the context of the types of dichotomous choices that are asked in time preference elicitation, though in a different domain. Manzini, Mariotti, and Mittone (2006) instead deal with the time domain.

proportion of decision-makers who, after having chosen a fancy French restaurant over a local Greek one, preferring it sooner rather than later, also chose the sequence (Greek dinner in one month and French dinner in two months) over the opposite sequence (French dinner in one month and Greek dinner in two months). This preference for increasingness can be motivated by "savouring": a decision-maker might like to postpone a pleasant activity so as to enjoy the "build-up" to it. As a flip side to this, "dread" would be reduced by anticipating an unpleasant task and reducing the time spent in contemplation of this unsavoury activity.[12] More generally, one can think of a plethora of potential relevant "attributes" of a sequence which might influence choice (see e.g. Read and Powell (2002), who study subjects' stated verbal motivation for their choices). These affective responses do not only involve sequences, of course, e.g. in the case of the choice of the optimal timing of a kiss of your favorite movie star. Goods and services may possess characteristics which make them idiosyncratically attractive or repulsive to respondents, and evoke feelings quite other than pure time preferences.

10.5.2 Soft Anomalies

In addition to the psychological effects mentioned earlier, framing effects may be rather substantial, too. Loewenstein (1988) observed a "delay/speed up" asymmetry, i.e. a difference in the willingness to pay for anticipating receipt of a good and that to postpone it. He showed that when subjects were asked to imagine that they owned a good (a video recorder in the experiment) available in one year, they would be prepared to pay only $54 on average in order to anticipate receipt, and obtain it now. On the other hand, when asked to imagine they actually owned the video recorder, subjects were asking on average a compensation of $126 in order to delay receiving it for a year. Loewenstein interpreted this as a framing effect,[13] a purely psychological phenomenon. He conjectured that if prompted to imagine that he owns a good that is immediately available, when asked how much he would have to be paid to delay receipt of the good, a decision-maker frames the delay as a loss. If, instead, the decision-maker is prompted to imagine that he owns a good available at a later date, when asked how much he would be willing to pay in order to anticipate collection, he would frame this last occurrence as a gain. Note that these types of result were found in both purely hypothetical scenarios as well as in an incentive-compatible one. If, as in prospect theory,[14] losses count for more than gains, then there is an asymmetry in discount rates elicited from the two choice frames. Agents are less willing to anticipate the gain than to postpone a loss; that is, they are more patient for speeding up than for delaying an outcome. As we explain more in detail

[12] An early formal model of these kinds of effects has been proposed by Loewenstein (1987).
[13] See Tversky and Kahneman (1981).
[14] See Kahneman and Tversky (1979).

in Section 10.6, however, these phenomena are not really a violation of standard discounting theorems, as they only impose restrictions on the shape of the utility function.

While the delay/speed-up asymmetry refers to differences in the implied discount rates, depending on the time when the good is available, the so-called magnitude effect refers to differences in the implied discount rates between large and small outcomes. It was first reported by Thaler (1981), who found that, in a hypothetical setting, on average, subjects, were indifferent between receiving \$15 immediately and \$60 in a year, and at the same time indifferent between receiving \$3000 immediately and \$4000 in a year. While the first choice (assuming linear utility) implies a 25 percent discount factor, the second implies a much larger implicit discount factor, of 75 percent. Shelley (1993) carried out a study of both the delay/speed-up asymmetry and the magnitude effect. She carried out a test for the possible combinations of gain, loss, and neutral frames with either a receipt or a payment. For receipts, she found that implied discount rates are higher for small amounts (\$40 and \$200) than for large amounts of money (\$1000 and \$5000), and for speed-up than for delay (time horizons considered were 6 months and 1 year for the small amounts, and 2 and 4 years for the large amounts). From an economist's perspective, the problem is that all these experiments were based on hypothetical choices, without real payments. However, they have been replicated also with real monetary payments (see e.g. Pender 1996). Like the previous anomaly, this can be reconciled with the EDM.

We have already discussed one of the main phenomena that violates one of the axioms (stationarity): namely, preference reversal. Intriguingly, in addition to the "direct" preference reversal we have considered, recently Sayman and Öncüler (2006) have found evidence of what they dub "reverse time inconsistency", whereby subjects who prefer a smaller, earlier reward when both options are in the future switch to the larger, later reward when the smaller option becomes imminent.

Thaler (1981) also observed evidence consistent with *discount rates declining with the time horizon*. That is, subjects were asked questions of the following type: What is the amount of money to be received at dates $t_1, t_2, \ldots, t_K$ that would make you indifferent to receiving x now? The implied discount rates (assuming linear utility) declined as the dates increased (e.g. they were 345 percent over a one-month horizon and 19 percent over a ten-year horizon).[15] There is a certain air of unreality about these values, and we shall say more about this later, when we consider the issue of risk aversion and of field, as opposed to hypothetical experiment, data. However, we emphasize now that even within the realm of experimental observations within an assumed linear utility model, Read (2001) uncovers contrary evidence. Discount rates appear to be constant across three consecutive eight-month periods.

[15] See also Benzion, Rapoport, and Yagil (1989) for an example in the case of hypothetical choices, and Pender (1996) for actual choices.

Rather, his evidence is consistent with subadditive discounting, as discussed in Section 10.3.3.

10.5.3 Sources of Data and Other Elicitation Issues

Since our focus is on the rationality or otherwise of decision-makers, we ought to consider whether it is possible to reconcile economic theory with either experimental evidence arising from experimental designs which are *incentive-compatible* or with empirical evidence from *field data* (which, using real-life choices, automatically avoid any worry about incentive compatibility), and with data that involve only monetary outcomes.

While the discrepancies between observations, and the unrealistic values found, suggest that some problems must be addressed in the elicitation procedures, the point is that paying subjects is in itself not necessarily enough to produce reliable data. What an incentive-compatible elicitation mechanism must do to be dependable is to induce people to reveal (what they perceive to be) their true evaluation of the good in question. Various methods have been used in domains different from time. In fact, the literature on the elicitation of "home-grown values" for all sorts of goods is vast. Traditionally, experimenters *induced* preferences (i.e. valuations of specific goods) in experimental subjects in order to assess the validity or otherwise of a given theoretical model. As the interest has moved towards assessing and eliciting subject preferences in choices among different goods, or in their valuation of some goods, various mechanisms have been introduced to tease out "home-grown" preferences from experimental subjects.

The most popular methods relied upon in the literature on the elicitation of preferences other than time preferences are English auction, second-price auction, and Becker–De Groot–Marschak procedure (BDM). For each of them bidding one's true value is a dominant strategy, and in many experimental settings instructions encourage bidders explicitly to understand and learn the dominant strategy (see e.g. Rutström 1998).

Let us consider them in turn:

- English (or ascending) auction: agents compete for obtaining a good. With the so-called clock implementation of the auction, the price of the good increases steadily over time. As time passes, participants can withdraw. When only one is left, he "wins" the object, and he alone pays the price at which he won.
- Vickrey (i.e. second-price sealed-bid) auction: subjects submit a single bid, secretly from all other participants. The one with the highest bid wins the object, but pays only the second-highest price. This is why it is strategically equivalent to the English auction, since in the latter the winner is the one who stays when the second-highest bidder gives in.

- BDM: this is also equivalent to the two previous auctions, although here bidders play "against" a probability distribution, rather than other subjects. Because of this the BDM procedure has the objectionable difficulty that it introduces a probability dimension to the problem. Subjects have to declare their willingness to pay for a good. Then a price is drawn from a uniform distribution, and if this price is higher than the willingness to pay, the agent gets nothing, whereas if it is lower, the agent pays the price drawn (so for a winning bidder it is as if he put forward the highest bid in a second-price auction, with the price drawn playing the role of the second-highest bid from a fictitious bidder).

All the above are strategically equivalent: so would it make any difference which one is used in the lab?

In auctions with induced preferences (i.e. where subjects are told what their valuation for a good is), Noussair, Roibin, and Ruffieux (2004) find that Vickrey auctions are more reliable in eliciting preferences than the BDM procedure. Again with induced values, Garratt, Walker, and Wooders (2007) find that in comparison with the usual student population, when using experienced eBay bidders as experimental subjects, the difference between over- and under-bidding is no longer significant (while the proportion of agents bidding their value is indistinguishable from standard lab implementation with students).

On the other hand, when preferences are not induced (i.e. they are "home grown"), Rutström (1998) finds that (average) bids are higher in the second-price auction than in either BDM or first-price auctions. Moreover, as noted by Harrison (1992), these elicitation methods suffer from serious incentive properties in the neighborhood of the truth-telling dominant strategy: deviations may be "cheap" enough that experimental subjects do not select the dominant strategy.

Although none of these auction methods has been applied until recently (see below) to time preferences, the systematic discrepancies between alternative methods to elicit preferences for goods suggest that different elicitation methods might also produce different estimates when applied to the time domain. The most relied-upon elicitation technique for time preferences at the moment consists in asking a series of questions, in table format, of the type "Do you prefer: (A) X today or (B) $X + x$ at time T?", where x is some additional monetary amount which increases steadily (from a starting value of zero) as the subject considers the sequence of questions (see Coller and Williams 1999, and Harrison, Lau and Williams 2002). A decision-maker would start switching from selecting A to selecting B from one specific choice onwards, making it possible to infer the discount factor.[16] This table method has been used with additional variations: namely, an additional piece of

[16] To be precise, one can infer only a *range* for the discount factor, whose width depends on the size of the progressive increments of the additional monetary increments x.

information (e.g. giving for each choice the implicit annual discount/interest rate implied by each choice and the prevalent market rate in the real economy) in order to reduce the extent to which subjects anchor their choices to their own experience outside the lab and unknown to the experimenters. Coller and Williams (1999) found discount rates to be much lower than previously found, once this kind of censoring is taken into account.

A very recent experimental study by Manzini, Mariotti, and Mittone (2007) has made the first comparative analysis of the table method, the BDM, and the Vickrey auction in a choice over time setting. Preliminary results show a similarity of elicited values between the latter two methods, but a marked difference between them and the table method.

However, one must be aware that all choice experiments involving questions about money–date pairs reveal discount factors for *money* only in an unequivocal way. It is often implicitly assumed that such experiments also reveal the discount factor for *consumption*, but this interpretation requires the assumption that the money offered in the experiment is consumed immediately: subjects do not use capital markets to reallocate their consumption over time. This assumption is not outrageous (especially for small amounts, it may not be implausible that capital market considerations are ignored), but it certainly cannot be taken for granted without further study. Coller and Williams (1999) were the first to point out the possible censoring effects of capital markets on experimental subjects' responses. Cubitt and Read (2007) explore in great detail what exactly can be inferred from responses to the standard laboratory tasks on choice over time once it is admitted that subjects are able and willing to access imperfect[17] capital markets, so that the implicit laboratory rate of interest competes with market rates. They point out that the choice between two money–date pairs in the presence of capital markets is not really the choice between two points in the standard Fischer diagram,[18] but rather the choice between two whole *consumption frontiers*. As is intuitive, this fact greatly reduces the possibility of inference about discount factors for consumption.

A different but conceptually related reservation about the correct inferences to be drawn from experimental results comes from the recent work by Noor (2007). He observes that nothing excludes that experimental subjects integrating the laboratory rewards with the anticipated future levels of wealth. The striking implication is that if such future wealth levels are expected to change, all the main documented soft anomalies, including preference reversal, turn out to be compatible with the EDM. Intuitively, if the subject is more cash-constrained now than he expects to be at a later date, so that his need for money is higher now than it is expected to be in the future, he may well choose according to the pattern of

[17] i.e. with the borrowing interest rate possibly differing from the lending rate.

[18] In which consumption levels at two distinct dates are represented on each axis on the plane.

the preference reversal phenomenon, while still making his choices on the basis of a constant, and not declining, discount factor. In a precise sense, the EDM is shown in this work to have no empirical content unless integration with expected future wealth is excluded a priori. However, Noor also suggests an experimental design including risky prospects as outcomes, which is sufficiently rich to test the EDM.

Furthermore, from *field data*, Harrison Lau, and Williams (2002) find that, unlike previous claims of non-constant discount factors, although discount factors do depend on household characteristics, within each homogeneous group, discount factors over a one and three year horizon are indeed constant. But this is not all: as we mentioned earlier, the concavity of the utility function may explain apparent anomalies, thus calling for *both* the time preferences *and* the preference for the good whose receipt is delayed to be elicited simultaneously. If a single utility value $u(x, t)$ is elicited, rather than a separate assessment of the time and outcome components, it seems fair to argue that the concavity of the utility function might conflate into the estimate of discount factors. Starting from this consideration, Andersen, Harrison, Lau, and Rutström (2007) show that the implausibly high estimates of discount factors previously uncovered fall substantially once the concavity of the utility function is taken into account in the estimation, and both risk and time preferences are elicited from experimental subjects. The difference is quite dramatic: whereas under the assumption of risk neutrality the point estimate of the yearly discount factor is roughly 25 percent, it falls sixfold to about 4 percent once risk aversion is (correctly) accounted for.[19]

Summing up, then, when it comes to preferences over outcome–date pairs, once the correct estimation techniques are used and concavity of the utility function is allowed for, the wildly varying discount factor estimates fall to more manageable ranges of variation and "realistic" values.

10.5.4 Hard Anomalies

There are other observed violations of the EDM which are more fundamental in the sense that, unlike preference reversal, they seem to contradict the basic assumption of maximization of any economically reasonable objective function. One notable instance is that human decision-makers have been shown to make *intransitive* choices. Although most data in this direction come from choices under risk, the evidence available for time preferences, though limited, is clear in suggesting that violations of transitivity are more frequent in this domain. Tversky, Slovic, and Khaneman (1990) show that a substantial 15 percent of subjects exhibited cyclical

[19] Field data from the retirement options offered to retired military personnel in the USA (see Warner and Pleeter 2001) suggest higher than expected discount rates. However, see Harrison and List (2004) for a critique of the "heroic" extrapolation method used.

patterns of choice that could not be explained by "framing effects"—and in an experiment which was not designed to uncover cycles. When the issue of cycles in choice is addressed directly, the evidence is even more striking: Roelofsma and Read (2000) found that the *majority* of intertemporal choices were intransitive. Cyclical choice is thus one "solid" anomaly that cannot be accommodated within any discounting model.

While incentive-compatible experimental investigations of choices over outcome–date pairs form a small but nonnegligible literature, experimental investigations of choices over reward *sequences* are extremely thin on the ground in the economics literature, especially with financially motivated subjects. Arguably, this is because the difficulties highlighted above are exacerbated by payment of experimental subjects having to take place over weeks if not months. The unreliability of data from experiments based on hypothetical choices seems to be driving the recent increase in incentive-compatible experimental designs.

The first experimental paper on preferences over *monetary* sequences of outcomes is Loewenstein and Sicherman (1991). In a survey of members of the public entering a museum, interviewers asked participants to choose among hypothetical alternative profiles of either wages or savings plan over a number of years. Loewenstein and Sicherman (1991) found evidence of preference for sequences of increasing monetary payments (versus constant or decreasing ones). They explain this finding by pointing to a preference for maintaining the "current" consumption level, so that wages should be nondecreasing. Admittedly, these were hypothetical questions, and some respondents motivated their preference for increasing sequences with inflation. In addition, the framing of these questions as salary profiles might evoke an improvement in one's career, or just be what one would generally expect (i.e. affective response). However, other authors have found evidence of preferences for constant and even decreasing sequences of outcomes over time (e.g. Chapman 1996; Gigliotti and Sopher 1997; Guyse, Keller, and Epple 2002). The domain of choice seems also to be important (e.g. there are differences in observed choices depending on whether or not the sequences are of money, or health or environmental outcomes; see e.g. Chapman 1996 and Guyse, Keller and Epple 2002).

Manzini, Mariotti, and Mittone (2006) asked subjects to make binary choices among all possible pairs of monetary sequences, with an increasing, constant, decreasing, or "jump" (i.e. end effect) pattern, both in a paid condition (where subjects do indeed receive the sums corresponding to the sequence chosen) and an unpaid condition (where choices are hypothetical). Previous experimental evidence on reward sequences suggests that the general trend of the sequence (increasing or decreasing) is relevant to making decisions. However, in this case the data provide much weaker evidence than Loewenstein and Prelec's (1991) in support of their view that "sequences of outcomes that decline in value are greatly disliked" (p. 351). It is found that, even in the simple decision problems studied,

where monetary sequences can be clearly ordered according to their trends, simply choosing according to the heuristics that favors the "increasingness" of the trend does a rather poor job of explaining the data. The modal subject and choice are "rational", in the sense of being compatible with positive time preference combined with preference for income smoothing (concave utility function). Therefore, while choice incompatible with EDM is observed, it is not to the extent that the existing literature suggests. When there are no affective factors involved (such as, for example, the sense of dread for choices relating to health, or the sense of failure involved in a decreasing wage profile), some theory of positive discounting can provide a rough approximation of the choice patterns. However, a nonnegligible proportion of subjects (around 30 percent) choose in ways that are incompatible with any form of positive discounting (exponential, hyperbolic, or otherwise). These subjects violate the basic economic assumption that for a good, the sooner the better, suggesting that other mechanisms beyond discounting are at work. That is, Loewenstein and Prelec's pioneering findings do capture, beside affective factors, some of the heuristic considerations that people use when evaluating "neutrally" (without affects) money sequences. Moreover, the study finds that "irrational" choices present a systematic pattern, not encountered previously. Of these, the most striking are the association between certain types of rational choices and irrational choices (those who prefer a decreasing to a constant sequence are disproportionately concentrated among those who also prefer a constant to an increasing sequence) and the association between irrational choices of a different type (choosing an increasing over a decreasing sequence is very strongly associated with choosing an increasing over a constant sequence). Such patterns cannot be generated by any discounting model, or by such a model augmented with random independent mistakes.

One last puzzling experimental finding that we wish to highlight is due to Rubinstein (2003). In a hypothetical setting he finds a "single outcome/sequence of outcomes" type of preference reversal of the kind highlighted by Loewenstein and Prelec (1993) in a different domain: in a classroom experiment a majority of students preferred to receive a payment of 997 monetary units at a later date t^* than 1000 at an even later date $t^* + 1$; but when choosing between sequences of four payments of a constant amount of either 997 starting at t' or 1000 starting at $t' + 1$, the latter sequence was now preferred, contradicting the theory of hyperbolic discounting.

That is, subjects exhibited the following type of behavior: they chose x to be received at some date t^* versus the larger sum $x + z$ to be received at date $t^* + 1$ (they were impatient and preferred a smaller reward earlier rather than a larger reward later) but chose the sequence

$$a = ((x + z, t' + 1), (x + z, t' + 2), (x + z, t' + 3), (x + z, t' + 4))$$

over the sequence

$$b = ((x, t'), (x, t' + 1), (x, t' + 2), (x, t' + 3)),$$

where $t' + 4 \leq t^* + 1$. This contradicts not only the EDM but also the HDM, and in fact any model of discounting based on diminishing impatience (declining discount rates): if the subject were impatient at the late date t^* and not willing to trade off one unit of delay for an additional reward z, he should have been unwilling to perform all four tradeoffs of this type involved in the comparisons between sequences. Rubinstein's explanation is based on similarity: given choices between sequences of alternatives with two distinct attributes, the decision-maker uses a procedure whereby first of all he tries to rank alternatives based on dominance (i.e. greater outcome and earlier time of receipt); if this is not decisive, he looks for similarities between the two dimensions, trying to discriminate based on evident differences. If none can be found, then he chooses based on some different criterion.[20] As we will see in Section 10.6, however, other theories, too, can explain this phenomenon.

10.6 Empirical "Anomalies" and Theory

As we have already made clear, some of the anomalous types of behavior discussed in the empirical literature are not "hard" anomalies after all, in the sense that they can be easily accommodated within the axiom set of Theorem 2 in Section 10.2. This is the case for the main phenomenon of *preference reversal* described in Section 10.3.1. So we begin by tackling "soft anomalies" first (delay/speed-up asymmetry, magnitude effect, and inverse preference reversal), and then move to the "hard anomalies" (cycles in choice and the "single outcome/sequence of outcomes" preference reversal).

The *delay/speed-up asymmetry* does not even violate by itself any of the Fishburn–Rubinstein axioms for exponential discounting, and in this sense it is not an anomaly. Imagine for notational simplicity that the object whose receipt is to be delayed or anticipated is an amount of money x. So the effect can be written within the EDM, in a noncontradictory way, as:

$$u(x - K) = \delta^t u(x)$$

$$u(x) = \delta^t u(x + P),$$

where K and P are, respectively, the amount of money that the agent is willing to pay to anticipate the receipt of the money which he is entitled to receive at date t,

[20] See also Rubinstein (1988).

and the amount of money that the agent requires to delay to date t the receipt of the amount of money to which he is entitled now.

The *magnitude effect,* like the delay/speed-up asymmetry, also does not violate any of the Fishburn–Rubinstein axioms for exponential discounting. As noted by Ok and Masatlioglu (2007),[21] for example, the exact Thaler's numbers reported before are compatible with the EDM with a $\delta = 0.95$ and a concave utility function $u(x) = x^{0.42} + 45.9$ defined on the positive real line.

However, although formally this phenomenon is indeed compatible with exponential discounting, some observations are in order. One might argue that although *some* utility function can necessarily be found (given the nonviolation of the axioms) that fits the (few) observations, additional constraints might be desirable for the utility function, and these constraints might create an incompatibility with the observed effects. For example, it is often argued informally (and often simply taken for granted) that for small amounts the utility function ought to be linear. In this case, the EDM is incompatible with the magnitude effect (using Thaler's numbers, it would require for example $\delta = \frac{15}{60} = \frac{250}{350} = \frac{3000}{4000}$). But in this case the magnitude effect (which involves only two dates) is also incompatible with the HDM, and indeed with *any* separable discounting model axiomatized in Theorem 2 of Fishburn and Rubinstein: we would still obtain a contradiction of the type

$$\frac{x}{y} = \delta(t) = \frac{x'}{y'} \quad \text{but also} \quad \frac{x'}{y'} < \frac{x}{y}.$$

So it seems to us that the real point about such effects is that either they do not constitute an EDM anomaly, or if they do (because of the linearity of utility) they also constitute an anomaly for (much) more general discounting models, notably including HDM. It is incompatible with the linear version of relative discounting, too, and it can be made compatible with a linear utility version of Manzini and Mariotti's (σ, δ) model only in a variant in which the vagueness term σ is made to depend on the outcome (see Manzini and Mariotti 2006 for details).

In addition, it is fair to say that the evidence is still too scant. In order to fit a utility function and check its "reasonableness" (or compatibility with independent data on concavity-convexity), one would need many more observations in different regions of the time and outcome space. At best, the existing observations might be simply suggestive of the fact that human decision-makers use certain yet to be discovered "heuristic" procedures when judging outcome–date pairs. Certainly the magnitude effect, even together with the assumption of exponential discounting, is not necessarily and intrinsically related to diminishing marginal utility. For example, suppose $u(x, t) = \delta^t x^\alpha$, where $\alpha \in (0, 1)$. Then let J and K be the compensations required by a decision-maker to delay by one period from now the receipt of x and y, respectively, that is $\delta(x + J)^\alpha = x^\alpha$ and $\delta(y + K)^\alpha = y^\alpha$.

[21] See Ok and Masatlioglu (2007) in the (2003) version.

This implies

$$\frac{\delta(y+K)^{\alpha}}{\delta(x+J)^{\alpha}} = \frac{y^{\alpha}}{x^{\alpha}} \Leftrightarrow \frac{x}{x+J} = \frac{y}{y+K},$$

so that the magnitude effect (which requires that if, say, $y < x$, then $\frac{y}{y+K} < \frac{x}{x+J}$) *never* obtains. It seems unlikely that simple changes of functional forms with respect to the EDM will be descriptively adequate in general. However, as we noted already, a new literature is emerging which attempts to estimate both the shape of the utility function and discount factors at the same time (see Andersen, Harrison, Mortensen, and Rutström 2007): this type of research may in due course shed additional light on the issue of magnitude effects.

Sayman and Öncüler's (2006) *negative preference reversal* is a soft anomaly that cannot be accommodated within the class of HDM. Indeed, if $x < y$, the preferences $(x, \tau) \succ (y, t+\tau)$ and $(y, t) \succ (x, 0)$ correspond to $\delta(\tau)u(x) > \delta(t+\tau)u(y)$ and $\delta(t)u(y) > u(x)$ (in a separable discounting model). This is obviously incompatible with HDM, since it requires decreasing (rather than increasing) discount rates. To the contrary, this type of preference can be accommodated within the model of vague time preferences.

For instance, in the simple (σ, δ) representation with the TPM we need

$$x\delta^{\tau} \leq y\delta^{t+\tau} + \sigma \quad \text{and} \quad y\delta^{t+\tau} \leq x\delta^{\tau} + \sigma$$
$$y\delta^{t} > x + \sigma$$

so that $(x, \tau) \succ^{*} (y, t+\tau)$ by the secondary criterion, while $(y, t) \succ^{*} (x, 0)$ by the primary criterion. Negative preference reversal is also compatible with Ok and Masatlioglu's RDM. Here we need

$$\delta(t, 0) = \frac{1}{\delta(0, t)} > \frac{u(x)}{u(y)} > \delta(t+\tau, \tau).$$

When interpreted as revealed preferences, observed *cycles in binary choice* are a harder anomaly to deal with because they violate the fundamental axiom of Order. For outcome–date pairs, the HDM cannot explain cycles, whereas alternative theories (most notably Ok and Masatlioglu's RDM and our own theory of vague time preferences) can. Similarly, Read's subadditive discounting and Rubinstein's similarity-based decision-making are also consistent with the phenomenon.

Consider three alternatives (x, r), (y, s), and (z, t), with $x < y < z$ and $r < s < t$. In the RDM it is perfectly possible to have $u(x) > \delta(t, r)u(z)$, $u(z) > \delta(s, t)u(y)$, but $u(y) > \delta(r, s)u(x)$. This could happen, for example, if there is very little perceived difference between t and s and between s and r, but the difference between t and r is perceived as significant (imagine $r < s < t$ and $x < y < z$). Then the latter two inequalities attest to the fact that the differences $z - y$ and $y - x$ are enough to compensate for the small delays, but the difference $z - x$ is not enough to compensate for the "large" delay. Note, however, that in the model this "compounding

of small differences" effect is not allowed to hold for the outcome dimension, but just for the time dimension. With these assumptions it is also easy to see how Rubinstein's similarities might work: if t and s are similar, and so are r and s, while all other comparisons are perceived as different, then (z, t) is preferred over (y, s), (y, s) is preferred over (x, r), and because r and t are not similar, nor are x and z, it is enough for the unspecified completing criterion to pick (x, r) over (z, t) to complete the cycle.

Subadditive discounting, too, can explain cyclical choices. Consider the three time intervals $[0, r]$, $[r, s]$, and $[s, t]$, and let the discount factors over an interval $[a, b]$ be denoted by δ_{ab}. In the presence of subadditivity we have that $\delta_{0t} > \delta_{0r}\delta_{rs}\delta_{st}$. So to generate the cycle in choice where (y, s) is chosen over (z, t) which is chosen over (x, r) which is chosen over (y, s), assume that the decision-maker splits the time intervals involved in each binary comparison taking into account the common delay. Then we need:

$$\begin{aligned} u(y)\delta_{0s} &> u(z)\delta_{0s}\delta_{st}, \\ u(z)\delta_{0r}\delta_{rt} &> u(x)\delta_{0r}, \\ u(x)\delta_{0r} &> u(y)\delta_{0r}\delta_{rs}, \end{aligned}$$

which simplifies to

$$\begin{aligned} u(y) &> u(z)\delta_{st}, \\ u(z)\delta_{rt} &> u(x), \\ u(x) &> u(y)\delta_{rs}. \end{aligned}$$

The last two inequalities imply $u(z)\delta_{rt} > u(y)\delta_{rs}$, which is compatible with the first inequality as long as $u(z)\delta_{rt} > u(z)\delta_{st}\delta_{rs}$; that is, if $\delta_{rt} > \delta_{rs}\delta_{st}$, which is precisely what subadditive discounting entails.

The vague theory of time preferences is also consistent with intransitivities. For instance, in the OPM we need

$$\begin{aligned} &x\delta^r > z\delta^t + \sigma, \\ &y\delta^s \leq z\delta^t + \sigma \quad \text{and} \quad z\delta^t \leq y\delta^s + \sigma, \\ &y\delta^s \leq x\delta^r + \sigma \quad \text{and} \quad x\delta^r \leq y\delta^s + \sigma, \end{aligned}$$

so that $(x, r) \succ^* (z, t)$, but $(x, r) \sim (y, s)$ and $(y, s) \sim (z, t)$ by the primary criterion. However, by the secondary criterion, $(y, s) \succ^* (x, r)$ and $(z, t) \succ^* (y, s)$, thereby producing a cycle. Obviously any theory that can cope with cycles can cope with "direct" preference reversal, so we skip the details.

Next, consider the "*outcome–sequence*" *type of preference reversal* presented by Rubinstein (2003) for the case of monetary sequences, and Loewenstein and Prelec (1993) in a different domain. Here once again the caveat must be that both these papers present results from hypothetical questions. This notwithstanding, we report

them here because, as we discussed earlier, these pose a harder challenge to conventional theories. Obviously Rubinstein's own similarity considerations provide an explanation for this observed phenomenon. In addition, the model of vague time preferences for sequences can provide an alternative explanation. For the sequences described in the previous section, suppose for example that the secondary criterion is the natural one proposed by Rubinstein himself, namely "Pareto dominance" between the outcome sequences, and that

$$\delta^{t^*} u(x) > \delta^{t^*+1} u(x+z) + \sigma(x+z, t^*+1)$$

$$\sum_{i=0}^{3} \delta^{t'+i} u(x) \leq \sum_{i=1}^{3} \delta^{t'+i} u(x+z) + \sigma(a).$$

In this case the preference between outcome–date pairs can be explained by present discounted utility (primary criterion), and the preference between sequences can be explained by the secondary criterion.

Finally, the evidence of moderate preference for increasing sequences discussed in Manzini, Mariotti, and Mittone (2006) is also consistent with the model proposed therein, as well, obviously, as with the model by Fishburn and Edwards (1997).

10.7 Concluding Remarks

In the last twenty years a growing body of experimental evidence has posed a challenge to the standard exponential discounting model of choice over time. Attention has focused on some specific "anomalies", notably preference reversal and declining discount rates, leading to the formulation of the model of hyperbolic discounting which is finding increasing favor in the literature. As we have seen, it is debatable whether some of the most focused upon anomalies should indeed be classified as such, or whether they are really the most challenging ones for conventional theory. If they violate any axiomatic property, this is Stationarity, which is not strongly defendable even on normative, let alone descriptive, grounds. A group of theoretical ideas is beginning to emerge which can address not only violations of Stationarity, but even more challenging observed phenomena.

At the same time, at the empirical level much progress is being made on two fronts: "sophisticated" estimation of discount factors (e.g. considering censoring factors) and simultaneous presence of discounting and risk aversion, a traditionally much neglected issue until very recently. The results are quite stunning, implying as they do a serious reconsideration of the previous estimates. On the other hand,

other recent work challenges the conventional interpretation given to responses to standard experimental choice tasks.

While in the theory of choice under risk there exist rationality axioms that exert a strong normative appeal, this is less clearly the case for choice over time. Thus, it is natural to move towards models of *procedural* rationality as opposed to normative rationality, given that the latter lacks a clear notion. Some work already exists in this direction, but much more remains to be done in this fascinating area, as we are still quite far from a clear-cut "best" theory.

References

Ainslie, G. (1975). Specious Reward: A Behavioral Theory of Impulsiveness and Impulsive Control. *Psychological Bulletin*, 82, 463–96.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2007). Eliciting Risk and Time Preferences. Mimeo, University of Central Florida.

Benoit, Jean-Pierre, and Ok, Efe A. (2007). Delay Aversion. *Theoretical Economics, Society for Economic Theory*, 2/1, 71–113.

Benzion, U., Rapoport A., and Yagil, J. (1989). Discount Rates Inferred from Decisions: An Experimental Study. *Management Science*, 35, 270–84.

Chapman, G. (1996). Expectations and Preferences for Sequences of Health and Money. *Organizational Behavior and Human Decision Processes*, 67, 59–75.

Coller, M., and Williams, M. B. (1999). Eliciting Individual Discount Rates. *Experimental Economics*, 2, 107–27.

Cubitt, R., and Read, D. (2007). Can Intertemporal Choice Experiments Elicit Time Preferences for Consumption. *Experimental Economics*, online at <http://www.springerlink.com/content/v47u1v31w5306870/>.

Cummings, R., Harrison, G. W., and Rutström, E. E. (1995). Homegrown Values and Hypothetical Surveys: Is a Dichotomous Choice Approach Incentive Compatible? *American Economic Review*, 85, 260–6.

Elster, J., and Loewenstein, G. (1992). *Choice over Time*. New York: Russell Sage Foundation.

Fishburn, P. J., and Edwards, W. (1997). Discount Neutral Utility Models for Denumerable Time Streams. *Theory and Decision*, 43, 139–66.

—— and Rubinstein, A. (1982). Time Preference. *International Economic Review*, 23, 677–94.

Frederick, S., G. Loewenstein, G., and O'Donoghue, T. (2002). Time Discounting and Time Preferences: A Critical Review. *Journal of Economic Literature*, 40, 351–401.

Garratt, R., Walker, M., and Wooders, J. (2007). Behavior in Second-Price Auctions by Highly Experienced eBay Buyers and Sellers. Mimeo, University of California, Santa Barbara.

Gigliotti, G., and Sopher, B. (1997). Violations of Present Value Maximization in Income Choice. *Theory and Decision*, 43, 45–69.

Guyse, J., Keller, L., and Epple, T. (2002). Valuing Environmental Outcomes: Preferences for Constant or Improving Sequences. *Organizational Behavior and Human Decision Processes*, 87, 253–77.

Harrison, G. W. (1992). Theory and Misbehavior of First Price Auctions: Reply. *American Economic Review*, 82, 1426–43.

—— and Lau, M. (2005). Is the Evidence for Hyperbolic Discounting in Humans Just an Experimental Artefact? *Behavioral and Brain Sciences*, 28, 657.

—— and List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, 42, 1009–55.

—— Lau, M., and Williams, M. B. (2002). Estimating Individual Discount Rates in Denmark: A Field Experiment. *American Economic Review*, 92, 1606–17.

Hertwig, R., and Ortmann, A. (2001). Experimental Practices in Economics: A Methodological Challenge for Psychologists? *Behavioral Brain Science*, 24, 383–451.

Kahneman, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 363–91.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, 112, 443–77.

Loewenstein, G. (1987). Anticipation and the Valuation of Delayed Consumption. *Economic Journal*, 97, 666–84.

—— (1988). Frames of Mind in Intertemporal Choice. *Management Science*, 34, 200–14.

—— and Prelec, D. (1991). Decision Making over Time and under Uncertainty: A Common Approach. *Management Science*, 37, 770–86.

—— —— (1992). Anomalies in Intertemporal Choice: Evidence and Interpretation. *Quarterly Journal of Economics*, 107, 573–97.

—— —— (1993). Preferences for Sequences of Outcomes. *Psychological Review*, 100, 91–108.

—— and Sicherman, N. (1991). Do Workers Prefer Increasing Wage Profiles? *Journal of Labour Economics*, 9, 67–84.

—— and Thaler, R. (1989). Intertemporal Choice. *Journal of Economic Perspectives*, 3, 181–93.

Manzini, P., and Mariotti, M. (2006). A Vague Theory of Choice over Time. *Advances in Theoretical Economics*, 6/1, article 6.

—— —— (2007). Sequentially Rationalizable Choice. *American Economic Review*, 97, 1824–39.

—— —— and Mittone, L. (2006). Choosing Monetary Sequences: Theory and Experimental Evidence. CEEL WP 1–06 and IZA DP no. 2129.

—— —— —— (2007). The Elicitation of Time Preferences. Research report for ESRC grant RES-000-22-1636, mimeo.

Noor, J. (2007). Hyperbolic Discounting and the Standard Model. Mimeo, Boston University.

Noussair, C., Roibin, S., and Ruffieux, B. (2004). Revealing Consumers' Willingness to Pay: A Comparison of the BDM Mechanism and the Vickrey Auction. *Journal of Economic Psychology*, 25, 725–41.

Ok, Efe A., and Masatlioglu, Y. (2003). A General Theory of Time Preferences. Mimeo, Department of Economics, New York University.

—— —— (2007). A Theory of (Relative) Discounting. *Journal of Economic Theory*, 137/1, 214–45.

Ortmann, A., and Hertwig, R. (2005). Monetary Incentives: Usually Neither Necessary Nor Sufficient? CERGE-EI Working Paper no. 307.

PENDER, J. (1996). Discount Rates and Credit Markets: Theory and Evidence from Rural India. *Journal of Development Economics*, 50, 257–96.

PHELPS, E., and POLLACK, R. (1968). On Second Best National Savings and Game-Equilibrium Growth. *Review of Economic Studies*, 35, 201–8.

PIGOU, A. C. (1920). *The Economics of Welfare*. London: Macmillan.

PLOTT, C., and ZEILER, K. (2005). The Willingness to Pay/Willingness to Accept Gap, the Endowment Effect, Subject Misconceptions and Experimental Procedures for Eliciting Valuations. *American Economic Review*, 95, 530–45.

READ, D. (2001). Is Time-Discounting Hyperbolic or Subadditive? *Journal of Risk and Uncertainty*, 23, 5–32.

—— (2005). Monetary Incentives, What are they Good For? *Journal of Economic Methodology*, 12, 265–7.

—— and POWELL, M. (2002). Reasons for Sequence Preferences. *Journal of Behavioral Decision Making*, 15, 433–60.

ROELOFSMA, P. H., and READ, D. (2000). Intransitive Intertemporal Choice. *Journal of Behavioral Decision Making*, 13, 161–77.

RUBINSTEIN, A. (1988). Similarity and Decision Making under Risk (Is there a Utility Theory Resolution to the Allais Paradox?). *Journal of Economic Theory*, 46, 145–53.

—— (2001). A Theorist's View of Experiments. *European Economic Review*, 45, 615–28.

—— (2003). Is it Economics and Psychology? The Case of Hyperbolic Discounting. *International Economic Review*, 44, 1207–16.

RUTSTRÖM, E. E. (1998). Home-Grown Values and Incentive Compatible Auction Design. *International Journal of Game Theory*, 27, 427–41.

SAMUELSON, P. (1937). A Note On the Measurement of Utility. *Review of Economic Studies*, 4, 155–61.

SAYMAN, S., and ÖNCÜLER, A. (2006). Reverse Time Inconsistency. Mimeo, INSEAD.

SCHOLTEN, M., and READ, D. (2006). Discounting by Intervals: A Generalized Model of Intertemporal Choice. *Management Science*, 52, 1424–36.

SHELLEY, M. (1993). Outcome Signs, Question Frames and Discount Rates. *Management Science*, 39, 806–15.

STROTZ, R. H. (1956). Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies*, 23, 165–80.

THALER, R. (1981). Some Empirical Evidence on Dynamic Inconsistency. *Economics Letters*, 8, 201–7.

TVERSKY, A. (1969). Intransitivity of Preferences. *Psychological Review*, 76, 31–48.

—— and KAHNEMAN, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211, 453–8.

—— and KOEHLER, D. (1994). Support Theory: Nonextensional Representation of Subjective Probability Theory. *Psychological Review*, 101, 547–67.

—— SATTATH, S., and SLOVIC, P. (1988). Contingent Weighting in Judgement and Choice. *Psychological Review*, 95, 371–84.

—— SLOVIC, P., and KAHNEMAN, D. (1990). The Causes of Preference Reversal. *American Economic Review*, 80, 204–17.

WARNER, J. T., and PLEETER, S. (2001). The Personal Discount Rate: Evidence from Military Downsizing Programs. *American Economic Review*, 91, 33–53.

CHAPTER 11

IMITATION AND LEARNING

CARLOS ALÓS-FERRER
KARL H. SCHLAG

11.1 Introduction

Imitation is one of the most common forms of human learning. Even if one abstracts from explicit evolutionary or cultural arguments, it is clear that the pervasiveness of imitation can be explained only if it often leads to desirable results. However, it is easy to conceive of situations where imitating the behavior of others is *not* a good idea; that is, imitation is a form of boundedly rational behavior. In this chapter we survey research which clarifies the circumstances in which imitation is desirable. Our approach is partly normative, albeit we do not rely on the standard Bayesian belief-based framework, but rather view imitation as a belief-free behavioral rule.

There are many reasons why imitation may be a "good strategy". Some of them are:

a. To free-ride on the superior information of others.
b. To save calculation and decision-taking costs.
c. To take advantage of being regarded as similar to others.
d. To aggregate heterogeneous information.
e. To provide a coordination device in games.

Sinclair (1990) refers to (a) as "information-cost saving". The argument is that, by imitating others, the imitator saves the information-gathering costs behind the

observed decisions. Examples range from children's learning[1] to research and development strategies.[2] Analogously, (b) refers to the fact that imitation economizes information-processing costs. A classical model building on this point is Conlisk (1980). Another nice example is Rogers (1989), who shows in an evolutionary model how an equilibrium proportion of imitators arises in a society when learning the true state is costly. Pingle and Day (1996) discuss experimental evidence showing that subjects use imitation (and other modes of "economizing behavior") in order to avoid decision costs.

Examples for (c) abound in ecology, where the phenomenon is called *mimicry*. In essence, an organism mimics the outward characteristics of another one in order to alter the behavior of a third organism (e.g. a predator). In line with this argument, Veblen (1899) describes lower social classes imitating higher ones through the adoption of fashion trends. Clearly, a case can be made for imitation as signaling (see e.g. Cho and Kreps 1987). In a pooling equilibrium, some agents send a specific signal in order to be mistaken for agents of a different type. In the model of Kreps and Wilson (1982) a flexible chain store that has the option to accommodate entry imitates the behavior of another chain store that can only act tough and fight entry.

The first three reasons we have just discussed are centered on the individual decision-maker. We will focus on the remaining two, which operate at a social level. There are two kind of examples in category (d). Banerjee (1992) and the subsequent herding literature show that the presence of intrinsic information in observed choices can lead rational individuals to imitate others, even disregarding conflicting private signals. A conceptually related example is Squintani and Välimäki (2002). Our objective here, though, is to provide a less rational perspective on (d) and (e) by showing whether and how individuals who aim to increase payoffs would choose to imitate.

We remind the reader that our approach is not behavioral in the sense of selecting a behavioral rule motivated by empirical evidence and finding out its properties. Although we do not shy away from empirical or experimental evidence on the actual form and relevance of imitation, in this chapter we consider abstract rules of behavior and aim to find out whether some of them have "nicer" properties than others. Thus, our approach to bounded rationality allows for a certain sophistication of the agents. To illustrate, the two central results below are as follows.

[1] Experiments in psychology have consistently shown that children readily imitate behavior exhibited by an adult model, even in the absence of the model. See e.g. Bandura (1977). More generally, Bandura (1977, p. 20) states that "[L]earning would be exceedingly laborious, not to mention hazardous, if people had to rely solely on the effects of their own actions to inform them what to do. Fortunately, most human behavior is learned observationally through modeling: from observing others one forms an idea of how new behaviors are performed, and on later occasions this coded information serves as a guide for action."

[2] The protection of innovators from imitation by competitors is the most commonly mentioned justification for patents. Interestingly, Bessen and Maskin (2007) argue that "in a dynamic world, imitators can provide benefit to both the original innovator and to society as a whole".

First, some imitation rules are better than others, e.g. rules which prescribe blind imitation of agents who perform better are dominated by a rule that incorporates the degree of better performance in the imitation behavior. That is, the reluctance to switch when the observed choices are only slightly better than their own might not be due to switching costs, but rather to the fact that the payoff sensitivity of the imitation rule allows the population to learn the best option. Second, in a large class of strategic situations (games), the results of imitation present a systematic bias from "rational" outcomes, while still ensuring coordination.

11.2 Social Learning in Decision Problems

Consider the following model of social learning in decision problems. There is a population of decision-makers (DMs) who independently and repeatedly face the same decision in a sequence of rounds. The payoff obtained by choosing an action is random and drawn according to some unknown distribution, independently across DMs and rounds. Between choices, each DM receives information about the choices of other DMs and decides according to this information which action to choose next. We restrict attention to rules that specify what to choose next relying only on one's own experience and observations in the previous round.[3] Our objective is to identify simple rules which, when used by all DMs, induce increasing average payoffs over time.

More formally, a decision problem is characterized by a finite set of actions A and a payoff distribution P_i with finite mean π_i for each action $i \in A$. That is, P_i determines the random payoff generated when choosing action i, and π_i is the expected payoff of action i. We will make further assumptions on the distributions P_i below. Action j is called *best* if $j \in \arg\max\{\pi_i, i \in A\}$. Further, denote by $\Delta = \{(x_i)_{i\in A} | \sum_{i\in A} x_i = 1,\ x_i \geq 0\ \forall i \in A\}$ the set of probability distributions (mixed actions) on A.

We consider a large population of DMs. Formally, the set of DMs might be either finite or countably infinite. Schlag (1998) explicitly works with a finite framework, while Schlag (1999) considers a countably infinite population. Essentially, the main results do not depend on the cardinality, but the interpretation of both the model and the considered properties does change (slightly). Thus, although the analysis is simpler in the countably infinite case, it is often convenient to keep a large but finite population intuition in mind.

[3] This "Markovian" assumption is for simplicity. Bounded memory is simply a (realistic) way to constrain the analysis to simple rules.

All DMs face the same decision problem. While we are interested in repeated choice, for our analysis it is sufficient to consider two consecutive rounds. Let p_i be the proportion of individuals[4] who choose action i in the first round. Let $\bar{\pi} = \sum p_i \pi_i$ denote the average expected payoff in the population. Let p_i' denote the (expected) proportion of individuals who choose action i in the next (or second) round.

We consider an exogenous, arbitrary $p = (p_i)_{i \in A} \in \Delta$ and concentrate on learning between rounds. In the countably infinite case, the proportion of individuals in the population choosing action j is equal to the proportion of individuals choosing j that a given DM faces. In a finite population one has to distinguish in the latter case whether or not the observing individual is also choosing j.

11.2.1 Single Sampling and Improving Rules

Consider first the setting in which each individual observes one other individual between rounds. We assume *symmetric sampling:* the probability that individual ω observes individual ω' is the same as vice versa. This assumption arises naturally from a finite population intuition, when individuals are matched in pairs who then see each other. A particular case is that of *random sampling*, where the probability that a DM observes some other DM who chose action j is p_j.[5]

A learning rule F is a mapping which describes the DM's choice in the next round as a function of what she observed in the previous round. We limit attention to rules that do not depend on the identity of those observed, only on their choice and success. For instance, if each DM observes the choice and payoff of exactly one other DM, a learning rule can be described by a mixed action $F(i, x, j, z) \in \Delta$, where $F(i, x, j, z)_k$ is the probability of choosing action k in the next round after playing action i, getting payoff x, and observing an individual choosing action j that received payoff z.

Clearly, the new proportions p_i' are a function of p, F, and the probabilities with which a DM observes other DMs. For our purposes, however, it will be enough to keep the symmetric sampling assumption in mind.

A learning rule F is *improving* if the average expected payoff in the population increases for all p when all DMs use F, i.e. if $\sum p_i' \pi_i \geq \sum p_i \pi_i$, for all possible payoff distributions. If there are only two actions, this is equivalent to requiring that

[4] For the countably infinite case, the p_i are computed as the limits of Cesàro averages, i.e. $p_i = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} g_i^k$ where the population is taken to be $\{1, 2, \ldots\}$ and $g_i^k = 1$ if DM k chooses action i, zero otherwise. This entails the implicit assumption that the limit exists.

[5] Note that "uniform sampling" is impossible within a countably infinite population. For details on random matching processes for countably infinite populations, see Boylan (1992), whose constructions can be used to show existence of the sampling procedures described here.

$\pi_i > \pi_j$ implies $p_i' \geq p_i$.[6] We will provide a characterization of improving rules which, surprisingly, requires the rule to be a form of imitation. A rule F is *imitating* if a DM switches only to observed actions, formally, if $F(i, x, j, y)_k > 0$ implies $k \in \{i, j\}$. In particular, if their own and the observed choice are identical, then the DM will choose the same action again.

For imitating rules, it is clear that the change in the population expected payoff depends only on the expected net switching behavior between two DMs who see each other. $F(i, x, j, z)_j - F(j, z, i, x)_i$ specifies this net switching behavior when one individual who chose i and received x sees an individual who chose j and received z. The expected net switching behavior is then given by

$$F(i, j)_j - F(j, i)_i = \iint (F(i, x, j, z)_j - F(j, z, i, x)_i) dP_i(x) dP_j(y). \qquad (1)$$

Thus, an imitating rule F is improving if (and only if) the expected net switching behavior from i to j is positive whenever $\pi_j > \pi_i$. The "only if" part comes from the fact that one can always consider a state where only i, j are used.

There are of course many different imitating rules. We consider a first example. An imitating rule F is called *Imitate If Better (IIB)* if $F(i, x, j, z)_j = 1$ if $z > x$, $F(i, x, j, z)_j = 0$ otherwise ($j \neq i$). This rule, which simply prescribes imitating the observed DM if her payoff is larger than one's own payoff, is possibly the most intuitive of the imitation rules.

The following illustrative example is taken from Schlag (1998, p. 142). Suppose the payoff distributions are restricted to be of the following form. Action i yields payoff $\pi_i + \epsilon$, where π_i is deterministic but unknown, and ϵ is random independent noise with mean 0. In statistical terminology, we consider distributions underlying each choice that differ only according to a location parameter. We refer to this case as the *idiosyncratic noise* framework. We make no assumptions on the support of ϵ; in particular, the set of achievable payoffs may be unbounded.

Schlag (1998) shows that Imitate If Better is improving under idiosyncratic noise. To see why, assume $\pi_j > \pi_i$. Consider first the case where two DMs who see each other have received the same payoff shock ϵ. Since $F(i, \pi_i + \epsilon, j, \pi_j + \epsilon)_j - F(j, \pi_j + \epsilon, i, \pi_i + \epsilon)_i \geq 0$, we find that net switching has the right sign. Now consider the case where the two DMs receive different shocks, ϵ_1 and ϵ_2. By symmetric sampling, and since payoff shocks follow the same distribution, this case has the exact same likelihood as the opposite case where each DM receives the shock that the other DM receives in the considered case. Thus we can just add these two cases

[6] Suppose that a DM replaces a random member of an existing population and observes the previous action and payoff of this individual before observing the performance of others. Then the improving condition requires that the DM is expected to attain a larger payoff than the one she replaced. This interpretation is problematic in a countably infinite population, due to the nonexistence of uniform distributions.

for the computation in (1),

$$(F(i, \pi_i + \epsilon_1, j, \pi_j + \epsilon_2)_j - F(j, \pi_j + \epsilon_1, i, \pi_i + \epsilon_2)_i) + (F(i, \pi_i + \epsilon_2, j, \pi_j + \epsilon_1)_j - F(j, \pi_j + \epsilon_2, i, \pi_i + \epsilon_1)_i), \quad (2)$$

and the conclusion follows because both terms in parentheses are nonnegative for IIB.

The case of idiosyncratic noise is of course particular; however, it provides us with a first intuition. We now consider an alternative, more reasonable restriction on the payoffs. The distributions P_i are allowed to be qualitatively different, but we assume that they have support on a common, nondegenerate, bounded interval. Without loss of generality, the payoff interval can be assumed to be [0, 1] after an affine transformation. The explicit assumption is that payoffs are known to be contained in [0, 1]. We refer to this case as the *bounded payoffs* framework.

We introduce now some other rules, whose formal definitions are adapted to the bounded payoffs case (IIB is of course also well-defined). An imitating rule F is called

- *Proportional Imitation Rule (PIR)* if $F(i, x, j, z)_j = \max\{z - x, 0\}$ $(j \neq i)$,
- *Proportional Observation Rule (POR)* if $F(i, x, j, z)_j = z$ $(j \neq i)$,
- *Proportional Reviewing Rule (PRR)* if $F(i, x, j, z)_j = 1 - x$ $(j \neq i)$.

PIR seems a plausible rule, making imitation depend on how much better the observed DM performed. POR has a smaller intuitive appeal, because it ignores the DM's own payoff. PRR is based on an aspiration-satisfaction model. The DM is assumed to be satisfied with her action with probability equal to the realized payoff; e.g. she draws an aspiration level from a uniform distribution on [0, 1] at the beginning of each round. A satisfied DM keeps her action while an unhappy one imitates the action used by the observed individual. Schlag (1998) shows:

Proposition 1. *Under bounded payoffs, a rule F is improving if and only if it is imitating and for each $i, j \in A$ such that $i \neq j$ there exists $\sigma_{ij} \in [0, 1]$ with $\sigma_{ij} = \sigma_{ji}$ and $F(i, x, j, z)_j - F(j, z, i, x)_i = \sigma_{ij} \cdot (z - x)$ for all $x, z \in [0, 1]$.*

The "if" statement is easily verified. The intuition behind the "only if" statement is as follows. Assume F is improving. We first show why F is imitating, so why $F(i, x, j, z)_k = 0$ for all other actions $k \neq i, j$. The improving condition must hold in all states, so assume that all DMs are choosing action i or j. Now consider an individual who chose i and received x and who observed j receiving z. It could be that all other actions $k \neq i, j$ always generate payoff 0 while $\pi_i = \pi_j > 0$. This means that everyone is currently choosing a best action, and if a nonnegligible set of individuals switches to any other action, then the average payoff decreases in expected terms. Hence our hypothetical DM should not switch.

Given that an improving rule is imitating, as already observed, the relevant quantities for the change in population expected payoffs are the expected net switching

rates given by (1). Again, if $\pi_j > \pi_i$, then the expected net switching behavior from i to j has to be positive. Of course neither π_i nor π_j is observable. But note that the sign of $\pi_j - \pi_i$ can change when the payoffs that occur with positive probability under P change slightly (recall that the improving condition requires an improvement in population payoffs for all possible payoff distributions). Thus the net switching behavior needs to be sensitive to small changes in the received payoffs. It turns out that this sensitivity has to be linear in order for average switching behavior based on realized payoffs to reflect differences in expected payoffs. The linearity is required by the need to translate behavior based on observations into behavior in terms of expected payoffs.

Given Proposition 1, it is easy to see that average payoffs increase more rapidly for larger σ_{ij}. The upper bound of 1 on σ_{ij} is due to the restriction that F describes probabilities. Note that if payoffs were unbounded, or at least if there were no known bound, then the above argument would still hold; however the restriction that F describes a probability would imply that $\sigma_{ij} = 0$ for all $i \neq j$. Hence, improving rules under general unbounded payoffs generate no aggregate learning as average payoffs do not change over time (due to $\sigma_{ij} = 0$). As we have seen, this is in turn no longer true if additional constraints are placed on the set of possible payoff distributions.

A direct application of Proposition 1 shows that, for bounded payoffs, PIR, PRR, and POR are improving, while IIB is not. For example, let $A = \{1, 2\}$, and consider P such that $P_1(1) = 1 - P_1(0) = \lambda$ and $P_2(x) = 1$ for some given $x \in (0, 1)$. Then $F(2, 1)_1 = \lambda$ and $F(1, 2)_2 = 1 - \lambda$, so $F(2, 1)_1 - F(1, 2)_2 = 2\lambda - 1$. Consequently, $x < \lambda < \frac{1}{2}$ and $p_1 \in (0, 1)$ implies $\pi_1 > \pi_2$ but $\bar{\pi}' < \bar{\pi}$.

Higher values of σ_{ij} induce a larger change in average payoffs. Thus it is natural to select among the improving rules those with maximal σ_{ij}. Note that PIR, POR and PRR satisfy $\sigma_{ij} = 1$ for all $i \neq j$. Thus all three can be regarded among the "best" improving rules. Each of these three rules can be uniquely characterized among the improving rules with $\sigma_{ij} = 1$ for all $i \neq j$ (see Schlag 1998). PIR is the unique such rule where the DM never switches to an action that realized a lower payoff. This property is very intuitive, although it is not necessary for achieving the improving property. However, it does lead to the lowest variance when the population is finite. PRR is the unique improving rule with $\sigma = 1$ that does not depend on the payoff of the observed individual. Hence it can be applied when individuals have less information available, i.e. when they observe the action and not the payoff of someone else. Last, POR is the unique such rule that does not depend on own payoff received. The fact that own payoff is not necessary for maximal increase in average payoffs among the improving rules is here an interesting finding that adds insights to the underlying structure of the problem.

Of course, Proposition 1 depends on the assumption of bounded payoffs. As we have illustrated, IIB is not payoff-improving under bounded payoffs, but it is under idiosyncratic noise. Additionally, the desirability of a rule depends on the criterion used. Proposition 1 focuses on the improving property. Oyarzun and Ruf

(2007) study an alternative property. A learning rule is *first-order monotone* if the proportion of DMs who play actions with first-order stochastic dominant payoff distributions is increasing (in expected terms) in any environment. They show that both IIB and PIR have this property. Actually, all improving rules have this property. This follows directly from Lemma 1 in Oyarzun and Ruf (2007), which establishes that first-order monotonicity can be verified from two properties: (i) imitating and (ii) positive net switching to first-order stochastically dominant actions. Oyarzun and Ruf (2007) go on to show that no "best rule" can be identified within this class of rules; the intuition is that the class is too large, or, in other words, that the concept of first-order monotonicity is too weak.

11.2.2 Population Dynamics and Global Learning

Consider a countably infinite population, and suppose that all DMs use the same improving rule with $\sigma_{ij} = 1$ for $i \neq j$, e.g. one of the three improving rules mentioned above: PIR, PRR, POR. Further, assume random sampling. Direct computation from Proposition 1 (see Schlag 1998) shows that

$$p_i' = p_i + (\pi_i - \bar{\pi}) \cdot p_i, \quad \text{and}$$
$$\bar{\pi}' = \bar{\pi} + \sum_i (\pi_i - \bar{\pi})^2 \cdot p_i = \bar{\pi} + \frac{1}{2} \sum_{i,j} (\pi_i - \pi_j)^2 p_i p_j. \tag{3}$$

If the learning procedure is iterated, one obtains a dynamical system in discrete time. Its analysis shows that for any interior initial condition (i.e. if $p \gg 0$), in the long run the proportion of individuals choosing a best action tends to 1.

We see in (3) that the expected increase in choice of action i is proportional to the frequency of the individuals currently choosing action i and to the difference between the expected payoff of action i and the average expected payoff among all individuals. The expected dynamic between learning rounds is thus a discrete version of what is known as the (*standard*) *replicator dynamics* (Taylor 1979; Weibull 1995).

For a finite population, the dynamics become stochastic, as one cannot implicitly invoke a law of large numbers. In (3), p' has to be replaced by its expectation Ep', and the growth rate has to be multiplied by $N/(N-1)$, where N is the number of individuals. Schlag (1998) provides a finite-horizon approximation result relating the dynamics for large but finite population and (3). The explicit analysis of the long-run properties of the finite-population dynamics has not yet been undertaken.

Improving captures local learning much in the spirit of evolutionary game theory, where payoff monotone selection dynamics are considered as the relevant class (cf. Weibull 1995). Selection dynamics refers to the fact that actions not chosen at present will also not be chosen in the future. Imitating rules lead to such

dynamics. When there are only two actions, monotonicity is equivalent to requiring that the proportion of those playing the best action is strictly increasing over time whenever both actions are initially present. Thus an improving rule generates a particular payoff monotone dynamics. This is particularly clear for PIR, PRR, and POR in view of (3).

Alternatively, one may be interested in global learning in terms of the ability to determine which action is best in the long run. Say that a rule is a *global learning rule* if for any initial state all DMs choose a best action in the long run. Say that a rule is a global *interior learning rule* if for any initial state in which each action is present all DMs choose a best action in the long run. The following result is from Schlag (1998).

Proposition 2. *Assume random sampling, countably infinite population, and consider only two actions. A rule is a global interior learning rule if and only if it is improving with* $\sigma_{12} > 0$. *There is no global learning rule.*

A cursory examination of (3) shows that PIR, PRR, and POR are global interior learning rules. For an arbitrary global interior learning rule, note that the states in which all DMs use the same action have to be absorbing, in order to enable them to be possible long-run outcomes. In the interior, as there are only two actions, global learning requires that the number of those using the best action increases strictly. This is slightly stronger than the improving condition, hence the additional requirement that $\sigma_{12} > 0$.

The fact that no improving rule is a global learning rule is due to the imitating property. When all DMs start out choosing a bad action, then they will keep doing so forever, as there is no information about the alternative action. But, global learning requires some DM to switch if the action is not a best action.[7]

If the population is finite, then there is also no global interior learning rule. Any state in which all DMs choose the same action is absorbing and can typically be reached with positive probability. Thus, not all DMs will be choosing a best action in the long run. However, this is simply due to the fact that the definition of global learning is not well-suited to explicitly stochastic frameworks. It would be reasonable to consider rare exogenous mistakes or mutations and then to investigate the long run when these mistakes become small, as in Kandori *et al.* (1993) and Young (1993) (see also Section 11.3). Alternatively, one could allow for nonstationary rules where DMs experiment with unobserved actions. Letting this experimentation vanish over time at the appropriate rate one can achieve global learning. The proof technique involves stochastic approximation (e.g. see Fudenberg and Levine 1998, appendix of ch. 4).

We comment briefly on a scenario involving global learning with *local interactions*. Suppose that each individual is indexed by an integer, and that learning

[7] The argument relies on the fact that the rule is stationary, i.e. it cannot condition on the period.

occurs only by randomly observing one of the two individuals indexed with adjacent integers. Consider two actions only, fix $t \in \mathbb{Z}$ and assume that all individuals indexed with an integer less than or equal to t choose action 1, while those with an index strictly higher than t choose action 2. Dynamics are particularly simple under PIR. Either individual t switches or individual $t+1$ switches, thus maintaining over time a divide in the population between choice of the two actions. Since $F(1,2)_2 - F(2,1)_1 = \pi_2 - \pi_1$, we obtain that $\pi_2 > \pi_1$ implies $F(1,2)_2 > F(2,1)_1$. Thus, the change in the position of the border between the two regions is governed by a random walk, hence in the long run all will choose a best action. PIR is a global interior learning rule. A more general analysis for more actions or more complex learning neighborhoods is not yet available.

11.2.3 Selection of Rules

Up to now we have assumed that all DMs use the same rule and have then evaluated performance according to change in average expected payoffs within the population. Here we briefly discuss whether this objective makes sense from an individual perspective. We present three scenarios. First, there is a *global individual learning* motivation. Suppose each DM wishes to choose the best action among those present in the long run. Then it is a Nash equilibrium that each individual chooses the same global learning rule. Second, there is an *individual improving* motivation. Assume that a DM enters a large, finite population, by randomly replacing one of its members where the entering DM is able to observe the previous action and payoff of the member replaced. If the entering DM wishes to guarantee an increase in expected payoffs in the next round, as compared to the status quo of choosing the action of the member replaced, then the DM can choose an improving rule regardless of what others do. The role of averaging over the entire population is now played by the fact that replacement is random in the sense that each member is equally likely to be replaced and the improving condition is evaluated *ex ante* before entry.

Last, we discuss briefly a setting with *evolution of rules*, in which rules are selected according to their performance. A main difficulty with such a setting is that there is no selection pressure once two rules choose the same action, provided selection is based on performance only. In particular, this means that there is no evolutionarily stable rule.

Björnerstedt and Schlag (1996) investigate a simple setting with two actions in which only two rules are present at the same time. An individual stays in the population with probability proportional to the last payoff achieved, and exits otherwise. Individuals entering the population sample some existing individual at random and adopt the rule and action of that individual. Neutral stability is investigated, defined as the ability to remain in an arbitrarily large fraction, provided sufficiently

many are present initially. It turns out that a rule is neutrally stable in all decision problems if and only if it is strictly improving (improving with $\sigma_{12} > 0$). Note that this result holds regardless of which action is being played among those using the strictly improving rule. Even if some are choosing the better action, it can happen, when the alternative rule stubbornly chooses the worse action, that in the long run all choose the worst action. Rules that sometimes experiment do not sustain the majority in environments where the action they experiment with is a bad action. Rules that are imitative but not strictly improving lose the majority against alternative rules that are not imitative.

A general analysis with multiple rules has not yet been undertaken.

11.2.4 Learning from Frequencies

We now consider a model in which DMs do not observe the individual behavior of others but instead know the population frequencies of each action. That is, each individual observes only the population state p. A rule F becomes a function of p; that is, $F(i, x, p)_j$ is the probability of choosing j after choosing i and obtaining payoff x when the vector of population frequencies is equal to p.

Notice that PRR can be used to create an improving rule even if there is no explicit sampling of others. Each DM can simply apply PRR to determine whether to keep the previous action or to randomly choose an action used by the others, selecting action j with probability p_j. Formally, $F(i, x, p)_i = x + (1 - x)p_i$ and $F(i, x, p)_j = (1 - x)p_j$ if $j \neq i$. The results of this rule are indistinguishable from the situation where there is sampling and then PRR is applied. Thus the resulting dynamic is the one described in (3).

This can be improved upon, though. We present an improving rule that outperforms any rule under single sampling in terms of change in average payoffs when there are two actions. The aim is to eliminate the inefficiencies created under single sampling when two individuals choosing the same action observe each other. The idea is to act as if there is some mediator that matches everyone in pairs such that the number of individuals seeing the same action is minimal. Suppose $p_1 \geq p_2 > 0$. All individuals choosing action 2 are matched with an individual choosing action 1. There are a total of p_2 pairs of individuals matched with a different action; thus an individual choosing action 1 is matched with one choosing action 2 with probability p_2/p_1. After being matched, PRR is used. Thus, individuals using action 2 who are in the minority switch if not happy. Individuals using action 1 that are not happy switch with probability p_2/p_1. Formally, F is imitating, and for $p \gg 0$, $F(2, y, p)_1 = 1 - y$, and $F(1, x, p)_2 = (p_2/p_1)(1 - x)$. Consequently,

$$p_1' = p_2(1 - \pi_2) + p_1\left(1 - \frac{p_2}{p_1}(1 - \pi_1)\right)$$

and hence

$$p_i' = p_i + \frac{1}{\max\{p_1, p_2\}} p_i(\pi_i - \bar{\pi}). \tag{4}$$

In particular, F is improving and yields a strictly higher increase in average payoffs than under single sampling whenever not all actions present yield the same expected payoff. For instance, when there are two actions, then the growth rate is up to twice that of single sampling. However, its advantage vanishes as the proportion of individuals choosing a best action tends to 1.

11.2.5 Absolute Expediency

There is a close similarity between learning from frequencies and the model of learning of Börgers *et al.* (2004), which centers on a refinement of the improving concept. A learning rule F is *absolutely expedient* (Lakshmivarahan and Thathachar 1973) if it is improving and $\sum p_i' \pi_i > \sum p_i \pi_i$ unless $\pi_i = \pi_j$ for all i, j. That is, the DM's expected payoff increases strictly in expected terms from one round to the next for every decision problem, unless all actions have the same expected payoff. Note that, as a corollary of Proposition 1, when exactly one other DM is observed, absolutely expedient rules are those where, additionally, $\sigma_{ij} > 0$ for all distinct $i, j \in A$.

Börgers *et al.* (2004) consider a single individual who updates a mixed action (i.e. a distribution over actions) after realizing a pure action based on it. They search for a rule that depends on the mixed action, the pure action realized, and the payoff received, that is absolutely expedient. Thus, expected payoff in the next round is supposed to be at least as large as the expected payoff in the current round, strictly larger unless all actions yield the same expected payoff. For two actions Börgers *et al.* (2004) show that a rule is absolutely expedient if and only if there exist $B_{ii} > 0$ and A_{ii} such that $F(i, x, p)_i = p_i + p_j(A_{ii} + B_{ii}x)$. They show that there is a best (or dominant) absolutely expedient rule that achieves higher expected payoffs than any other. This rule is given by $B_{ii} = 1/\max\{p_1, p_2\}$ and $A_{ii} = -\min\{p_1, p_2\}/\max\{p_1, p_2\}$. A computation yields

$$E(p_1' | p) = p_1 + \frac{1}{\max\{p_1, p_2\}} p_1(\pi_1 - \bar{\pi}).$$

Note that the expected change is equal to the change under our rule shown in (4). This is no surprise, as any rule that depends on frequencies can be interpreted as a rule for individual learning, and vice versa. Note that Börgers *et al.* (2004) also show that there is no best rule when there are more than two actions.

Since Börgers *et al.* (2004) work with rules whose unique input is one's own received payoff, most imitating rules are excluded. Morales (2002) considers rules where the action and payoff of another individual are also observed, as in Schlag

(1998), although, like Börgers *et al.*, he considers mixed-action rules. He does not consider general learning rules, but rather focuses directly on imitation rules, in the sense that the probabilities attached to nonobserved actions are not updated.[8] Still, the main result of Morales (2002) is in line with Schlag (1998). An imitating rule is absolutely expedient if and only if it verifies two properties. The first one, unbiasedness, specifies that the expected movement induced by the rule is zero whenever all actions yield the same expected payoff. The second one, positivity, implies that the probability attached to the action chosen by the DM is reduced if the received payoff is smaller than the one of the observed individual, and increased otherwise. Specifically, as in Proposition 1, "the key feature is proportional imitation, meaning that the change in the probability attached to the played strategy is proportional to the difference between the received and the sampled payoff" (Morales 2002, p. 476).

Morales (2002) identifies the "best" absolutely expedient imitating rule, which is such that the change in the probability of one's own action is indeed proportional to the payoff difference, but the proportionality factor is the minimum of the probabilities of one's own action and the sampled one. Absolute expediency, though, does not imply imitation. In fact, there is a non-imitating, absolutely expedient rule which, in this framework, is able to outperform the best imitating one. This rule is the classical reinforcement learning model of Cross (1973). The reason behind this result is that, in a framework in which the DM plays mixed actions, an imitating rule is not allowed to update the choice probabilities in the event that the sampled action is the same as the chosen one, while Cross's rule uses the observed payoffs in order to update the choice probabilities.

11.2.6 Multiple Sampling

Consider now the case where each DM observes $M \geq 2$ other DMs. We define the following imitation rules. *Imitate the Best* (*IB*) specifies imitating the action chosen by the observed individual who was most successful. *Imitate the Best Average* (*IBA*) specifies considering the average payoff achieved by each action sampled and then choosing the action that achieved the highest average. For both rules, if there is a tie among several best-performing actions, the DM randomly selects among them, except if his own action is one of the best-performing. In the latter case, the DM does not switch.

The *Sequential Proportional Observation Rule* (*SPOR*) is the imitation rule which applies POR sequentially as follows. Randomly select one individual among those observed, including those that chose the same action as you did. Imitate her action with probability equal to her payoff. With the complementary probability,

[8] Morales (2005) shows that no pure-action imitation rule can lead a DM towards optimality for given, fixed population behavior. Recall that in Proposition 1 the rule is employed by the population as a whole.

randomly select another individual among those observed and imitate her action with probability equal to her payoff. Otherwise, select a new individual. That is, randomly select without replacement among the individuals observed, imitate their action with a probability equal to the payoff, and stop once another DM is imitated. Do not change action if the last selected individual is not imitated.[9]

For $M = 1$, IBA and IB reduce to Imitate If Better, while SPOR reduces to POR Note also that when only two payoffs are possible, say 0 and 1, then SPOR and IB yield equivalent behavior. The only difference concerns who is imitated if there is a tie, but this does not influence the overall expected change in play.

Consider, first, idiosyncratic noise. Schlag (1996) shows that IB is improving, but IBA is not. To provide some intuition for this result, consider the particular case where there are only two actions, 1 and 2, which yield the same expected payoff. Further, suppose that noise is symmetric and takes only two values. Suppose M is large, and consider a group of M DMs in which one of them chooses action 1 and all the other $M - 1$ choose action 2. We will investigate net switching between the two actions within this group. If a rule is improving, then net switching must be equal to 0. Suppose all DMs use IBA. The average payoff of action 2 will most likely be close to its mean if M is large. Action 1 achieves the highest payoff with probability equal to $\frac{1}{2}$, in which case all individuals switch to action 1 unless of course all individuals choosing action 2 also attained the highest payoff. On the other hand, action 1 achieves the lowest payoff with probability $\frac{1}{2}$, in which case the individual who chose action 1 switches. Thus, it is approximately equally likely that all DMs end up choosing action 1 or action 2. It is important to take into account the number of individuals switching. When all DMs switch to action 1, there are $M - 1$ switches; when all switch to action 2, there is just a single switch. This imbalance causes IBA not to be improving. An explicit example is easily constructed (see Schlag 1996, ex. 11). In fact, IBA has the tendency to equalize the proportions of the two actions when the difference in expected payoffs is small (Schlag 1996, thm. 12).

Now suppose that all DMs use IB. Since both actions achieve the same expected payoff, and only two payoffs are possible, then the net switching is zero, because IB coincides with SPOR with only two possible payoffs. It is equally likely for each individual to achieve the highest payoffs. In $M - 1$ cases some individual choosing action 2 achieves the highest payoff, in which case only one individual switches, while in one case it is the individual who chose action 1, and $M - 1$ individuals switch. The latter becomes rare for large M. On average the net switching is zero. Intuitively, idiosyncratic payoffs allow one to infer which action is best by just looking at maximal payoffs.

[9] The basic definition of SPOR is due to Schlag (1999) and Hofbauer and Schlag (2000), except that they did not include the stopping rule.

Consider now bounded payoffs, i.e. payoffs that are known to be contained in a bounded closed interval which we renormalize to $[0, 1]$. The following proposition summarizes the results for the three considered rules.

Proposition 3. *Neither IBA nor IB is improving. SPOR is improving, with*

$$p_i' = p_i + \left(1 + (1 - \bar{\pi}) + \ldots + (1 - \bar{\pi})^{M-1}\right)(\pi_i - \bar{\pi})\, p_i$$

$$= p_i + \frac{1 - (1 - \bar{\pi})^M}{\bar{\pi}}(\pi_i - \bar{\pi})\, p_i. \tag{5}$$

The reason why neither IBA nor IB is improving is, as in the case of Imitate If Better, their unwillingness to adjust to small changes in payoffs. To see this, consider again the example where P satisfies $P_1(1) = 1 - P_1(0) = \lambda$ and $P_2(x) = 1$ for some given $x \in (0, 1)$. Consider a group of $M + 1$ DMs seeing each other, where one is choosing action 1 while the other M are choosing action 2. Then behavior is the same under IB as under IBA, $F(2, 1, 2, \ldots, 2)_1 = \lambda$ and $F(1, 2, \ldots, 2)_2 = 1 - \lambda$. The net switching in this group from action 2 to action 1 is equal to $M\lambda - (1 - \lambda)$ which clearly does not reflect which action is best. Notice, however, that improving requires that $\pi_1 > (<)\pi_2$ implies $MF(2, 1, 2, \ldots, 2)_1 \geq (\leq)F(1, 2, \ldots, 2)_2$ when $p_2 < 1$ but $p_2 \approx 1$.

As an illustration, and still within this example, consider SPOR when $M = 2$. We find that $F(2, 1, 2)_1 = \frac{1}{2}\lambda + \frac{1}{2}(1 - x)\lambda$ and $F(1, 2, 2)_2 = x + (1 - x)x$. So net switching from 2 to 1 is equal to $2F(2, 1, 2)_1 - F(1, 2, 2)_2 = (2 - x)(\lambda - x)$ which is ≥ 0 if $\pi_1 = \lambda \geq \pi_2 = x$, which ensures the necessary condition for being improving.

The expression (5) for SPOR is derived in Hofbauer and Schlag (2000). If the first sampled DM is imitated, then it is as if POR is used, and hence $p_i' = p_i + (\pi_i - \bar{\pi})p_i$. However, if the first DM is not imitated, in practice she receives a further chance of being imitated, hence $(\pi_i - \bar{\pi})\, p_i$ is added on again, only discounting for the probability $1 - \bar{\pi}$ of getting a second chance, etc. Note that when $\bar{\pi}$ is small, then the growth rate of SPOR is approximately M times that of single sampling, while when $\bar{\pi}$ is large, the growth rate depends only marginally on M. Given (5), it is clear that SPOR is improving. Note also that the growth rate of SPOR is increasing with M. The dynamics that obtains as M tends to infinity is called the *adjusted replicator dynamics*, due to the scaling factor in the denominator (Maynard Smith 1982).

For $M \geq 2$, there are other strictly improving rules. Indeed, unlike in the single sampling case, a unique class of best rules cannot be selected as under single sampling. An interesting research agenda would be to identify appropriate criteria for selecting among improving rules with multiple sampling.

We conclude with some additional comments on IB and IBA. First, note that IB is improving if only two payoffs are possible. This follows immediately from our previous observation that SPOR and IB yield the same change in average payoffs whenever payoffs are binary valued.

Second, for any given p and any given decision problem, if M is sufficiently large, then $\bar{\pi}' \geq \bar{\pi}$ under IBA. This is because, due to the law of large numbers, most individuals will switch to a best action. In fact, this result can be established uniformly for all decision problems with $|\pi_1 - \pi_2| \geq d$ for a given $d > 0$. This statement does not hold for IB, though. For sufficiently large M, and provided distributions are discrete, the dynamics under IB are solely driven by the largest payoff in the support of each action, which need not be related to which action has the largest mean.

11.2.7 Correlated Noise

It is natural to imagine that individuals who observe each other also face similar environments. This leads us to consider environments where payoffs are no longer realized independently. Instead, they can be correlated. We refer to this as *correlated noise*, as opposed to independent noise.

Consider the following model of correlated noise. There is a finite number Z of states. Let $q_{\alpha\beta}$ be the probability that a DM is in state α and observes a DM in state β, with $q_{\alpha\beta} = q_{\beta\alpha}$. States are not observable by the DMs. The probability that a given DM is in state α is given by $q_\alpha = \sum_\beta q_{\alpha\beta}$. Apart from the finiteness of states, the previous case with independent payoffs corresponds to the special case where $q_{\alpha\beta} = q_\alpha q_\beta$ for all α, β. Another extreme is the case with *perfectly correlated noise*, where $q_{\alpha\beta} > 0$ implies $\alpha = \beta$. Let $\pi_{i,\alpha}$ be the deterministic payoff achieved by action i in state α. So $\pi_i = \sum_\alpha q_\alpha \pi_{i,\alpha}$ is the expected payoff of action i. Consider first the setting of bounded payoffs.

Proposition 4. *Assume that payoffs $\pi_{i,\alpha}$ are known to be contained in [0, 1]. If $M = 1$, then the improving rules under independent noise are also improving under correlated noise. If $M \geq 2$, then there exists $(q_{\alpha\beta})_{\alpha\beta}$ such that SPOR is not improving.*

To understand this result, consider first $M = 1$, and let F be improving under independent noise. Then net switching between action i and action j is given by

$$\sum_{\alpha\beta} q_{\alpha\beta}(F(i, \pi_{i,\alpha}, j, \pi_{j,\beta})_j - F(j, \pi_{j,\beta}, i, \pi_{i,\alpha})_i)$$

$$= \sum_\alpha q_{\alpha\beta}\sigma_{ij}(\pi_{j,\alpha} - \pi_{i,\alpha}) = \sigma_{ij}(\pi_j - \pi_i),$$

and hence F is also improving under correlated noise. Clearly, it is the linearity of net switching behavior in both payoffs that ensures improving under correlated noise. Analogously, it is precisely the lack of linearity when $M \geq 2$ that leads to a violation of the improving condition for SPOR. For, let $M \geq 2$, and consider a group of $M + 1$ DMs in which one DM chooses action 1 and the rest choose action

2. Suppose $q_{\alpha\alpha} = q_{\beta\beta} = \frac{1}{2}$, $\pi_{1,\alpha} = 0$, and $\pi_{2,\alpha} = 1$. The DM who chose action 1 is the only one who switches, thus the net switching from action 1 to action 2 in state α is $s_{12}(\alpha) = 1$. If in state β we have $\pi_{1,\beta} = 1$ and $\pi_{2,\beta} = 0$, thus $s_{12}(\beta) = -M$. Consequently, the expected net switching from action 1 to action 2 is equal to $\frac{1}{2}(1 - M)$, which means that SPOR is not improving.

Idiosyncratic payoffs can be embedded in this model with correlated noise. In fact, we can build a model such that a first-order stochastic dominance relationship emerges by assuming that the order of actions is the same in each state. Specifically, we say that payoffs *satisfy common order* if for all α, β it holds that $\pi_{i,\alpha} \geq \pi_{j,\alpha} \Leftrightarrow \pi_{i,\beta} \geq \pi_{j,\beta}$. The same calculations used in Section 11.2.1 (see Eq. 2) then show that if payoffs are known to satisfy common order, then IIB is improving.

We now turn to population dynamics. In order to investigate the dynamics in a countably infinite population, one has to specify how the noise is distributed within the population. The resulting population dynamics can be deterministic or stochastic. We discuss these possibilities and provide some connections to the literature.

Consider first the case where the proportion of individuals in each state is distributed according to $(q_\alpha)_{\alpha \in A}$. Then the population dynamics is deterministic, and we obtain the equivalence of improving with $\sigma_{ij} > 0$ and global interior learning (see Proposition 2). The special case where payoffs are perfectly correlated is closely related to Ellison and Fudenberg (1993, sect. II), which we now briefly present. In their model there are two actions, payoffs are idiosyncratic, in each period a fraction of the population is selected and receives the opportunity to change their action, and an agent revising observes the population average payoffs of each action and applies IBA (i.e. it is as if they had an infinite sample size, so $M = \infty$). The consequence is that the time averages of the proportion of DMs using each action converge to the probability with which those actions are best. That is, global learning does not occur, since in general the worst action can outperform the best one with positive probability due to noise. As stated above, global learning occurs when $M = 1$ and all use IB. Thus it is the ineffectiveness of the learning rule combined with the information available that prevents global learning.

Ellison and Fudenberg (1993, sect. III) enrich the model by adding popularity weighting (a form of payoff-independent imitation) to the decision rule. Specifically, a DM who receives revision opportunity chooses action 1 if $\pi_1 + \epsilon_1 \geq \pi_2 + \epsilon_2 + m(1 - 2p_1)$, where m is a popularity parameter. That is (for $m > 0$), if action 1 is "popular" ($p_1 > \frac{1}{2}$), then it is imitated even if it is slightly worse than action 2, and vice versa. Then they assume further that the distribution of the payoff shock $\epsilon_1 - \epsilon_2$ is uniform on an interval $[-a, a]$ and find that global learning (convergence to the best action with probability 1) requires $a - \Delta\pi \leq m \leq a + \Delta\pi$,

where $\Delta\pi = \pi_1 - \pi_2$.[10] In contrast, if payoffs are bounded, and all use the rule described in Section 11.2.4, then global learning emerges under the same informational conditions: agents revising know average payoffs of each action and population frequencies.

Now assume that all DMs are in the same state in each period. Then the dynamic is stochastic, and convergence is governed by logarithmic growth rates. Ellison and Fudenberg (1995) consider such a model with idiosyncratic and aggregate shocks, where DMs apply IBA to a sample of M other agents. For $M = 1$, IBA is equivalent to IB (up to ties which play no role). Global learning occurs when the fraction of DMs who receive revision opportunities is small enough and the two actions are not too similar. The intuition is that in this case the ratio of the logarithmic growth rates has the same sign as the difference in expected values. For $M \geq 2$, global learning no longer necessarily obtains when few DMs revise and actions are sufficiently different. Given the relationship between expected change and logarithmic growth rates mentioned above, this inefficiency is due to the fact that IBA is not improving under idiosyncratic payoffs.

11.3 Game Playing

We now turn to imitation and learning in strategic environments. While above we were interested in whether the population would learn which action is best, we are now interested in whether play approaches a Nash equilibrium or another suitable convention.

11.3.1 Imitation of Kin, Play against Others

There is a straightforward way to translate the framework we considered above to game-playing and thus allow us to investigate imitation in general (here two-person) games.

Consider a two-player game with two actions for each player. Associate a different population to each player role in the game. Each round, individuals are randomly matched in pairs, one from each population, to play the game. Between rounds,

[10] Juang (2001) studies the initial conditions under which an evolutionary process on rules will lead the population to select popularity weighting parameters ensuring global learning. In a society with two groups of agents, these conditions require that either one group adopts the optimal parameter from the beginning, or the optimal parameter lies between those of both groups. That is, "a society does not have to be precise to learn efficiently, as long as the types of its members are sufficiently diverse" (Juang 2001, p. 735).

individuals observe the choice and success of someone else within the same population. That is, there is explicitly no strategic interaction between individuals who observe each other. Individuals belonging to the same population are *ex ante* in identical situations, and hence imitation can be useful. Further, from the point of view of an individual, the other population's play can be viewed as inducing a distribution of outcomes for each possible (own) action. That is, we are (myopically) back in the framework considered above.

Suppose that all individuals in the same population use the same rule. The improving condition becomes the objective to increase average payoffs in one population, for all distributions of own and opponent play and for all games, *under the (mistaken) belief that the opponent population's behavior does not change.* Equivalently, the rule must induce a better reply dynamics in each population. Thus we need no separate analysis of which rules are selected and can proceed immediately to the analysis of dynamics.

Consider the dynamics which result when each population uses a single rule: namely, a strictly improving single sampling rule, or SPOR, when $M \geq 2$. It is easily shown that if play starting in the interior converges, then it converges to a Nash equilibrium. This is, of course, unsurprising, since the rules aim to increase individual payoffs and the multi-population setting abstracts from strategic effects. If there is no expected movement, then each action chosen yields the same expected payoff. For an interior initial state, actions not chosen in the limit must achieve lower expected payoff (see Weibull 1995 for a formal argument). However, trajectories need not converge to a single point.

By a standard argument, reducing the time between rounds and the proportion of individuals that observe someone else between rounds, one obtains a continuous-time dynamics. Strictly improving rules under single sampling induce the standard replicator dynamics of Taylor (1979). Convergence from an interior starting point to a Nash equilibrium holds in all types of 2×2 games except in those that have best reply structure, as in Matching Pennies. In this Matching Pennies type of game the replicator dynamics are known to cycle forever around the interior Nash equilibrium. Hofbauer and Schlag (2000) show that the dynamics under SPOR with $M \geq 2$ starting in the interior converge to the Nash equilibrium. Observing two others is sufficient to lead the population to the Nash equilibrium from an interior initial state in all 2×2 games. However, this result should not be expected to hold in all more general games. For instance, it will not hold in the generalized Shapley game of Hofbauer and Swinkels (1995; cf. Balkenborg and Schlag 2007) that involves a $2 \times 2 \times 2$ game among three populations.

Of course, the true dynamics is not continuous but discrete, driven by the jumps associated with a strictly positive proportion of individuals changing actions. Hofbauer and Schlag (2000) investigate the discrete dynamics induced by SPOR and show that the Nash equilibrium of Matching Pennies is repellent for all M. Under single sampling the population state spirals outwards to the boundary. When

$M \geq 2$, then the dynamics will circle close to and around the Nash equilibrium if sufficiently few individuals observe the play of others between rounds.[11]

Pollock and Schlag (1999) consider individuals who know the game they play, so uncertainty is only about the distribution of actions. They investigate conditions on a single sampling rule that yield a payoff monotone dynamics in a game that has a cyclic best response structure as in Matching Pennies. They find that the rule has to be imitating, and that the continuous version of the population dynamics will have—like the standard replicator dynamics—closed orbits around the Nash equilibrium. They contrast this with the finding that there is no rule based only on a finite sample of opponent play that will lead to a payoff monotone dynamics. This is due to the fact that information on success of play has to be stored and recalled in order to generate a payoff monotone dynamics.

Dawid (1999) considers two populations playing a battle-of-the-sexes game, where each agent observes a randomly selected other member of the same population and imitates the observed action if the payoff is larger than their own and the gap is large enough. For certain parameter values, this model includes PIR. The induced dynamics is payoff monotone. In games with no risk-dominant equilibrium, there is convergence towards one of the pure-strategy coordination equilibria unless the initial population distribution is symmetric. In the latter case, depending on the model's parameters, play might converge either to the mixed-strategy equilibrium or to periodic or complex attractors. If one equilibrium is risk-dominant, it has a larger basin of attraction than the other one.

11.3.2 Imitating your Opponents

In the following we consider the situation where player roles are not separated. There is a symmetric game, and agents play against and learn from agents within the same population. Environments where row players cannot be distinguished from column players include oligopolies and financial markets. Here it makes a difference whether we look for rules that increase average payoffs or those that induce a better reply dynamics.

Consider, first, the objective to induce a better reply dynamics. Rules that we characterized as being improving in decision problems have this property. To induce a (myopic) better reply dynamic means that, if play of other agents does not change, an individual agent following the rule should improve payoffs. Thus this condition is identical with the improving condition for decision problems. Specifically, a rule induces a better reply dynamic if and only if it is improving in decision problems. The condition of bounded payoffs translates into considering the set of all games with payoffs within these bounds. The decision setting with

[11] Cycling can have a descriptive appeal, for such cycles might describe fluctuations between costly enforcement and fraud (e.g. see Cressman *et al.* 1998).

idiosyncratic payoffs translates into games where all pure strategies can be ordered according to dominance.

Now turn to the objective of finding a rule that always increases average payoffs. Ania (2000) presents an interesting result showing that this is not possible unless average payoffs remain constant. The reason is as follows. When a population of players is randomly matched to play a Prisoner's Dilemma, in a state with mostly cooperators and only a few defectors, increase in average payoffs requires that more defectors switch to cooperate than vice versa. However, note that the game might just as well not be a Prisoner's Dilemma, but one in which mutual defection yields a superior payoff to mutual cooperation. Then cooperators should switch more likely to defect than vice versa. Note that the difference between these two games does not play a role when there are mostly cooperators, and hence the only way to solve the problem is for there to be no net switching. Thus, the strategic framework is fundamentally different from the individual decision framework of, for example, Schlag (1998).

Given this negative result, it is natural to investigate directly the connection between imitation dynamics and Nash equilibria. The following dynamics, which we will refer to as the *perturbed imitation dynamics*, has played a prominent role in the literature. Each period, players receive revision opportunities with a given, exogenous probability $0 < 1 - \delta \leq 1$; that is, δ measures the amount of inertia in individual behavior. When allowed to revise, players observe either all or a random sample of the strategies used and payoffs attained in the last period (always including their own) and use an imitation rule, e.g. Imitate the Best. Additionally, with an exogenous probability $0 < \epsilon < 1$, players mutate (make a mistake) and choose a strategy at random, all strategies having positive probability. Clearly, the dynamics is a Markov chain in discrete time, indexed by the mutation probability. The "long-run outcomes" (or stochastically stable states) in such models are the states in the support of the (limit) invariant distribution of the chain as ϵ goes to zero. See Kandori *et al.* (1993) or Young (1993) for details.

The first imitation model of this kind is due to Kandori *et al.* (1993), who show that when N players play an underlying two-player, symmetric game in a round-robin tournament, the long-run outcome corresponds to the symmetric profile where all players adopt the strategy of the risk-dominant equilibrium, even if the other pure-strategy equilibrium is payoff-dominant. A clever robustness test was performed by Robson and Vega-Redondo (1996), who show that when the round-robin tournament is replaced by random matching, the perturbed IB dynamics leads to payoff-dominant equilibria instead.

We concentrate now on proper N-player games. When considering imitation in games, it is natural to restrict attention to symmetric games: that is, games where the payoff of each player k is given through the same function $\pi(s_k|s_{-k})$, where s_k is the strategy of player k, s_{-k} is the vector of strategies of other players, all strategy spaces are equal, and $\pi(s_k|s_{-k})$ is invariant to permutations in s_{-k}.

The consideration of N-player, symmetric games immediately leads to a departure from the framework in the previous sections. First, DMs imitate their opponents, so that there is no abstracting away from strategic considerations. Second, the population size has to be N; that is, we are dealing with a finite population framework, and no large population limit can be meaningfully considered for the resulting dynamics.

It turns out that the analysis of imitation in N-player games is tightly related to the concept of finite population ESS (Evolutionarily Stable Strategy), which is different from the classical infinite population ESS. This notion was developed by Schaffer (1988). A finite population ESS is a strategy such that, if it is adopted by the whole population, then any single deviant (mutant) will fare worse than the incumbents after deviation. Formally, it is a strategy a such that $\pi(a|b, a, \overset{N-2}{\ldots}, a) \geq \pi(b|a, \overset{N-1}{\ldots}, a)$ for any other strategy b. An ESS is strict if this inequality is always strict. Note that, if a is a finite population ESS, the profile $(a, \ldots, a)$ does not need to be a Nash equilibrium. Instead of maximizing the payoffs of any given player, an ESS maximizes relative payoffs—the difference between the payoffs of the ESS and those of any alternative "mutant" behavior.[12]

An ESS a is (strictly) globally stable if

$$\pi(a|b, \overset{m}{\ldots}, b, a, \overset{N-m-1}{\ldots}, a.)(>) \geq \pi(b|b, \overset{m-1}{\ldots}, b, a, \overset{N-m}{\ldots}, a)$$

for all $1 \leq m \leq N-1$; that is, if it resists the appearance of any fraction of such experimenters. We obtain:

Proposition 5. *For an arbitrary, symmetric game, if there exists a strictly globally stable finite population ESS a, then $(a, \ldots, a)$ is the unique long-run outcome of all perturbed imitation dynamics where the imitation rule is such that actions with maximal payoffs are imitated with positive probability and actions with worse payoffs than one's own are never imitated, e.g. IB or PIR.*

Alós-Ferrer and Ania (2005*b*) prove this result for IB. However, the logic of their proof extends to all the rules mentioned in the statement. The intuition is as follows. If the dynamics starts at $(a, \ldots, a)$, any mutant will receive worse payoffs than the incumbents, and hence will never be imitated. However, starting from any symmetric profile $(b, \ldots, b)$, a single mutant to a will attain maximal payoffs, and hence be imitated with positive probability. Thus, the dynamics flows towards $(a, \ldots, a)$.

Schaffer (1989) and Vega-Redondo (1997) observe that, in a Cournot oligopoly, the output corresponding to a competitive equilibrium—the output level that maximizes profits at the market-clearing price—is a finite population ESS. That is, a firm deviating from the competitive equilibrium will make lower profits than

[12] An ESS may correspond to spiteful behavior, i.e. harmful behavior that decreases the survival probability of competitors (Hamilton 1970).

its competitors after deviation. Actually, Vega-Redondo's proof shows that it is a strictly, globally stable ESS. Additionally, Vega-Redondo (1997) shows that the competitive equilibrium is the only long-run outcome of a learning dynamics where players update strategies according to Imitate the Best and occasionally make mistakes (as in Kandori *et al.* 1993).

Possajennikov (2003) and Alós-Ferrer and Ania (2005*b*) show that the results for the Cournot oligopoly are but an instance of a general phenomenon. Consider any *aggregative game*, i.e. a game where payoffs depend only on individual strategies and an aggregate of all strategies (total output in the case of Cournot oligopolies). Suppose there is strategic substitutability (*submodularity*) between individual and aggregate strategy. For example, in Cournot oligopolies the incentive to increase individual output decreases, the higher the total output in the market. Define an aggregate-taking strategy (ATS) to be one that is individually optimal, given the value of the aggregate that results when all players adopt it. Alós-Ferrer and Ania (2005*b*) show the following:

Proposition 6. *Any ATS is a finite population ESS in any submodular, aggregative game. Further, any strict ATS is strictly globally stable, and the unique ESS.*

This result has a natural counterpart in the supermodular case (strategic complementarity), where any ESS can be shown to correspond to aggregate-taking optimization.[13]

As a corollary of the last two propositions, any strict ATS of a submodular aggregative game is the unique long-run outcome of the perturbed imitation dynamics with e.g. IB, hence implying the results in Vega-Redondo (1997).

These results show that, in general, imitation in games does not lead to Nash equilibria. The concept of finite population ESS, and not Nash equilibrium, is the appropriate tool to study imitation outcomes.[14] In some examples, though, the latter might be a subset of the former. Alós-Ferrer *et al.* (2000) consider Imitate the Best in the framework of a Bertrand oligopoly with strictly convex costs. Contrary to the linear costs setting, this game has a continuum of symmetric Nash equilibria. Imitate the Best selects a proper subset of those equilibria. As observed by Ania (2008), the ultimate reason is that this subset corresponds to the set of finite population ESS.[15]

[13] Leininger (2006) shows that, for submodular aggregative games, every ESS is globally stable.

[14] For the inertia-less case, this assertion depends on the fact that we are considering rules which depend only on the last period's outcomes. Alós-Ferrer (2004) shows that, even with just an additional period of memory, the perturbed IB dynamics with $\delta = 0$ selects all symmetric states with output levels between, and including, the perfectly competitive outcome and the Cournot–Nash equilibrium.

[15] Alós-Ferrer and Ania (2005*a*) study an asset market game where the unique pure-strategy Nash equilibrium is also a finite population ESS. They consider a two-portfolio dynamics on investment strategies where wealth flows with higher probability into those strategies that obtained higher realized payoffs. Although the resulting stochastic process never gets absorbed in any population profile, it can be shown that, whenever one of the two portfolios corresponds to the ESS, a majority of traders adopt

The work just summarized focuses mainly on Imitate the Best. As seen in Proposition 5, there are no substantial differences if one assumes PIR instead. The technical reason is that the models mentioned above are finite population models with vanishing mutation rates. For these models, results are driven by the existence of a strictly positive probability of switching, not by the size of this probability. Behavior under PIR is equivalent to that of any other imitative rule in which imitation takes place only when observed payoff is strictly higher than own payoff. Whether or not net switching is linear plays no role. Rules like IBA and SPOR would produce different results, though, although a general analysis has not yet been undertaken.

We would like to end this chapter by reminding the reader that our aim has been to concentrate on learning rules, and in particular imitating ones, that can be shown to possess appealing optimality properties. However, we would like to point out that a large part the literature on learning in both decision problems and games has been more descriptive. Of course, from a behavioral perspective we would expect certain, particularly simple rules like IB or PIR to be more descriptively relevant than others. For example, due to its intricate definition, we think of SPOR more as a benchmark. Huck *et al.* (1999) find that the informational setting is crucial for individual behavior. If provided with the appropriate information, experimental subjects do exhibit a tendency to imitate the highest payoffs in a Cournot oligopoly. Apesteguía *et al.* (2007) elaborate on the importance of information and also report that the subjects' propensity to imitate more successful actions is increasing in payoff differences as specified by PIR. Barron and Erev (2003) and Erev and Barron (2005) discuss a large number of decision-making experiments and identify several interesting behavioral traits which oppose payoff maximization. First, the observation of high (foregone) payoff weighs heavily. Second, alternatives with the highest recent payoffs seem to be attractive even when they have low expected returns. Thus, IB or PIR might be more realistic than IBA.

References

Alós-Ferrer, C. (2004). Cournot vs. Walras in Oligopoly Models with Memory. *International Journal of Industrial Organization*, 22, 193–217.

—— and Ania, A. B. (2005*a*). The Asset Market Game. *Journal of Mathematical Economics*, 41, 67–90.

it in the long run. The dynamics can also be interpreted as follows: each period, an investor updates her portfolio. The probability that this revision results in an investor switching from the first portfolio to the second, rather than vice versa, is directly proportional to the difference in payoffs between the portfolios. That is, those probabilities follow PIR.

——— ——— (2005*b*). The Evolutionary Stability of Perfectly Competitive Behavior. *Economic Theory*, 26, 497–516.

——— ——— and Schenk-Hoppé, K. R. (2000). An Evolutionary Model of Bertrand Oligopoly. *Games and Economic Behavior*, 33, 1–19.

Ania, A. B. (2000). *Learning by Imitation when Playing the Field*. Working Paper 0005, Department of Economics, University of Vienna.

——— (2008). Evolutionary Stability and Nash Equilibrium in Finite Populations, with an Application to Price Competition. *Journal of Economic Behavior and Organization*, 65/3, 472–88.

Apesteguía, J., Huck, S., and Oechssler, J. (2007). Imitation—Theory and Experimental Evidence. *Journal of Economic Theory*, 136, 217–35.

Balkenborg, D., and Schlag, K. H. (2007). On the Evolutionary Selection of Nash Equilibrium Components. *Journal of Economic Theory*, 133, 295–315.

Bandura, A. (1977). *Social Learning Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Banerjee, A. (1992). A Simple Model of Herd Behavior. *Quarterly Journal of Economics*, 107, 797–817.

Barron, G., and Erev, I. (2003). Small Feedback-Based Decisions and their Limited Correspondence to Description-Based Decisions. *Journal of Behavioral Decision Making*, 16, 215–33.

Bessen, J., and Maskin, E. (2007). Sequential Innovation, Patents, and Imitation. *The Rand Journal of Economics*, forthcoming.

Björnerstedt, J., and Schlag, K. H. (1996). On the Evolution of Imitative Behavior. Discussion Paper No. B–378, Sonderforschungsbereich 303, University of Bonn.

Börgers, T., Morales, A., and Sarin, R. (2004). Expedient and Monotone Rules. *Econometrica*, 72/2, 383–405.

Boylan, R. T. (1992). Laws of Large Numbers for Dynamical Systems with Randomly Matched Individuals. *Journal of Economic Theory*, 57, 473–504.

Cho, I.-K. and Kreps, D. (1987). Signaling Games and Stable Equilibria. *Quarterly Journal of Economics*, 102, 179–221.

Conlisk, J. (1980). Costly Optimizers versus Cheap Imitators. *Journal of Economic Behavior and Organization*, 1, 275–93.

Cressman, R., Morrison, W. G., and Wen, J. F. (1998). On the Evolutionary Dynamics of Crime. *Canadian Journal of Economics*, 31, 1101–17.

Cross, J. (1973). A Stochastic Learning Model of Economic Behavior. *Quarterly Journal of Economics*, 87, 239–66.

Dawid, H. (1999). On the Dynamics of Word of Mouth Learning with and without Anticipations. *Annals of Operations Research*, 89, 273–95.

Ellison, G., and Fudenberg, D. (1993). Rules of Thumb for Social Learning. *Journal of Political Economy*, 101, 612–43.

——— ——— (1995). Word of Mouth Communication and Social Learning. *Quarterly Journal of Economics*, 110, 93–125.

Erev, I., and Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning among Cognitive Strategies. *Psychological Review*, 112, 912–31.

Fudenberg, D., and Levine, D. K. (1998). *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Hamilton, W. (1970). Selfish and Spiteful Behavior in an Evolutionary Model. *Nature*, 228, 1218–20.

HOFBAUER, J., and SCHLAG, K. H. (2000). Sophisticated Imitation in Cyclic Games. *Journal of Evolutionary Economics*, 10/5, 523–43.

—— and SWINKELS, J. (1995). A Universal Shapley-Example. Unpublished MS, University of Vienna and Washington University in St Louis.

HUCK, S., NORMANN, H. T., and OECHSSLER, J. (1999). Learning in Cournot Oligopoly—An Experiment. *Economic Journal*, 109, C80–C95.

JUANG, W.-T. (2001). Learning from Popularity. *Econometrica*, 69, 735–47.

KANDORI, M., MAILATH, G., and ROB, R. (1993). Learning, Mutation, and Long Run Equilibria in Games. *Econometrica*, 61, 29–56.

KREPS, D., and WILSON, R. (1982). Reputation and Imperfect Information. *Journal of Economic Theory*, 27, 253–79.

LAKSHMIVARAHAN, S., and THATHACHAR, M. A. L. (1973). Absolutely Expedient Learning Algorithms for Stochastic Automata. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, 281–6.

LEININGER, W. (2006). Fending Off One Means Fending Off All: Evolutionary Stability in Submodular Games. *Economic Theory*, 29, 713–19.

MAYNARD SMITH, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

MORALES, A. J. (2002). Absolutely Expedient Imitative Behavior. *International Journal of Game Theory*, 31, 475–92.

—— (2005). On the Role of Group Composition for Achieving Optimality. *Annals of Operations Research*, 137, 378–97.

OYARZUN, C., and RUF, J. (2007). Monotone Imitation. Unpublished MS, Texas A&M and Columbia University.

PINGLE, M., and DAY, R. H. (1996). Modes of Economizing Behavior: Experimental Evidence. *Journal of Economic Behavior and Organization*, 29, 191–209.

POLLOCK, G., and SCHLAG, K. H. (1999). Social Roles as an Effective Learning Mechanism. *Rationality and Society*, 11, 371–97.

POSSAJENNIKOV, A. (2003). Evolutionary Foundations of Aggregate-Taking Behavior. *Economic Theory*, 21, 921–8.

ROBSON, A. J., and VEGA-REDONDO, F. (1996). Efficient Equilibrium Selection in Evolutionary Games with Random Matching. *Journal of Economic Theory*, 70, 65–92.

ROGERS, A. (1989). Does Biology Constrain Culture?. *American Anthropologist*, 90, 819–31.

SCHAFFER, M. E. (1988). Evolutionarily Stable Strategies for a Finite Population and a Variable Contest Size. *Journal of Theoretical Biology*, 132, 469–78.

—— (1989). Are Profit-Maximisers the Best Survivors?. *Journal of Economic Behavior and Organization*, 12, 29–45.

SCHLAG, K. H. (1996). Imitate Best vs Imitate Best Average. Unpublished MS, University of Bonn.

—— (1998). Why Imitate, and if so, How? A Boundedly Rational Approach to Multi-Armed Bandits. *Journal of Economic Theory*, 78, 130–56.

—— (1999). Which One Should I Imitate?. *Journal of Mathematical Economics*, 31, 493–522.

SINCLAIR, P. J. N. (1990). The Economics of Imitation. *Scottish Journal of Political Economy*, 37, 113–44.

SQUINTANI, F., and VÄLIMÄKI, J. (2002). Imitation and Experimentation in Changing Contests. *Journal of Economic Theory*, 104, 376–404.

Taylor, P. D. (1979). Evolutionarily Stable Strategies with Two Types of Players. *Journal of Applied Probability*, 16, 76–83.
Veblen, T. (1899). *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: The Macmillan Company.
Vega-Redondo, F. (1997). The Evolution of Walrasian Behavior. *Econometrica*, 65, 375–84.
Weibull, J. (1995). *Evolutionary Game Theory*. Cambridge, MA: MIT Press.
Young, P. (1993). The Evolution of Conventions. *Econometrica*, 61/1, 57–84.

CHAPTER 12

DIVERSITY

KLAUS NEHRING
CLEMENS PUPPE

12.1 Introduction

How much species diversity is lost in the Brazilian rainforest every year? Is France culturally more diverse than Great Britain? Is the range of car models offered by BMW more or less diverse than that of Mercedes-Benz? And more generally: What is diversity, and how can it be measured?

This chapter critically reviews recent attempts in the economic literature to answer this question. As indicated, the interest in a workable theory of diversity and its measurement stems from a variety of different disciplines. From an economic perspective, one of the most urgent global problems is the quantification of the benefits of ecosystem services and the construction of society's preferences over different conservation policies. In this context, biodiversity is a central concept that still needs to be understood and appropriately formalized. In welfare economics, it has been argued that the range of different life-styles available to a person is an important determinant of this person's well-being (see e.g. Chapter 15 below). Again, the question arises as to how this range can be quantified. Finally, the definition and measurement of product diversity in models of monopolistic competition and product differentiation constitute an important and largely unresolved issue since Dixit and Stiglitz's (1977) seminal contribution.

We thank Stefan Baumgärtner, Nicolas Gravel, and Yongsheng Xu for helpful comments and suggestions.

The central task of a theory of diversity is properly to account for the similarities and dissimilarities between objects. In the following, we present some basic approaches to this problem.[1]

12.2 Measures Based on Dissimilarity Metrics

A natural starting point for thinking about diversity is based on the intuitive inverse relationship between diversity and similarity: the more dissimilar objects are among themselves, the more diverse is their totality. Clearly, this approach is fruitful only to the extent to which our intuitions about (dis)similarity are more easily accessible than those about diversity. In the following, we distinguish the different concrete proposals according to the nature of the underlying dissimilarity relation: whether it is understood as a binary, ternary, or quaternary relation, and whether it is used as a cardinal or only an ordinal concept.

12.2.1 Ordinal Notions of Similarity and Dissimilarity

Throughout, let X denote a finite universe of objects. As indicated in the introduction, the elements of X can be as diverse objects as biological species, ecosystems, life-styles, brands of products, the flowers in the garden of your neighbor, etc. The simplest notion of similarity among the objects in X is the dichotomous distinction according to which two elements are either similar or not, with no intermediate possibilities. Note that in almost all interesting cases such binary similarity relations will not be transitive. Pattanaik and Xu (2000) have used this simple notion of similarity in order to define a ranking of sets in terms of diversity, as follows. A *similarity-based partition* of a set $S \subseteq X$ is a partition $\{A_1, \ldots, A_m\}$ of S such that, for each partition element A_i, all elements in A_i are similar to each other. Clearly, similarity-based partitions thus defined are in general not unique. As a simple example, consider the universe $X = \{x, y, z\}$ and suppose that x and y, as well as y and z are similar, but x and z are not similar. The singleton partition (i.e. here: $\{\{x\}, \{y\}, \{z\}\}$) always qualifies as a similarity-based partition. In addition, there are the following two similarity-based partitions in the present example: namely, $\{\{x, y\}, \{z\}\}$ and $\{\{x\}, \{y, z\}\}$. Pattanaik and Xu (2000) propose to take the minimal cardinality of a similarity-based partition of a set as an ordinal measure of its diversity.

[1] For recent alternative overviews, see Baumgärtner (2006) and Gravel (2008).

Evidently, the ranking proposed (and axiomatized) by Pattanaik and Xu (2000) is very parsimonious in its informational requirements. Inevitably, this leads to limitations in its applicability, since differential degrees of similarity often appear to have a significant effect on the entailed diversity. To enrich the informational basis while sticking to the ordinal framework, Bervoets and Gravel (2007) have considered a quaternary similarity relation that specifies which *pairs* of objects are comparably more dissimilar to each other than other pairs of objects.[2] Bervoets and Gravel (2007) axiomatize the "maxi-max" criterion according to which a set is more diverse than another if its two most dissimilar elements are more dissimilar than those of the other set.[3] One evident problem with this approach (and the ordinal framework, more generally) is that it cannot account for tradeoffs between the number and the magnitude of binary dissimilarities. Intuitively, it is by no means evident that a set consisting of two maximally dissimilar elements is necessarily more diverse than a set of many elements all of which are pairwise less dissimilar. In a recent contribution, Pattanaik and Xu (2006) introduce a relation of "dominance in (ordinal) dissimilarity" and axiomatically characterize the class of rankings that respect it. While this avoids the conclusion that two very dissimilar objects are necessarily more diverse than many pairwise less dissimilar objects, it does not help in deciding which of the two situations offers more diversity in any concrete example. In order to properly account for the relevant tradeoffs, one needs cardinal dissimilarity information, to which we turn now.

12.2.2 Cardinal Dissimilarity Metrics

In a seminal contribution, Weitzman (1992) has proposed to measure diversity based on a cardinal dissimilarity metric, as follows. Let $d(x, y)$ denote the dissimilarity between x and y, and define the *marginal diversity* of an element x at a given set S by

$$v(S \cup \{x\}) - v(S) = \min_{y \in S} d(x, y). \tag{1}$$

Given any valuation of singletons (i.e. sets containing only one element), and given any ordering of the elements $x_1, \ldots, x_m$, Eq. 1 recursively yields a diversity value

[2] Denoting the quaternary relation by Q, the interpretation of $(x, y)Q(z, w)$ is thus that x and y are more dissimilar to each other than z and w. Bossert, Pattanaik, and Xu (2003) have also considered relations of this kind and observed that the dichotomous case considered above corresponds to the special case in which Q has exactly two equivalence classes.

[3] The maximal distance between any two elements is often called the *diameter* of a set. The ranking of sets according to their diameter has also been proposed in the related literature on freedom of choice by Rosenbaum (2000). In the working paper version, Bervoets and Gravel (2007) also consider a lexicographic refinement of the "maxi-max" criterion.

$v(S)$ for the set $S = \{x_1, \ldots, x_m\}$.[4] The problem is that the resulting value in general depends on the ordering of the elements. Weitzman (1992) observes this, and shows that Eq. 1 can be used to assign a unique diversity value $v(S)$ to each set S if and only if d is an *ultrametric*, i.e. a metric with the additional property that the two greatest distances between three points are always equal.[5] To overcome the restrictiveness of Eq. 1, Weitzman (1992) has also proposed a more general recursion formula. However, the entailed diversity evaluation of sets has the counter-intuitive property that the marginal diversity of an object can *increase* with the set to which it is added (see Section 12.3.1 below for further discussion). An ordinal ranking in the spirit of Weitzman's general recursion formula has been axiomatically characterized by Bossert, Pattanaik, and Xu (2003).

The fact that the validity of Eq. 1 is restricted to ultrametrics reveals a fundamental difficulty in the general program to construct appropriate diversity measures from binary dissimilarity information (see Van Hees 2004 for further elaboration of this point). There do not seem to exist simple escape routes. For instance, ranking sets according to the average dissimilarity, i.e. $v(S) = \sum_{\{x,y\}\subseteq S} d(x, y)/\#S$, is clearly inappropriate, due to the discontinuity when points get closer to each other and merge in the limit; other measures based on the sum of the dissimilarities have similar problems. We therefore turn to an alternative approach that has been suggested in the literature.

12.3 The Multi-Attribute Model of Diversity

In a series of papers (Nehring and Puppe 2002, 2003, 2004*a*, 2004*b*), we have developed a *multi-attribute approach* to valuing and measuring diversity. Its basic idea is to think of the diversity of a set as derived from the number and weight of the different *attributes* possessed by its elements.[6] Due to its generality, the multi-attribute approach allows one to integrate and compare different proposals of how

[4] Indeed, by Eq 1. we have $v(\{x_1, \ldots, x_k\}) = \min_{i=1,\ldots,k-1} d(x_k, x_i) + v(\{x_1, \ldots, x_{k-1}\})$ for all $k = 2, \ldots, m$. Thus, given the ordering of elements, $v(\{x_1, \ldots, x_m\})$ can be recursively determined from the dissimilarity metric and the value $v(\{x_1\})$.

[5] Such metrics arise naturally, e.g. in evolutionary trees, as shown by Weitzman (1992); see Sect. 12.3.2 below for further discussion.

[6] Measures of diversity that are based (explicitly or implicitly) on the general idea of counting attributes ("features", "characteristics") have been proposed frequently in the literature; see among others, Vane-Wright, Humphries and Williams (1991); Faith (1992, 1994); Solow, Polasky and Broadus (1993); Weitzman (1998); and the volumes edited by Gaston (1996) and Polasky (2002).

diversity is based on binary dissimilarity information, and to ask questions such as "When, in general, can diversity be determined by binary information?"

12.3.1 The Basic Framework

As a simple example in the context of biodiversity, consider a universe X consisting of three distinct species: whales (wh), rhinoceroses (rh), and sharks (sh). Intuitively, judgments on the diversity of different subsets of these species will be based on their possessing different *features*. For instance, whales and rhinos possess the feature "being a mammal", while sharks possess the feature "being a fish". Let F be the totality of all features deemed relevant in the specific context, and denote by $R \subseteq X \times F$ the "incidence" relation that describes the features possessed by each object; i.e. $(x, f) \in R$ whenever object $x \in X$ possesses feature $f \in F$. A sample of elements of R in our example is thus (wh, f_{mam}), (rh, f_{mam}), and (sh, f_{fish}), where f_{mam} and f_{fish} denote the features "being a mammal" and "being a fish", respectively. For each relevant feature $f \in F$, let $\lambda_f \geq 0$ quantify the value of the realization of f. Upon normalization, λ_f can thus be thought of as the relative importance, or *weight* of feature f. The *diversity value* of a set S of species is defined as

$$v(S) := \sum_{f \in F : (x,f) \in R \text{ for some } x \in S} \lambda_f. \tag{2}$$

Hence, the diversity value of a set of species is given by the total weight of all different features possessed by some species in S. Note especially that each feature occurs at most once in the sum. In particular, each single species contributes to diversity the value of all those features that are not possessed by any already existing species.

The relevant features can be classified according to which sets of objects possess them, as follows. First are all idiosyncratic features of the above species, the sets of which we denote by $F_{\{wh\}}$, $F_{\{rh\}}$, and $F_{\{sh\}}$, respectively. Hence, $F_{\{wh\}}$ is the set of all features that are possessed exclusively by whales, and analogously for $F_{\{rh\}}$ and $F_{\{sh\}}$. For instance, sharks being the only fish in this example, $F_{\{sh\}}$ contains the feature "being a fish". On the other hand, there will typically exist features jointly possessed by several objects. For any subset $A \subseteq X$ denote by F_A the set of features that are possessed by *exactly* the objects in A; thus, each feature in F_A is possessed by all elements of A and not possessed by any element of $X \setminus A$. For instance, whales and rhinos being the only mammals in the example, the feature "being a mammal" belongs to the set $F_{\{wh,rh\}}$. With this notation, (2) can be rewritten as

$$v(S) := \sum_{A \cap S \neq \emptyset} \sum_{f \in F_A} \lambda_f. \tag{2'}$$

Intuitively, any feature shared by several objects corresponds to a similarity between these objects. For instance, the joint feature "mammal" renders whales and rhinos similar with respect to their taxonomic classification. Suppose, for the moment, that the feature of "being a mammal" is in fact the only non-idiosyncratic feature deemed relevant in our example, and let λ_{mam} denote its weight. In this case, (2) or (2′) yields $v(\{wh, sh\}) = v(\{wh\}) + v(\{sh\})$; i.e. the diversity value of whale and shark species together equals the sum of the value of each species taken separately. On the other hand, since $v(\{wh, rh\}) = v(\{wh\}) + v(\{rh\}) - \lambda_{mam}$, the diversity value of whale and rhino species together is *less* than the sum of the corresponding individual values by the weight of the common feature "mammal". This captures the central intuition that the diversity of a set is reduced by similarities between its elements.

It is useful to suppress explicit reference to the underlying description F of relevant features by identifying features *extensionally*. Specifically, for each subset $A \subseteq X$ denote by $\lambda_A := \sum_{f \in F_A} \lambda_f$ the total weight of all features with extension A, with the convention that $\lambda_A = 0$ whenever $F_A = \emptyset$. With this notation, (2′) can be further rewritten as

$$v(S) = \sum_{A \cap S \neq \emptyset} \lambda_A. \qquad (2'')$$

The totality of all features $f \in F_A$ will be identified with their extension A, and we will refer to the subset A as a particular *attribute*. Hence, a set A viewed as an attribute corresponds to the family of all features possessed by exactly the elements of A. For instance, the attribute $\{wh\}$ corresponds to the conjunction of all idiosyncratic features of whales ("being a whale"), whereas the attribute $\{wh, rh\}$ corresponds to "being a mammal".[7] The function λ that assigns to each attribute A its weight λ_A, i.e. the total weight of all features with extension A, is referred to as the *attribute weighting function*. The set of *relevant* attributes is given by the set $\mathcal{A} := \{A : \lambda_A \neq 0\}$. An element $x \in X$ *possesses* the attribute A if $x \in A$, i.e. if x possesses one, and therefore all, features in F_A. Furthermore, say that an attribute A is *realized* by the set S if it is possessed by at least one element of S, i.e. if $A \cap S \neq \emptyset$. According to (2″), the diversity value $v(S)$ is thus the total weight of all attributes realized by S.

A function v of the form (2″) with $\lambda_A \geq 0$ for all A is called a *diversity function*, and we will always assume the normalization $v(\emptyset) = 0$. Clearly, any given attribute weighting function $\lambda \geq 0$ determines a particular diversity function via formula (2″). Conversely, any given diversity function v *uniquely* determines the corresponding collection λ_A of attribute weights via "conjugate Moebius

[7] Subsets of X thus take on a double role as sets to be evaluated in terms of diversity on the one hand, and as weighted attributes, on the other. In order to distinguish these roles notationally, we will always denote generic subsets by the symbol "A" whenever they are viewed as attributes, and by the symbol "S" otherwise.

inversion".[8] In particular, any given diversity function v unambiguously determines the corresponding family Λ of relevant attributes. This basic fact allows one to describe properties of a diversity function in terms of corresponding properties of the associated attribute weighting function.

An essential property of a diversity function is that the marginal value of an element x *decreases* in the size of existing objects; formally, for all S, T and x,

$$S \subseteq T \Rightarrow v(S \cup \{x\}) - v(S) \geq v(T \cup \{x\}) - v(T). \tag{3}$$

Indeed, using (2″), one easily verifies that

$$v(S \cup \{x\}) - v(S) = \sum_{A \ni x,\, A \cap S = \emptyset} \lambda_A,$$

which is decreasing in S due to the nonnegativity of λ. Property (3), known as *submodularity*, is a very natural property of diversity; it captures the fundamental intuition that it becomes harder for an object to add to the diversity of a set the larger that set already is.[9]

Any diversity function naturally induces a notion of pairwise dissimilarity between species. Specifically, define the *dissimilarity from x to y* by

$$d(x, y) := v(\{x, y\}) - v(\{y\}). \tag{4}$$

The dissimilarity $d(x, y)$ from x to y is thus simply the marginal diversity of x in a situation in which y is the only other existing object. Using (2″), one easily verifies that

$$d(x, y) = \sum_{A \ni x,\, A \not\ni y} \lambda_A;$$

that is, $d(x, y)$ equals the weight of all attributes possessed by x but not by y. Note that, in general, d need not be symmetric, and thus fails to be a proper metric; it does, however, always satisfy the triangle inequality. The function d is symmetric if and only if $v(\{x\}) = v(\{y\})$ for all $x, y \in X$; i.e. if and only if single objects have identical diversity value.

A decision-theoretic foundation of our notion of diversity can be given along the lines developed by Nehring (1999*b*). Specifically, it can be shown that a von Neumann–Morgenstern utility function v derived from ordinal expected utility preferences over distributions of sets of objects is a diversity function, i.e. admits a nonnegative weighting function λ satisfying (2″), if and only if the underlying preferences satisfy the following axiom of "indirect stochastic dominance": a distribution of sets p is (weakly) preferred to another distribution q whenever,

[8] Specifically, one can show that if a function v satisfies (2″) for all S, then the attribute weights are (uniquely) determined by $\lambda_A = \sum_{S \subseteq A} (-1)^{\#(A \setminus S)+1} \cdot v(X \setminus S)$; see Nehring and Puppe (2002, fact 2.1).

[9] A somewhat stronger property, called *total* submodularity, in fact characterizes diversity functions; see Nehring and Puppe (2002, fact 2.2).

for all attributes A, the probability of realization of A is larger under p than under q (see Nehring 1999*b* and Nehring and Puppe 2002 for details). In this context, distributions of sets of objects can be interpreted in two ways: either as the uncertain consequences of conservation policies specifying (subjective) survival probabilities for sets of objects, or as describing (objective) frequencies of sets of existing objects, e.g. as the result of a sampling process. In terms of interpretation, different preferences over probabilistic lotteries describe different *valuations* of diversity (or, equivalently, of the realization of attributes). By contrast, different rankings of frequency distributions correspond to different ways of *measuring* diversity. The multi-attribute approach is thus capable of incorporating either the valuation or the measurement aspect of diversity.[10]

12.3.2 Diversity as Aggregate Dissimilarity

In practical applications, one will have to construct the diversity function from primitive data. One possibility is, of course, first to determine appropriate attribute weights and to compute the diversity function according to (2″). Determining attribute weights is a complex task, however, since there are as many potential attributes as there are nonempty *subsets* of objects, i.e. $2^n - 1$ when there are n objects. An appealing alternative is to try to derive the diversity of a set from the pairwise dissimilarities between its elements. This is a much simpler task since, with n objects, there are at most $n \cdot (n-1)$ nonzero dissimilarities. The multi-attribute approach makes it possible to determine precisely when the diversity of a set can be derived from the pairwise dissimilarities between its elements. The central concept is that of a "model of diversity".

A nonempty family of attributes $\mathcal{A} \subseteq 2^X \setminus \{\emptyset\}$ is referred to as a *model* (*of diversity*). A diversity function v is *compatible* with the model $\mathcal{A}$ if the corresponding set Λ of relevant attributes is contained in $\mathcal{A}$; i.e. if $\Lambda \subseteq \mathcal{A}$. A model thus represents a *qualitative* a priori restriction: namely, that no attributes outside $\mathcal{A}$ can have strictly positive weight. For instance, in a biological context, an example of such an a priori restriction would be the requirement that all relevant attributes are biological taxa, such as "being a vertebrate", "being a mammal", etc. This requirement leads to an especially simple functional form of any compatible diversity function, as follows. Say that a model $\mathcal{A}$ is *hierarchical* if, for all $A, B \in \mathcal{A}$ with $A \cap B \neq \emptyset$, either $A \subseteq B$ or $B \subseteq A$. In Nehring and Puppe (2002) it is shown that a diversity function v is compatible with a hierarchical model if and only if, for all S,

$$v(S \cup \{x\}) - v(S) = \min_{y \in S} d(x, y),$$

[10] For an argument that the measurement of diversity presupposes some form of value judgment, see Baumgärtner (2008).

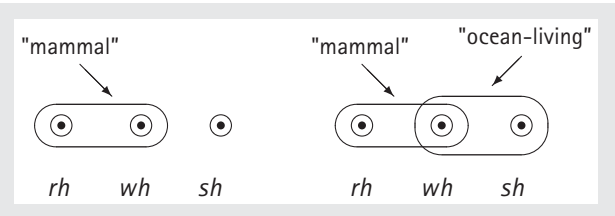

Fig. 12.1. Hierarchical versus linear organization of attributes.

where d is defined from v via (4). This is precisely Weitzman's recursion formula (1) the only difference being that no symmetry of d is required here. Thus, Weitzman's original intuition turns out to be correct exactly in the hierarchical case.[11]

A more general model that still allows one to determine the diversity of arbitrary sets from the binary dissimilarities between its elements is the *line model*. Specifically, suppose that the universe of objects X can be ordered from left to right in such a way that all relevant attributes are connected subsets, i.e. intervals. This structure emerges, for instance, in the above example once one includes the nontaxonomic attribute "ocean-living" possessed by whales and sharks (see Figure 12.1). A diversity function v is compatible with this *line model* if and only if, for all sets $S = \{x_1, \ldots, x_m\}$ with $x_1 \leq x_2 \leq \ldots \leq x_m$,

$$v(S) = v(\{x_1\}) + \sum_{i=2}^{m} d(x_i, x_{i-1}) \tag{5}$$

(see Nehring and Puppe 2002, thm. 3.2).

When, in general, is diversity determined by binary information alone? Say that a model $\mathcal{A}$ is *monotone in dissimilarity* if, for any compatible diversity function v and any S, the diversity $v(S)$ is uniquely determined by the value of all single elements in S and the pairwise dissimilarities within S, and if, moreover, the diversity $v(S)$ is a monotone function of these dissimilarities. Furthermore, say that a model $\mathcal{A}$ is *acyclic* if for no $m \geq 3$ there exist elements $x_1, \ldots, x_m$ and attributes $A_1, \ldots, A_m \in \mathcal{A}$ such that, for all $i = 1, \ldots, m-1$, $A_i \cap \{x_1, \ldots, x_m\} = \{x_i, x_{i+1}\}$, and $A_m \cap \{x_1, \ldots, x_m\} = \{x_m, x_1\}$. Thus, for instance, in the case $m = 3$, acyclicity requires that there be no triple of elements such that each pair of them possesses an attribute that is not possessed by the third element. A main result of Nehring and

[11] Another example of a hierarchical model emerges by taking the "clades" in the evolutionary tree, i.e. for any species x the set consisting of x and all its descendants, as the set of relevant attributes. For a critique of the "cladistic model" and an alternative proposal, the "phylogenetic tree model", see Nehring and Puppe (2004*b*).

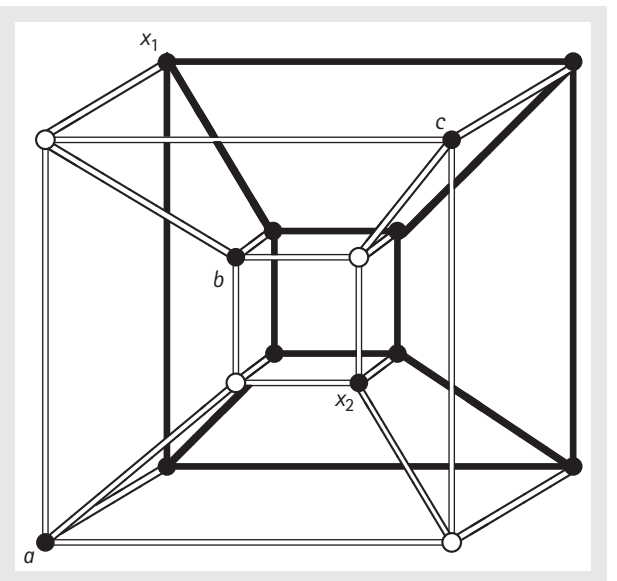

Fig. 12.2. Two metrically isomorphic subsets of the 4-hypercube.

Puppe (2002) establishes that a model of diversity is monotone in dissimilarity if and only if it is acyclic.[12]

An important example of a non-acyclic model is the *hypercube model*, which takes the set of all binary sequences of length K ("the K-dimensional hypercube") as the universe of objects and assumes all relevant attributes to be subcubes (i.e. subsets forming a cube of dimension $k \leq K$).[13] The hypercube model is clearly not acyclic (see Nehring and Puppe 2002, sect. 3.3). To illustrate the possible violations of monotonicity in dissimilarity in the hypercube model, consider the following five points in the 4-hypercube: $a = (0, 0, 0, 0)$, $b = (0, 0, 1, 1)$, $c = (1, 0, 1, 0)$, $x_1 = (0, 1, 1, 0)$ and $x_2 = (1, 0, 0, 1)$ (see Figure 12.2). If all subcubes of the same dimension have the same (positive) weight, then the dissimilarity $d(y, z)$ is uniquely determined by the Hamming distance between y and z.[14] Now consider the sets $S_1 = \{a, b, c, x_1\}$ and $S_2 = \{a, b, c, x_2\}$. The two sets are metrically isomorphic, since any element in either set has Hamming distance 2 from any other element in the same set. Nevertheless S_1 is unambiguously more diverse, since S_2 is entirely contained in the three-dimensional subcube spanned by all elements with a "0" in

[12] The necessity of acyclicity hinges on a weak regularity requirement, see Nehring and Puppe (2002, sect. 6).

[13] The hypercube model seems to be particularly appropriate in the context of sociological diversity. In this context, individuals are frequently classified according to binary characteristics such as "male vs. female", "resident vs. non-resident", etc.

[14] By definition, the Hamming distance between two points in the hypercube is given by the number of coordinates in which they differ.

the second coordinate (the white front cube in Figure 12.2). By contrast, S_1 always gives a choice between "0" and "1" in each coordinate.

12.3.3 On the Application of Diversity Theory

In the context of biodiversity a key issue is the choice of an appropriate conservation policy such as investment in conservation sites, restrictions of land development, anti-poaching measures, or the reduction of carbon dioxide emission. This can be modeled along the following lines. A policy determines at each point in time a probability distribution over sets of existing species and consumption. Formally, a policy p can be thought of as a sequence $p = (p^t)_{t \geq 0}$, where each p^t is a probability distribution on $2^X \times \mathbf{R}_+^N$ with $p^t(S^t, c^t)$ as the probability that at time t the set S^t is the set of existing species and c^t is the consumption vector. Denoting by P the set of feasible policies, society's problem can thus be written as

$$\max_{p \in P} \int_0^\infty e^{-\delta t} \cdot E_{p^t}[v(S^t) + u(c^t)]dt, \tag{6}$$

where δ denotes the discount rate and E_p the expectation with respect to p. The objective function in (6) is composed of utility from aggregate consumption $u(c^t)$, and the existence value $v(S^t)$ from the set S^t of surviving species; its additively separable form is assumed here for simplicity.

Diversity theory tries to help us determine the intrinsic value we put on the survival of different species, which is represented by the function v. The probabilities p^t reflect society's expectations about the consequences of its actions; these, in turn, reflect our knowledge of economic and ecological processes. For instance, the role of keystone species that are crucial for the survival of an entire ecosystem will be captured in the relevant probability distribution. Thus, the value derived from the presence of such species *qua* keystone species enters as an indirect rather than an intrinsic utility.[15]

As a simple example, consider two species y and z each of which can be saved forever (at the same cost); moreover, suppose that it is not possible to save both of them. Which one should society choose to save? Assuming constant consumption *ceteris paribus*, the utility gain at t from saving species x, given that otherwise the set S^t of species survives, is

$$v(S^t \cup \{x\}) - v(S^t) = \sum_{A \ni x,\, A \cap S^t = \emptyset} \lambda_A.$$

[15] Alternatively, the multi-attribute framework can also be interpreted in terms of *option value*, as explained in Nehring and Puppe (2002, p. 1168). As a result, measures of biodiversity based on that notion, such as the one proposed in Polasky, Solow, and Broadus (1993), also fit into the framework of the multi-attribute model.

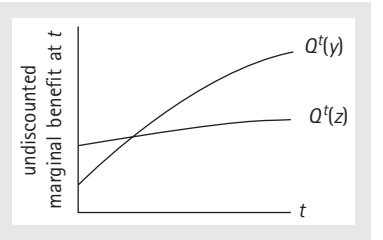

Fig. 12.3. Streams of expected marginal benefits.

Denote by $Q^t(x) := \sum_{A \ni x} \lambda_A \cdot prob(A \cap S^t = \emptyset)$ the expected marginal value at t of saving x, which is given by the sum of the weights of all attributes possessed by x multiplied by the probability that x is the unique species possessing them. The expected present value of the utility gain from saving x is given by

$$\int_0^\infty e^{-\delta t} \cdot Q^t(x) dt.$$

For concreteness, let y be one of the few species of rhinoceroses, and z a unique endemic species which currently has a sizeable number of fairly distant relatives. In view of the fact that all rhino species are currently endangered, this leads to the following tradeoff between maximizing diversity in the short run and in the long run. Saving the endemic species z yields a significant short-run benefit, while the expected benefit from safeguarding the last rhino species would be very high. This suggests the qualitative behavior of the streams of intertemporal benefits accruing from the two policies shown in Figure 12.3. The strong increase in the expected marginal value of saving y stems from the fact that, due to the limited current number of rhinos, the extinction probability of their unique attributes becomes high as t grows. Clearly, the rhino species y should be saved if the discount rate is low enough; otherwise, z should be saved. The decision thus depends on three factors: the discount rate, the value of the relevant attributes at stake, and the probability of the survival of close relatives over time.

12.4 Abstract Convexity and the Geometry of Similarity

12.4.1 Convex Models Described by Structural Similarity Relations

A key issue in applications of diversity theory is the danger of combinatorial explosion, since the number of conceivable attributes, and hence the upper bound

on the number of independent value assessments, grows exponentially with the number of objects. Nehring (1999*a*) proposes a general methodology of taming this combinatorial explosion, refining the idea of a model as a family of (potentially relevant) attributes $\mathcal{A} \subseteq 2^X \setminus \{\emptyset\}$ introduced in Section 12.3.2.

The key idea is to assume that the family of potentially relevant attributes is *patterned* in an appropriate way. Such patterning is important for two related reasons. First, excluding an isolated attribute rather than a patterned set of attributes typically does not correspond to an interpretable restriction on preferences.[16] Second, an isolated exclusion of an attribute will not capture a well-defined structural feature of the situation to be modeled.

Nehring (1999*a*) argues that an appropriate notion of pattern is given by that of an "abstract convex structure" in the sense of abstract convexity theory.[17] To motivate it, consider the case of objects described in terms of an ordered, "one-dimensional" characteristic such as mass for species or latitude for habitats. Here, the order structure motivates a selection of attributes of the form "weighs no more than 20 grams"; "weighs at least 1 ton", "weighs between 3 and 5 kilograms", that is; of *intervals* of real numbers. This selection defines the "line model" introduced in Section 12.3.2; it rules out, e.g., the conceivable attribute "weighs an odd number of grams".

Any family of relevant attributes $\mathcal{A}$ induces a natural ternary *structural similarity relation* $T_{\mathcal{A}}$ on objects as follows: y is *at least as similar to z as x is to z* if y shares all relevant attributes with z that x shares with z. In the line model, e.g., in which all attributes are intervals, the weight "5 kilograms" shares all attributes with the weight "10 kilograms" that the weight "1 kilogram" does; by contrast, the weight "1 ton" does not share all attributes common to "10 kilograms" and "1 kilogram". Likewise, in a hierarchical model in which the set of relevant attributes of species is given by biological taxonomy, a chimpanzee is at least as similar to human as a pig is, since the chimpanzee shares all taxonomic attributes with a human that a pig does.

A family of attributes can now be defined as "patterned" if it is determined by its similarity geometry $T_{\mathcal{A}}$. To do this, one can associate with any ternary relation T on X (i.e. any $T \subseteq X \times X \times X$) an associated family $\mathcal{A}_T$ by stipulating that $A \in \mathcal{A}_T$ if, for any $(x, y, z) \in T$, $\{x, z\} \subseteq A$ implies $y \in A$. A family of attributes

[16] In view of conjugate Moebius inversion (see Sect. 12.3.1 above), excluding a particular attribute A by imposing the restriction "$\lambda_A = 0$" is equivalent to a linear equality on v involving $2^{\#(X \setminus A)-1}$ terms which will lack a natural interpretation unless $\#(X \setminus A)$ is very small. In Nehring and Puppe (2004*a*) it is shown more specifically that this restriction can be viewed as a restriction on a $\#(X \setminus A)$-th order partial derivative (more properly: $\#(X \setminus A)$-th order partial difference) of the diversity function.

[17] Abstract convexity theory is a little-known field of combinatorial mathematics whose neighboring fields include lattice and order theory, graph theory, and axiomatic geometry. It is surveyed in the rich monograph by Van de Vel (1993).

$\mathcal{A}_T$ derived from some T satisfies three properties: Boundedness ($\emptyset, X \in \mathcal{A}$), Intersection Closure ($A, B \in \mathcal{A}$ implies $A \cap B \in \mathcal{A}$) and Two-Arity, to be defined momentarily. These three properties define a *convex model*. The second is the most important of the three. Translated into the language of attributes, it says that an arbitrary *conjunction* of relevant attributes is a relevant attribute. For example, if "mammal" and "ocean-living" are relevant attributes, so is the conjoint attribute "is a mammal and lives in the ocean". Note that this closure property is much more natural than closure under disjunction; for example, "is a mammal *or* lives in the ocean" is entirely artificial.[18]

The first two properties identify $\mathcal{A}$ as an abstract convex structure in the sense of abstract convexity theory (see Van de Vel 1993). In particular, the first two properties allow one to define, for any $S \subseteq X$ the (abstract) convex hull $co_{\mathcal{A}}(S) := \cap\{A \in \mathcal{A} : A \supseteq S\}$. Two-Arity says that $A \in \mathcal{A}$ whenever A contains, for any $x, y \in A$, their convex hull $co_{\mathcal{A}}(\{x, y\})$. It is easily verified that if the families $\mathcal{A}$ and $\mathcal{B}$ are convex models, so is $\mathcal{A} \cap \mathcal{B}$. It follows that for any family (model) $\mathcal{A} \subseteq 2^X \setminus \emptyset$, there exists a unique smallest superfamily $\mathcal{A}^*$ of $\mathcal{A}$ that is a convex model, the *convexity hull* of $\mathcal{A}$. Nehring (1999*a*) shows that $\mathcal{A}_{(T_{\mathcal{A}})} = \mathcal{A}^*$ for any $\mathcal{A}$; it follows that $\mathcal{A}$ is a convex model if and only if $\mathcal{A} = \mathcal{A}_{(T_{\mathcal{A}})}$. Thus convex models are exactly the models that are characterized by their associated qualitative similarity relation $T_{\mathcal{A}}$.

Structural similarity relations are characterized by transitivity and symmetry properties; symmetry in particular means that if y is at least as similar to z as x is to z, then y must also be at least as similar to x as z is to x. In view of these properties, structural similarity can be interpreted geometrically as betweenness ("y lies between x and z"). For example, structural similarity in the line model is evidently nothing but the canonical notion of betweenness on a line: y lies between x and z if and only if $x \geq y \geq z$ or $x \leq y \leq z$. A structural similarity relation can therefore be viewed as describing the *similarity geometry* of the space of objects. This endows a convex model with the desired qualitative interpretation.

12.4.2 Structural Similarity Revealed

Besides this direct conceptual significance, structural similarity relations are useful because they directly relate the structure of the support of λ to that of the diversity function itself. In the following, denote by $d(x, S) := v(S \cup \{x\}) - v(S)$ the marginal value of x at S (the "distinctiveness" of x from S). Say that x is *revealed as at least as similar to* z as y—formally, $(x, y, z) \in T_v$—if $d(x, \{y\}) = d(x, \{y, z\})$. To

[18] In a related vein, the philosopher Gärdenfors has argued in a series of papers (see e.g. Gärdenfors 1990) that legitimate inductive inference needs to be based on convex predicates.

understand the definition, note that

$$d(x, \{y\}) - d(x, \{y, z\}) = \sum_{A: x \in A, y \notin A} \lambda_A - \sum_{A: x \in A, y \notin A, z \notin A} \lambda_A = \sum_{A: x \in A, y \notin A, z \in A} \lambda_A.$$

By nonnegativity of λ, one always has $d(x, \{y\}) \geq d(x, \{y, z\})$; moreover, $d(x, \{y\}) > d(x, \{y, z\})$ if and only if a single term on the right-hand side is positive; i.e. if there exists an attribute $A \in \mathcal{A}$ that is common to x and z but not possessed by y. But this simply says that for any diversity function v the revealed similarity T_v is identical to the similarity associated with the family of relevant attributes $T_{\mathcal{A}}$,

$$T_v = T_{\mathcal{A}}.$$

This result has the following two important corollaries. The first characterizes compatibility with a convex model: for any convex model $\mathcal{A}$ and any diversity function with corresponding set $\mathcal{A}$ of relevant attributes,

$$\mathcal{A} \subseteq \mathcal{A} \Leftrightarrow T_v \supseteq T_{\mathcal{A}}. \tag{7}$$

The second corollary shows that the set of relevant attributes is revealed from T_v "up to abstract convexification": for any diversity function v, $\mathcal{A}^* = \mathcal{A}_{(T_v)}$.

The equivalence (7) is as powerful as it is simple, since it amounts to a universal characterization result for arbitrary convex models. For example, noting that for diversity functions, $(x, y, z) \in T_v$ is equivalent to the statement that $d(x, \{y\}) = d(x, S)$ for any S containing y, it allows one to deduce the line equation (5) and the hierarchy recursion (1) straightforwardly.

12.4.3 Application to Multidimensional Settings

An important application of (7) is to the characterization of multidimensional models in which X is the Cartesian product of component spaces, $X = \Pi_k X_k$; an example is the hypercube introduced in Section 12.3.2. In the context of biodiversity, multidimensional models arise naturally if diversity is conceptualized in functional, morphological,[19] or genetic, rather than, or in addition to, phylogenetic terms. In multidimensional settings, it is natural to require that any relevant attribute share this product structure as well; i.e. that $\mathcal{A} \subseteq \mathcal{A}_{sep}$, where $\mathcal{A}_{sep}$ is the set of all $A \subseteq X$ of the form $A = \Pi_k A_k$. Diversity functions with this property are called *separable*. Since $\mathcal{A}_{sep}$ is easily seen to be a convex model, the equivalence (7) can be applied to yield a straightforward characterization of separability that allows one to check

[19] The "charisma" of many organisms is closely associated with their anatomy and shape, as in the case of the horn of the rhino, the nobility of a crane, the grace of a rose, or the sheer size of a whale.

whether the restrictions on diversity values/preferences imposed by this mathematically convenient assumption are in fact reasonable. Indeed, $(x, y, z) \in T_{\mathcal{A}^{sep}}$ if and only if, for all k, $y^k \in \{x^k, z^k\}$. Thus separability amounts to the requirement that $d(x, \{y\}) = d(x, \{y, z\})$ for all x, y, z such that, for all $k \in K$, $x^k = z^k \Rightarrow y^k = x^k = z^k$.

Note the substantial gains in parsimony: while $X = \Pi_k X_k$ allows for $2^{\Pi_k \#X_k} - 1$ conceivable attributes, $\#\mathcal{A}_{sep} = \Pi_k \left(2^{\#X_k} - 1\right)$; in the case of the K-dimensional hypercube, for example, $\#\mathcal{A}_{sep} = 3^K$.

Under separability, it is further frequently natural (and mathematically extremely useful) to require *independence* across dimensions; i.e. for any $A = \Pi_k A_k$, $\lambda_A = \Pi_k \lambda^k_{A_k}$ for appropriate marginal attribute weighting functions λ^k; Nehring (1999*a*) provides simple characterizations of independence in terms of the diversity function and the underlying preference relation. Independence achieves further significant gains in parsimony, as now only $\sum_k \left(2^{\#X_k} - 1\right)$ independent attribute weights need to be determined; in the K-dimensional hypercube, for example, $3K$ such weights.

In spite of the obvious importance of multidimensional settings, to the best of our knowledge only the pioneering contributions by Solow, Polasky, and Broadus (1993) and Solow and Polasky (1994) have tried to model diversity in such settings; we do not survey their work in detail, since their measures are quite special and not well understood analytically.[20]

12.5 Absolute versus Relative Conceptions of Diversity

The literature is characterized by two competing intuitive, pre-formal conceptions of diversity that we shall term the "absolute" and the "relative". On the absolute conception, diversity is ontological richness; it has found clear formal expression in the multi-attribute model described in Section 12.3. On the relative conception, diversity is pure difference, heterogeneity. To illustrate the difference, consider the addition of some object z to the set of objects $\{x, y\}$. On the absolute conception, the diversity can never fall, even if z is a copy of x or very similar to it. By contrast, on the relative conception, the diversity may well fall; indeed, if one keeps adding

[20] The former paper represents objects as points in a finite-dimensional Euclidean space, and restricts relevant attributes to being balls in this space. The latter provides a lower bound on diversity values of arbitrary sets given the diversity values of sets with at most two elements; it also proposes taking these lower bounds as a possibly useful diversity measure based on distance information in its own right with an interesting statistical interpretation. It seems doubtful that this measure will ordinarily be a diversity function, and thus that it will admit a multi-attribute interpretation.

(near) copies of x, the resulting set would be viewed as nearly homogeneous and thus almost minimal in diversity.

In the literature, the relative conception has been articulated via indices defined on probability (i.e. relative frequency) distributions over *types* of objects. In a biological context, these types might be species, and the probability mass of a species may be given by the physical mass of all organisms of that species as a fraction of the total mass; in a social context, types might be defined by socioeconomic characteristics, and the probability mass of a type be given by the relative frequency of individuals with the corresponding characteristics.

Formally, let $\Delta(X)$ denote the set of all probability distributions on X, with $p \in \Delta(X)$ written as $(p_x)_{x\in X}$, where $p_x \geq 0$ for all x and $\sum_{x\in X} p_x = 1$. Thus, p_x is the fraction of the population of type $x \in X$. The support of p is the set of types with positive mass, $\operatorname{supp} p = \{x \in X : p_x > 0\}$. A *heterogeneity index* is a function $h : \Delta(X) \to \mathbf{R}$.[21] It is natural to require that h take values between 1 and $\#X$, as this allows an interpretation of "effective number of different types" (cf. Hill 1973). As developed in the literature, a heterogenity index is understood to rely on the frequency distribution over different types as the *only* relevant information; heterogeneity indices are thus required to be *symmetric,* i.e. invariant under arbitrary permutations of the p vector. This reflects the implicit assumption that all individuals are either exact copies or just different (by belonging to different types); all nontrivial similarity information among types is ruled out.

To be interpretable as a heterogeneity index, h must rank more "even" distributions higher than less even ones; formally, *Preference for Evenness* is captured by the requirement that h be quasi-concave. Note that Symmetry and Preference for Evenness imply that the uniform distribution $(\frac{1}{n}, \ldots, \frac{1}{n})$ has maximal heterogeneity.

A particular heterogeneity index h is characterized in particular by how it trades off the "richness" and the "evenness" of distributions. Roughly, richness measures how many different entities there are (with any nonzero frequency), while evenness measures how frequently they are realized. For instance, comparing the distributions $p = (0.6, 0.3, 0.1)$ and $q = (0.5, 0.5, 0)$, intuitively the former is richer while the latter is more even.

The most commonly used heterogeneity indices belong to the following one-parameter family $\{h_\alpha\}_{\alpha\geq 0}$, in which the parameter $\alpha \geq 0$ describes the tradeoff between richness and evenness:

$$h_\alpha(p) = \left(\sum_{x\in X} p_x^\alpha\right)^{\frac{1}{1-\alpha}}.$$

These indices (more properly, their logarithm) are known in the literature as "generalized" or "Renyi" entropies (Renyi 1961). Like much of the literature, we take

[21] We use this nonstandard terminology to distinguish heterogeneity indices clearly from diversity functions in terms of both their formal structure and their conceptual motivation.

these indices to have primarily ordinal meaning; the chosen cardinalization insures that uniform distributions of the form $(\frac{1}{m}, \ldots, \frac{1}{m}, 0, \ldots)$ have heterogeneity m. The class of generalized entropy indices $\{h_\alpha\}$ can be cleanly characterized axiomatically; for a nice exposition that draws on a closely related result on inequality measurement by Shorrocks (1984), see Gravel (2008).

A high α implies emphasis on frequent types, and thus a relatively strong weight on evenness over richness. Indeed, in the limit when α grows without bound, one obtains $h_\infty(p) = \frac{1}{\max_{x \in X} p_x}$; i.e. the frequency of the most frequent type determines heterogeneity completely.[22] At the other end of the spectrum ($\alpha = 0$), h_α simply counts the size of the support #(supp p): here evenness counts for nothing, and richness is everything. Besides the counting index, by far the most important in applications are the parameter values $\alpha = 1$ and $\alpha = 2$.

For $\alpha = 1$, the logarithm of $h_\alpha(p)$ (defined by an appropriate limit operation) is the Shannon–Wiener entropy, $\log_2 h_1(p) = -\sum_{x \in X} p_x \log_2 p_x$. An intuitive connection to a notion of diversity as disorder comes from its origin in coding theory, where it describes the minimum average number of bits needed to code without redundancy a randomly drawn member of the population.

For $\alpha = 2$, $h_2(p) = (\sum_{x \in X} p_x^2)^{-1}$ is an ordinal transform of the Simpson index (Simpson 1949) in the biological literature. Again, an intuitive link to some notion of heterogeneity can be established by noting that $\sum_{x \in X} p_x^2$ is the probability that two randomly and independently drawn elements of the population belong to the same class.

Despite their popularity, the conceptual foundations of generalized entropy indices remain to be clarified. We note three issues in particular. First, an important conceptual gap in the existing literature is the lack of a substantive interpretation of the parameter α. What does the parameter α represent? On what grounds should a diversity assessor choose one value of α rather than another? Could α represent a feature of the world? If so, what could that feature be? Alternatively, could α represent a feature of the assessor, a "taste" for richness versus evenness? Such a preference interpretation may be tempting for economists, especially in view of certain formal similarities to the theory of risk aversion. Note, however, that in the latter the degree of risk aversion can reasonably (if controversially) be explained, or at least related to, the speed at which the marginal (hedonic) utility decreases with income. The problem with the parameter α is the apparent lack of any such correlate; at least, no such correlate appears to have been suggested in the literature.

Second, the generalized entropy indices rely on a partitional classification of pairs of individuals as either completely identical or completely different. Intermediate degrees of similarity/dissimilarity are ruled out. But these are of evident importance for a relative conception of diversity no less than for an absolute one. In

[22] The index h_∞ is known as the Berger–Parker index (Berger and Parker 1970) in the biological literature.

applications, the need to fix a partition introduces a significant degree of arbitrariness into the measurement of heterogeneity.

Third, and perhaps most fundamentally, it is not clear whether the relative conception constitutes a fundamentally different notion of diversity, or whether it is in some way derivable from the absolute conception or, indeed, from a "diversity-free" notion altogether. An example of the latter is Weitzman's (2000) model of economically optimal crop variety in which he provides assumptions under which Shannon entropy can serve as a "generalized measure of resistance to extinction". To establish irreducibility, invocations of terms like "surprise" and "disorder" are clearly not enough.[23] While they may serve to visualize notions of (generalized) entropy, they do not establish the appropriateness of these as measures of diversity. Hill (1973, p. 428), for example, emphatically asserts that "the information-theoretic analogy is not illuminating".

In the remainder of this chapter, we sketch one way to make sense of relative diversity as derived from absolute diversity by "sampling". The sampling could represent a future evolution/survival process that selects a subset of the given set of individuals. Alternatively, the sampling may capture the diversity experienced by an embodied diversity consumer whose physical or mental eye is constrained by the limited capacity to take in and absorb the existing range of objects. For concreteness, think, for example, of a tourist on an ecotrip. Under both interpretations, the addition of a common organism may hinder the likelihood of survival (respectively of observation) of a less common one, in line with the Preference for Evenness intuition that is characteristic of the relative conception. To come up with a determinate and simple functional form, we assume a very stylized sampling process with fixed sample size, independent draws and replacement. By building on the multi-attribute model described in Section 12.3, the resulting family of indices allows one to capture nontrivial similarities in a very general manner. Furthermore, the sample size can serve as an interpretable parameter determining the richness–evenness tradeoff. The exposition is heuristic and hopes to stimulate further research in this important grey area of diversity theory.

Think of individual entities ("individuals") $y \in Y$ as described by their type $x \in X$ and a numeric label $i \in \mathbf{N}$. Thus the domain of individuals is given as $Y = X \times \mathbf{N}$. For a given set of individuals $S \subseteq Y$, it is convenient to write $S_x = S \cap (\{x\} \times \mathbf{N})$ for the subset of individuals in type x, and $q_x^S = \#S_x/\#S$ for the fraction of these individuals. Individual entities carry no diversity value of their own. That is, the diversity of S is given by the diversity of the set of extant types: $\widetilde{v}(S) = v(\{x : \#S_x \neq 0\})$, where $v : 2^X \to \mathbf{R}_+$ is represented by the attribute weighting function $\lambda \geq 0$.

[23] For an interpretation of product diversity in terms of "potential for surprise", see Baumgärtner (2004).

Now suppose that the "effective" diversity of some set S is determined by a sampling process. Specifically, assume that from the individuals in S, a fixed number of times $k \geq 1$ some individual is randomly drawn with replacement. The replacement assumption is chosen for mathematical convenience; in some settings, a sampling without replacement may be more realistic, but we believe the difference between the two scenarios to be minor in most cases. Note that, due to the assumed replacement, the sample size may well be strictly less than k. If r_T^k denotes the probability of obtaining $T \subseteq S$ as a result of sampling k times with replacement from S, then

$$v_k(S) = \sum_{T \subseteq S} r_T^k \tilde{v}(T)$$

defines the expected diversity of the sample.[24]

It is easily seen that in fact

$$v_k(S) = \sum_{A \in 2^X} \lambda_A \left(1 - (1 - \sum_{x \in A} q_x^S)^k\right); \tag{8}$$

indeed, note that $1 - \sum_{x \in A} q_x^S$ is the probability that the sampled individual does not belong to A, for a single draw; since draws are independent, the probability that some individual in the sample belongs to A is $1 - (1 - \sum_{x \in A} q_x^S)^k$.

Since the expected sampled diversity $v_k(S)$ is determined by the distribution of individuals over types given by the vector $\left(q_x^S\right)_{x \in X}$, one can think of v_k in terms of an associated heterogeneity index $h = w^{k,v}$, where, for any $p \in \Delta(X)$ with rational coefficients, $w^{k,v}(p) = v_k(S)$ for any S such that $q_x^S = p_x$ for all x; (8) yields the following simple representation in terms of an attribute weighting expression:

$$w^{k,v}(p) = \sum_{A \subseteq X} \lambda_A \left(1 - (1 - \sum_{x \in A} p_x)^k\right). \tag{9}$$

Note that, by Jensen's inequality, it follows immediately from (9) that $w^{k,v}$ is concave, hence a fortiori quasi-concave. This preference for evenness is explained naturally here by the increased chance of duplication of an individual of the same type in the sample with the prevalence of that type in the population.

Evidently, for any p, $w^{k,v}(p)$ increases with the sample size k; moreover, as the sample size becomes infinitely large, the sampled and underlying diversities become equal,

$$\lim_{k \to \infty} w^{k,v}(p) = v(\text{supp } p).$$

Thus, the sample size can be viewed as a parameter measuring the importance of rare types, thereby controlling the richness–evenness tradeoff: the larger the

[24] The exact expression for r_T^k is of no relevance; for example, r_T^k for $\#T = k$ equals $\left(\frac{1}{\#S}\right)^k k!$.

sample, the more can one take the realization of frequent types for granted, and the more rare types matter. Since v will not in general be symmetric, neither will be $w^{k,v}$; heterogeneity will thus no longer be maximized by uniform distributions. For example, if singletons have equal value, in the hierarchical model of Figure 12.1, maximization of sampled diversity entails an above-average fraction of sharks (to insure against the loss of the taxon "fish" that is uniquely realized by sharks).

It is instructive to consider the case of "zero similarity" that is implicitly assumed by the generalized entropy measures described above. This assumption can be made explicit here by taking the underlying diversity function to be the counting measure, $v(S) = \#S$ for all $S \subseteq X$. This yields the sampled diversity function $w^{k,\#}$ given by

$$w^{k,\#}(p) = \sum_{x \in X} \left(1 - (1 - p_x)^k\right).$$

The family of functions $\{w^{k,\#}\}$ has two points of intersection with the generalized entropy measures: the support count and the Simpson rule.[25]

This model of heterogeneity as sampled diversity invites generalizations. For example, instead of a fixed sample size, it would frequently be natural consider the sample size itself to be random. Inspired by (9), one can also take a more abstract route and consider indices of the form

$$h^{\phi,v}(p) = \sum_{A \subseteq X} \lambda_A \phi\left(\sum_{x \in A} p_x\right)$$

for some transformation function $\phi : [0, 1] \to [0, 1]$. Preference for Evenness is assured by concavity of ϕ; monotonicity of ϕ is not needed. An especially intriguing choice of ϕ is the entropic one $\phi = \phi_{ent}$, where $\phi_{ent}(q) = q \log q$. Since $h^{\phi_{ent},\#}(p)$ is the Shannon entropy of p, the indices $h^{\phi_{ent},v}$ can be viewed as *similarity-adjusted entropy indices.* Appealing as these look, their conceptual foundation is yet to be determined.

References

Baumgärtner, S. (2004). Diversity as a Potential for Surprise. Mimeo.

——(2006). Measuring the Diversity of What? And for What Purpose? A Conceptual Comparison of Ecological and Economic Biodiversity Indices. Mimeo.

——(2008). Why the Measurement of Species Diversity Requires Prior Value Judgments. In A. Kontoleon, U. Pascual, and T. Swanson (eds.), *Biodiversity Economics: Principles, Methods and Applications*, 293–310. Cambridge: Cambridge University Press.

[25] For the former, note that $w^{\infty,\#}(p) = \#(\text{supp } p) = h_0(p)$. For the latter, note that $w^{2,\#}(p) = 2 - \sum_{x \in X} p_x^2 = 2 - \frac{1}{w_2(p)}$; $w^{2,\#}$ thus ranks distributions in the same way as h_2 does.

Berger, W. H., and Parker, F. L. (1970). Diversity of Planktonic Foraminifera in Deep Sea Sediments. *Science*, 168, 1345–7.

Bervoets, S., and Gravel, N. (2007). Appraising Diversity with an Ordinal Notion of Similarity: An Axiomatic Approach. *Mathematical Social Sciences*, 53, 259–73.

Bossert, W., Pattanaik, P. K., and Xu, Y. (2003). Similarity of Options and the Measurement of Diversity. *Journal of Theoretical Politics*, 15, 405–21.

Dixit, A., and Stiglitz, J. (1977). Monopolistic Competition and Optimum Product Diversity. *American Economic Review*, 67, 297–308.

Faith, D. P. (1992). Conservation Evaluation and Phylogenetic Diversity. *Biological Conservation*, 61, 1–10.

—— (1994). Phylogenetic Pattern and the Quantification of Organismal Biodiversity. *Philosophical Transactions of the Royal Society*, *B*, 345, 45–58.

Gärdenfors, P. (1990). Induction, Conceptual Spaces and AI. *Philosophy of Science*, 57, 78–95.

Gaston, K. J. (ed.) (1996). *Biodiversity: A Biology of Numbers and Difference*. Oxford: Blackwell.

Gravel, N. (2008). What is Diversity?. In R. Gekker and M. van Hees (eds.), *Economics, Rational Choice and Normative Philosophy*. London: Routledge, forthcoming.

Hill, M. (1973). Diversity and Evenness: A Unifying Notation and its Consequences. *Ecology*, 54, 427–31.

Nehring, K. (1999*a*). Diversity and the Geometry of Similarity. Mimeo.

—— (1999*b*). Preference for Flexibility in a Savagian Framework. *Econometrica*, 67, 101–19.

—— and Puppe, C. (2002). A Theory of Diversity. *Econometrica*, 70, 1155–98.

—— —— (2003). Diversity and Dissimilarity in Lines and Hierarchies. *Mathematical Social Choice Sciences*, 45, 167–83.

—— —— (2004*a*). Modelling Cost Complementarities in Terms of Joint Production. *Journal of Economic Theory*, 118, 252–64.

—— —— (2004*b*). Modelling Phylogenetic Diversity. *Resource and Energy Economics*, 26, 205–35.

Pattanaik, P. K., and Xu, Y. (2000). On Diversity and Freedom of Choice. *Mathematical Social Choice Sciences*, 40, 123–30.

—— —— (2006). Ordinal Distance, Dominance, and the Measurement of Diversity. Mimeo.

Polasky, S. (ed.) (2002). *The Economics of Biodiversity Conservation*. Burlington, VT: Ashgate Publishers.

—— Solow, A., and Broadus, J. (1993). Searching for Uncertain Benefits and the Conservation of Biological Diversity. *Environmental and Resource Economics*, 3, 171–81.

Renyi, A. (1961). On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statics and Probability*, 547–61. Berkeley: University of California Press.

Rosenbaum, E. F. (2000). On Measuring Freedom. *Journal of Theoretical Politics*, 12, 405–21.

Shorrocks, A. F. (1984). Inequality Decomposition by Population Subgroups. *Econometrica*, 52, 1369–86.

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, 163, 688.

Solow, A., and Polasky, S. (1994). Measuring Biological Diversity. *Environmental and Ecological Statistics*, 1, 95–107.

Solow, A., Polasky, S., and Broadus, J. (1993). On the Measurement of Biological Diversity. *Journal of Environmental Economics and Management*, 24, 60–8.

Van de Vel, M. L. J. (1993). *Theory of Convex Structures*. Amsterdam: North-Holland.

Vane-Wright, R. I., Humphries, C. J., and Williams, P. H. (1991). What to Protect?—Systematics and the Agony of Choice. *Biological Conservation*, 55, 235–54.

Van Hees, M. (2004). Freedom of Choice and Diversity of Options: Some Difficulties. *Social Choice and Welfare*, 22, 253–66.

Weitzman, M. (1992). On Diversity. *Quarterly Journal of Economics*, 107, 363–405.

——— (1998). The Noah's Ark Problem. *Econometrica*, 66, 1279–98.

——— (2000). Economic Profitability versus Ecological Entropy. *Quarterly Journal of Economics*, 115, 237–63.

PART II

SOCIAL CHOICE AND WELFARE

This page intentionally left blank

CHAPTER 13

LIMITS OF UTILITARIANISM AS THE ETHICAL BASIS OF PUBLIC ACTION

PRASANTA K. PATTANAIK

13.1 Introduction

Though utilitarianism has been highly influential as a theory of public choice and as a theory of personal morality, it has faced searching criticism during much of its history. In welfare economics and the theory of social choice, where the utilitarian tradition was predominant for a long time, the last few decades have seen the development of distinctly nonutilitarian approaches, which have gained increasing acceptance among economists. In this chapter, I review some of the main objections that have been advanced in the literature by critics of utilitarianism as a theory of public action. Except in a tangential fashion, I shall not be concerned with utilitarianism as a theory of personal morality. Also, given limitations of space, I shall not cover the sizeable literature that analyzes axiomatically the formal structure of

I am grateful to Paul Anand, Clemens Puppe, and M. Salles for several helpful comments on an earlier draft.

utilitarianism.[1] Instead, my focus will be on certain central conceptual problems, which have come up repeatedly in critiques of utilitarianism.

The plan of this chapter is as follows. In Section 13.2, I discuss some basic features of utilitarianism. Sections 13.3, 13.4, and 13.5 outline some important limitations of utilitarianism discussed in the literature. I conclude in Section 13.6.

13.2 Some General Features of Utilitarianism

An ethical theory of public action specifies, either explicitly or implicitly, the considerations/factors that are relevant for assessing the rightness or wrongness of public actions. Though important, this specification, by itself, may not completely characterize the theory. Not only should the theory specify what factors are relevant for the ethical assessment of public actions, but, if it specifies several morally relevant factors which may conflict, it must also specify how these factors are to be weighed against each other where there is such a conflict. Thus, an ethical theory of public action can be criticized on at least three distinct grounds. First, one may believe that the theory excludes from the category of relevant factors some factors that should not be so excluded (I shall call this the problem of exclusion). Secondly, one may believe that the theory includes in the category of relevant factors some factors that should not be so included (I shall call it the problem of inclusion). Finally, even if one agrees with the specification of relevant factors as given by the theory, one may believe that the theory does not use an appropriate aggregation rule to reconcile possible conflicts between the different factors that the theory considers relevant (I shall call this the problem of aggregation).

All different versions of utilitarianism subscribe to consequentialism: namely, the ethical position that the rightness of actions must be judged exclusively by the goodness of their consequences.[2] The second important feature of utilitarianism is that, in judging the goodness of consequences, the only things that utilitarianism takes into account are the consequences for the welfare of the individuals under consideration. The welfare of an individual can be conceived in many ways. Utilitarianism takes a subjective view of an individual's welfare and identifies it with her utility. Finally, classical utilitarianism aggregates the utilities of different individuals by simply summing up the individual utilities and identifying the right action as an action that maximizes the sum of individual utilities among all feasible actions. Each of these three distinct steps involves issues of interpretation.

[1] For an excellent survey of this literature, the reader may refer to d'Aspremont and Gevers (2002).
[2] See Ch. 14 below for an axiomatic analysis of consequentialism and non-consequentialism.

First, even if we assume, as I assume in this chapter, that each public action or policy is associated with exactly one well-defined set of consequences, so that the problem of uncertainty regarding consequences is ruled out,[3] one still has to face the question of how to interpret the notion of consequences, and, depending on one's notion of consequences, consequentialism may or may not be a restrictive feature of utilitarianism. It is possible to interpret the notion of consequences with varying degrees of inclusiveness. Thus, if, in the absence of any law restricting the number of children that a couple may have, no couple decides to have more than two children, then, in one sense of the term "consequence", the consequences of having a law that restricts the number of children to no more than two for each family can be said to be the same as the consequences of not having such a law. If, however, one thinks that the freedom to determine the number of children is, in itself, morally relevant, then one can include the existence or nonexistence of such freedom in the specification of consequences. The use of this broader notion of consequence does not create any deep-rooted logical or conceptual problem so long as one makes explicit what one is doing. Similarly, if, for some reason, whether or not a particular public action is being taken by the public authority under consideration is, in itself, considered morally relevant (apart from any consequences that such action may have in the ordinary sense of the term "consequence"), then analytically one can include, in the description of the consequences of such action, the fact of the action being taken by the public authority under consideration. Thus, depending on how broadly one describes the consequences, judging the rightness of a public action exclusively by the goodness or badness of its consequences may or may not be a restrictive feature. Of course, as one continues to extend the notion of consequences for analytical purposes, one may move further and further away from the ordinary meaning of the term "consequences". Unless stated otherwise, I shall use the term "consequences" in a very broad sense (see Sen (1982, 1987*a*) for some persuasive arguments in favor of a consequentialist framework based on a broad notion of consequences).

It may be convenient to think of the consequences of public actions as social states, a social state being a complete description of the state of affairs that follows from the policies. The set of all possible social states will be denoted by $X = \{x, y \ldots\}$.

In assessing the consequences of a social policy, utilitarianism relies exclusively on information regarding the individuals' utilities. Two questions arise here. Who are the individuals whose utilities must be taken into account? Secondly, what exactly is utility? The answer to the first question is not always obvious. Even if we are considering public policies in a given political unit, say, a country, it is not clear why, from an ethical point of view, one should not consider the consequences of public

[3] For a discussion of some problems that uncertainty about consequences creates for utilitarianism, the reader may refer to Hammond (1980, 1981, 1982) and Hahn (1982).

action in this country on the citizens of another country. If the pollution generated in country A spills over into country B, then measures to control pollution in A will benefit the citizens of both countries. In this case, it is not clear that an ethical assessment of the consequences of pollution control measures in country A should confine itself to consequences for the citizens of A only. Also, insofar as the effect of public action now may have implications for the future generation as well as the present generation, one may like to include, in the relevant group of individuals, individuals belonging to the future generation as well as those belonging to the present generation. Finally, it is not obvious why one should take into account only the welfare of human beings and exclude all consideration of the welfare of nonhuman animals (see Singer 1975; 1979, pp. 48–71). I shall, however, avoid these complications here: I shall assume that we are concerned only with human beings and that the relevant group of individuals is fixed.[4] Let this fixed group of individuals be denoted by $N = \{1, \ldots, n\}$.

The notion of utility has been interpreted in various ways. For Jeremy Bentham, utility was the sensation of pleasure and the absence of the sensation of pain; but few writers now use the term in that sense. A second interpretation of "utility" runs in terms of happiness. While pleasure, or at least the absence of pain, can contribute to happiness, happiness seems to indicate a somewhat deeper and more durable mental state than the mere sensation of pleasure or absence of pain. A third sense in which the term "utility" has been used, is to indicate desire fulfillment. This is perhaps the interpretation most widely used now. It is different from the interpretation in terms of either happiness or pleasure. People desire many things because they believe, often correctly, that these things will bring them happiness; but we also often desire things that we know will not give us happiness. Nor is it true that everything that gives happiness is necessarily desired. Similarly, desire fulfillment is also very different from the sensation of pleasure. In positive economic theory, utility has a purely technical meaning: utility numbers corresponding to different options before an individual simply constitute a real-valued representation of the individual's preference ordering over the options. This, of course, begs the question of how one is to interpret the preference ordering itself. The usual interpretation of such a preference ordering runs in terms of desire: a person's preference for x over y means that the person desires x more than she desires y. Since utility numbers constitute a representation of the preference ordering, saying that the utility of x for the person is higher than the utility of y for her implies that having x would satisfy her desires to a greater extent than having y. If the utility numbers of an individual

[4] If the number of individuals is assumed to be fixed, then clearly there is no difference between maximizing the sum of utilities and maximizing the average of utilities. The ethical implications of these two procedures can be very different when we relax the assumption of a fixed number of individuals. For a discussion of some problems that arise when the size of the relevant population is variable, see Ch. 20 below.

constitute simply a representation of his preference ordering, then, of course, they have just an ordinal significance, and the comparison of utility differences, even for the same individual, would not have any meaning. In much of the literature on utilitarianism, however, utilities are treated as cardinal (the utility numbers of an individual are unique up to a positive linear transformation) and interpersonally comparable (full comparability).[5] Accordingly, I assume that we have n real-valued cardinal utility functions, $u_1, \ldots, u_n$, of the individuals in N, all defined over X, and that we have full interpersonal comparability of utilities.[6]

Finally, so far as the utility sum criterion is concerned, it is, of course, one specific way of aggregating individual utilities. Even if one confines oneself to individual utilities as the only relevant considerations, one has the option of using other aggregation rules. For example, one could use a suitably chosen function of the sum of all individual utilities and some index of inequality in the distribution of utilities. One can use the Rawlsian maximin rule that ranks social states by comparing the utility levels of worst-off individuals in different social states. Indeed, one can think of a large class of rules, with varying degrees of plausibility, for aggregating utilities. The aggregation procedure based on the summation of individual utilities is just one member of this class and needs justification.

That utilitarianism is consequentialistic in nature does not seem to me to be a particularly restrictive feature of utilitarianism. As I mentioned earlier, I shall use the notion of consequences in a very broad sense. Analytically, I do not see much difficulty in interpreting the notion broadly enough to allow incorporation in it of all the considerations that we consider to be relevant for judging the rightness of public actions or policies. In assessing the goodness of the consequences of public actions, however, utilitarianism focuses exclusively on the "quantities" of the individuals' utilities. This eschewal of all information other than that about utilities of individuals, together with the use of summation as the rule for aggregating individual utilities, constitutes the source of most difficulties with utilitarianism. Not only does utilitarianism throw away all information unrelated to the individuals' desires or preferences, it also throws away all information about the nature and sources of such desires and preferences themselves. This narrow informational base of utilitarianism generates what I have called the problems of exclusion and inclusion. The reliance on simple summation as the rule for aggregating individual utilities, in its turn, creates what I have called the problem of aggregation. In the next three sections, I consider some conspicuous difficulties in each of these three categories.

[5] For discussions of cardinal measurement of utility and different types of interpersonal comparisons of utility, see Sen (1970*a*, chs. 7 and 7*; 1979).

[6] For many purposes, cardinal utilities together with the weaker assumption of unit comparability will be enough, but I deliberately make a somewhat more demanding assumption to simplify the exposition.

13.3 The Problem of Exclusion

Many of our moral intuitions about the goodness of social states make use of information of various types, such as the information about who has how much power or freedom to do what, what has happened in the past to give someone an entitlement to something, and who has done what, etc., which do not have much to do with individual desires or preferences. By ignoring all such information, utilitarianism finds it difficult to accommodate many considerations, such as concerns about individual rights and freedom, which are usually considered morally important.

13.3.1 Individual Rights and Freedom

That strict adherence to the utilitarian criterion for public action can come into sharp conflict with some of our most cherished individual rights and freedom was recognized fairly early in the development of utilitarianism. John Stuart Mill (1859), a utilitarian himself, sought to identify, for each individual, an area of autonomy in which the society or the state should not have any say, and the principle he used to demarcate it from all other areas of life was not based on utilities. Mill wrote:

> The only part of the conduct of any one, for which he is amenable to society, is that which concerns others. For that part which merely concerns himself, his independence is, of right, absolute. Over himself, over his body and mind, the individual is sovereign.
>
> (Mill 1859: ch. 1, para. 9)

For Mill, an individual's tastes, pursuits, beliefs, and opinions, which did not involve "harm" to others, came within the area of life which concerned only the individual and over which the individual was to be the sovereign with the right to choose whatever he wanted to choose. Harm to others was not, however, defined by Mill in terms of reduction in other people's utilities:

> There are many who consider as an injury to themselves any conduct which they have a distaste for, and resent it as an outrage to their feelings . . . But there is no parity between the feelings of a person for his opinion, and the feeling of another who is offended at his holding it; no more than between the desire of a thief to take a purse, and the desire of the right owner to keep it. (Mill 1859: ch. 4, para. 11)

Thus, the part of an individual's life which "merely concerns himself" is identified on the basis of considerations other than utilities.

Since the desires of the individual with respect to matters in his protected private sphere are not to be overridden by conflicting desires of others, it is to be expected that the individual's right to freedom of choice in his personal sphere can conflict with the utilitarian ideal of maximizing the sum of utilities. What is surprising,

however, is that the existence of such rights is not compatible with one of the weakest implications of utilitarianism: namely, that if a social state (or social policy) x yields a higher level of utility to every individual in the society as compared to another social state, y, then x is better for the society than y. This is what Sen (1970*a*, 1970*b*) demonstrated in his celebrated result on the impossibility of the Paretian liberal. Sen's result is formulated in terms of a social decision rule, which, for every profile of preference orderings (defined over the set of all possible social states) of individuals in the society, specifies exactly one reflexive and connected binary social weak preference relation such that the strict preference relation corresponding to it is acyclic.[7] Sen demonstrated that if such a social decision rule satisfies the familiar Weak Pareto Criterion (i.e. the requirement that if every individual in the society prefers x to y, then the society must prefer x to y), then it must violate his condition of Minimal Liberalism. Sen's Minimal Liberalism requires that there must be at least two distinct individuals, i and j, and social states x, y, z, and w such that [$x \neq y$ and $z \neq w$] and [whenever i prefers x to y (resp. y to x), the society must prefer x to y (resp. y to x); and, whenever j prefers z to w (resp. w to z) the society must prefer z to w (resp. w to z)].[8] Minimal Liberalism captures a very weak version of Mill's notion of a protected sphere for an individual: x and y figuring in the statement of the condition are to be interpreted as two social states differing only with respect to some aspect of i's life over which the rest of the society is not allowed to have any say, and similarly in the case of z and w for j. Thus, basically what Sen shows is that even a very weak implication of utilitarianism, namely the Weak Pareto Criterion, conflicts with individuals' rights in any areas of their lives.[9] While Mill and, to some extent, Sen (1970*a*, 1970*b*) were concerned with an individual's rights with respect to his "personal" life, it is clear that the same tension between utilitarianism and individual rights can arise in the case of any other type of individual rights. In fact, the same problem can arise if individuals enjoy *de facto* freedom of choice in some areas of their lives without necessarily having a right to freedom of choice in those areas.[10]

Since it is difficult to deny the moral significance of individual rights and liberties, the possibility that they conflict with utilitarianism raises problems for utilitarianism. The question arises whether utilitarianism can justify the existence

[7] The strict preference relation P corresponding to a social weak preference relation is acyclic if and only if there do not exist social states $x_1, \ldots, x_m \in X$, such that $x_1 P x_2 \;\&\; x_2 P x_3 \;\&\; \ldots \;\&\; x_{m-1} P x_m \;\&\; x_m P x_1$.

[8] There has been some debate about whether Minimal Liberalism constitutes an appropriate way of formulating the notion of the individuals' right to freedom of choice in some areas of their lives; on this, see Sugden (1985); Gaertner, Pattanaik, and Suzumura (1992); Sen (1992); and Pattanaik (1996), among others.

[9] In a somewhat different framework, Kaplow and Shavell (2001) show that, if the social decision process satisfies a property of continuity, then the Weak Pareto Criterion conflicts with any attempt to incorporate a non-welfaristic value in the social decision process.

[10] An individual may enjoy certain *de facto* freedoms without those freedoms being backed by rights based on either laws or social conventions.

of any individual rights and liberties. One way in which utilitarians have tried to accommodate individual rights in their framework is by introducing a distinction between act utilitarianism and rule utilitarianism, and justifying rights in terms of rule utilitarianism. It is true, the argument goes, that a specific violation of a rule, say one embodied in a particular individual right, may increase aggregate utility in the long run, conforming to the rule generates more aggregate utility as compared to not having the rule. In that case the rule is justifiable in terms of maximization of aggregate utility in the long run, though a departure from the action prescribed by the rule in a specific situation may increase aggregate utility in that situation. To take an example, what is claimed is that, though, in a specific case, when other people have very strong feelings about *j*'s religion but *j* himself does not have such strong feelings, forcing *j* to convert to a particular religion may increase aggregate utility in the long run, considering all situations, aggregate utility will be higher if the right to religion is respected. It is not clear, however, in what sense a rule, such as the rule that every individual should have the right to practice the religion of her choice, maximizes aggregate utility in the long run, though a departure from the rule in a specific situation may increase aggregate utility in that situation. If, indeed, such a situation can arise under the rule, then one can think of an amended version of the rule that explicitly incorporates suitable exceptions in some of those situations where exceptions to the original rule can increase aggregate utilities. If it is possible to have such an amended version of the original rule, then the amended rule should be able to generate more aggregate utility in the long run than the original rule. If all such amendments are feasible, then, carried to its logical limit, the reasoning would suggest that the only rule that will maximize aggregate utility in the long run is also the rule that maximizes aggregate utility in every specific situation. It will not then be possible to justify any right or freedom for individuals by appealing to the distinction between maximization of aggregate utility in the long run (rule utilitarianism) and maximization of aggregate utility in each specific instance (act utilitarianism). One response to this objection may be that it is not possible for governments or public officials to operate by considering every possible case on its own merit (merit being judged in terms of aggregate utility), and that the only practical option is for public officials to use general rules. This, however, begs the question of why public officials cannot go through a case-by-case decision. One reason may be that the cost, in terms of resources, of collecting the relevant information in each individual case is high, while past experience may indicate that the gain in aggregate utility that results from going through such calculation in each specific case is usually low. Alternatively, making exceptions to the general rule in specific cases, even when justified in terms of gains in aggregate utility, may undermine the public's faith in the government and public officials, so that making some exceptions and then reverting to the original rule may not be politically feasible. Note that these arguments are based on certain empirical assumptions about the real world, and it is not clear that one

can always reconcile important individual rights with utilitarianism by appealing to such empirical circumstances.

If individual rights and liberties cannot be justified in terms of the maximization of aggregate utility, why should one consider them to be morally important? As we have seen, the justification that Mill (1859) had in mind for some individual rights and liberties appealed to the concept of a private area in an individual's life, where his choices did not do any "harm" to anybody else. Dworkin (1978, 1984) argues that a basic principle underlying utilitarianism itself provides justification for treating some rights as "trumps" against "unrestrained utilitarianism". As Dworkin points out, a basic presupposition of utilitarianism is that the preferences of different individuals must be treated symmetrically after taking into account their intensities, and that "the only reason for denying the fulfillment of one person's desire, whatever these are, is that more or intense desire must be satisfied" (cf. Dworkin 1984, p. 157). If utilitarianism denies a homosexual person the fulfillment of his desire for his sexual partner because the majority have a moral aversion to the fulfillment of the homosexual's desire, or if it goes against a Jewish person's preferences (after taking into account preference intensities) on the ground that the Nazis, who are in a majority, have a strong "political" preference against the fulfillment of the Jewish person's desires simply because he is Jewish, then utilitarianism will contradict its own basic presupposition of neutrality with respect to the different types of desires and preferences. Dworkin argues that if utilitarianism is to avoid contradicting its underlying principle of neutrality between different types of preferences after these preferences are weighted for their intensities, it must discount the majority's "political preferences" against the fulfillment of the Jewish individual's preferences and the majority's "moral preference" against the fulfillment of the homosexual individual's preferences regarding sexual partners. Equality of treatment of different "personal preferences" requires that political independence and moral independence of the individual, which the application of unrestrained utilitarianism may deny, be protected. Dworkin views rights, which grant appropriate powers and immunities to individuals, as a practical way of ensuring political and moral independence of individuals.

One difficulty in accommodating rights in the utilitarian framework is that utilitarianism takes the desires and preferences of people as given and then proceeds to evaluate all outcomes (including the consequences of the individuals' exercise of their rights and liberties) by considering the extent to which the outcomes fulfill these given desires and preferences. A strong argument for rights and freedoms, however, is that they give people the autonomy to develop their own objectives and preferences. It is by exercising her autonomy to choose from among alternative options that an individual develops her own objectives, interests, and preferences and defines her own individuality. John Stuart Mill seems to have a related point in mind when he defended individual liberties as follows:

> The human faculties of perception, judgment, discriminative feeling, mental activity, and even moral preference, are exercised only in making a choice. . . . these qualities [reasoning, judgment, discrimination, etc.] he requires and exercises exactly in proportion as the part of his conduct which he determines according to his own judgment and feelings is a large one. (Mill 1859: ch. 3, para. 4)

Thus, individual rights and freedoms may be valuable for various reasons unrelated to individual utilities, and with its exclusive concern with individual utilities, utilitarianism cannot possibly attach any weight to such value in its ethical accounting. To say this does not, of course, imply that, in the case of a conflict between rights and liberties, on the one hand, and the objective of maximizing aggregate utility, on the other, one must necessarily give priority to rights. What it implies is that, when such nonutility values of rights and freedom are taken into account, the assessment of public policies may be different from that yielded by utilitarianism.

13.3.2 Agent-Relative Responsibilities and Duties

In this chapter, we are concerned exclusively with utilitarianism as the basis of public policy rather than as a moral theory for personal conduct. Several familiar critiques of utilitarianism, as a principle for individual morality, are, however, based on examples specially constructed to illustrate that utilitarianism ignores some of our deep-rooted moral intuition about personal integrity. The question arises whether these critiques are inapplicable in our context. Consider a well-known example due to Williams (1973, pp. 97–9). George, who is opposed to biological warfare, has the option of working in a laboratory which conducts research in biological warfare and earning an income that he and his family need. If George does not take up the position, then someone else will definitely do so and will pursue the research on biological warfare with much greater zeal than George and would contribute to a greater extent to the death and destruction that such warfare brings about. Utilitarianism would hold that it would be right for George to accept the job in these circumstances. In doing so, utilitarianism would ignore one nonutility feature of the situation: namely, that by accepting the job George himself will directly promote biological warfare, and that if he does not accept the job, then *he* will not promote biological warfare. Many of us feel that this information about whether or not George himself is conducting research into biological warfare is morally relevant because "each of us is specially responsible for what *he* does, rather than for what other people do" (Williams 1973, p. 99). What is involved here is the personal integrity of George. Personal integrity requires George not to accept the job, though that would mean that someone else will do the same job with utility

consequences that may be worse. Thus, in Williams's example, the case against utilitarianism is built by appealing to a notion of agent-relative responsibility.

Goodwin (1995, ch. 4) raises an important issue here. Do notions of agent-relative duties and responsibilities have much relevance when one is considering the moral basis of public policy rather than the moral criteria for private conduct? Goodwin (1995) distinguishes between two types of duties and responsibilities. First, there are the agent's duties and responsibilities to do something, say *a*. Secondly, there are the agent's duties or responsibilities to see to it that *a* is done. Goodwin claims that, while the examples of agent-specific duties and responsibilities typically used in arguments against utilitarianism are usually of the first type, in the case of the government or public officials the duties and responsibilities are of the second type. The duty/responsibility of public officials is to see to it that people get enough to eat, but they do not have any duty or responsibility to feed the hungry themselves. The implication seems to be that, since the duties and responsibilities of governments and public officials are typically of the second type, and for such duties and responsibilities it does not matter who actually brings about the final state of affairs, agent relativity does not create any problem for utilitarianism when it is used as the principle for guiding actions of governments and public officials.

The distinction that Goodwin (1995) draws between duties and responsibilities of the first order and those of the second order is interesting. Nevertheless, the implication that he seeks to derive from it does not seem to be tenable. First, note that a country or society may face issues of integrity in a way that may not be very different from the situation that George faces in Williams's example. Consider the case where a country, *A*, faces a choice between two options with respect to a foreign country in which a group of religious fundamentalists are killing people of a different religious persuasion. One option is to send war planes to bomb strongholds of religious fundamentalists in country *B*; this will lead to the death of a small number of people, including some civilians, but will prevent the fundamentalists from inflicting terror and death on the rest of the population. The other option is not to do anything, which, in the long run, will involve a much larger number of deaths in country *B*. It is not clear that, in such situations, the only responsibilities and duties that the government in country *A* has are responsibilities or duties of the second order. Faced with the options described above, the citizens of country *A* and the government representing them may justifiably feel that, while the choice of the second option will involve a much larger number of deaths in country *B*, country *A* will not be directly responsible for such deaths in country *B*. To adapt Williams's apt summary of the moral emerging from the example involving George, one's moral intuition may be that the government of country *A* has a special responsibility for what it does in country *B* rather than for what some people in country *B* do to other people in country *B*.

13.3.3 Procedural Fairness and Some Other Nonutility Considerations

Besides rights, liberties, and integrity, there are many other moral concerns unrelated to individual utilities. People often attach much weight to procedural fairness. If the construction of a new dam will displace some people living in the area, then it seems fair that those people should be consulted before the decision is taken to construct the dam. This is so even if the government already has full information about the effect of the dam on those people's welfare. While the observance of the "due process" may have an instrumental role in increasing aggregate utility, the due process also seems to have an intrinsic value for many people.[11]

The notion of entitlement is another powerful moral concept which has very little to do with utility. Nozick's (1974) entitlement theory of distributive justice is one of the most well-known recent moral theories based on this concept, but one does not have to subscribe to Nozick's specific theory of entitlement to see the ethical appeal of the concept. The value judgments that there should be equal pay for equal work, that workers should get the product of their labor, and that it is exploitative to pay workers a wage lower than their marginal revenue product are all based on (diverse) notions of entitlement, and not on considerations of utility.

13.4 The Problem of Inclusion

I now consider whether by excluding all information other that about individual utilities, utilitarianism runs into the problem of inclusion: that is, the problem of admitting, as relevant, factors that should not be so admitted. Even if we agree that individual welfare is all that matters, should we identify an individual's welfare with the fulfillment of the individual's desires, without any consideration of the sources and nature of those desires? If we feel that one needs to discriminate between different types of desires of a person, then such discrimination clearly has to be made on the basis of information that cannot be had by considering only the (total) utilities of individuals. If all such information is declared inadmissible, then one may end up by identifying an individual's welfare with the overall fulfillment of her

[11] See Wailoo and Anand (2005) for evidence about the intrinsic value that people attach to procedural fairness in issues relating to health care. See also Anand (2003) for an interesting framework wherein considerations of procedures, rights, duties, and consequences are combined together in the specific context of health care.

desires when the fulfillment of some of these desires either should not count at all or should count less than the fulfillment of other desires.[12]

It is worth noting that utilitarians themselves have expressed reservations about identifying, in an unqualified fashion, the welfare of an individual with the fulfillment of her desires in general. Irrespective of whether they interpret utility as happiness or desire fulfillment, many utilitarians accept some qualifications to the general rule that the utility of an individual is what constitutes her welfare. Consider, for example, the qualifications that Harsanyi (1977*c*) introduces:

> It is well known that a person's preferences may be distorted by factual errors, ignorance, careless thinking, rash judgments, or strong emotions hindering rational choice, etc. Therefore, we may distinguish between a person's *explicit* preferences, i.e., his preferences as they actually *are* . . . and his 'true' preferences, i.e., his preferences as they *would* be under 'ideal conditions'.[13]

Some of the suggestions for disregarding the "explicit preferences" of individuals seem to be uncontroversial. For example, an individual's explicit preferences may be based on wrong information or ignorance. People may not want to evacuate an area despite explicit warnings of an impending disastrous cyclone because they mistakenly believe that, since there have been no disastrous cyclones in the area in living memory, the impending cyclone cannot possibly be disastrous. In such cases, very few people, including utilitarians, would like to identify the individuals' welfare with the fulfillment of their desires.

The issue of what type of preferences and desires should be taken into account when our concern is with an individual's welfare, however, raises many other difficulties that utilitarianism, with its emphasis on the fulfillment of desires in general, is poorly equipped to handle. I consider some of these problems.

(i) *Multiplicity of preferences*: The assumption that an individual has exactly one integrated system of preferences and desires over all the alternatives confronting him seems to be too simple a view of human beings. An individual may have a host of different preferences which may be in conflict with each other, and different preferences may come to the surface at different times, depending on the context. An individual typically belongs to different, possibly overlapping, groups that impose different obligations on her. Depending on the context, which may emphasize the individual's membership of one group rather than another, the individual's preferences may be different. Thus, as a member of her family, an individual may like to promote the career of her own child. On the other hand, if the individual is the head

[12] Very similar issues arise when an individual's welfare is identified with the individual's happiness rather than the fulfillment of her desires. Note that, in his well-known arguments for distinguishing between "higher pleasures" and "inferior pleasures", Mill (1861, ch. 2, pars. 4–8) was basically concerned with such issues.

[13] In fact, Harsanyi (1977*a*) would go even further and exclude "all clearly antisocial preferences, such as sadism, envy, resentment, and malice".

of an organization in which her son is an employee, she may not like to provide extra assistance to her son even in the absence of external constraints. Harsanyi (1953, 1955, 1977*b*) makes a broad distinction between subjective preferences and ethical preferences. Subjective preferences are the individual's ordinary preferences, while ethical preferences are characterized by an impersonal consideration of all individuals in the society. This is an important distinction, but it may not fully capture the richness of the different types of preferences that may reside in the same individual. There may be different types of subjective preferences and different degrees of impersonality. Consider, for example, the notion of subjective preferences. Individual *A* knows that it is not in his interest to consume a large amount of rich food. Nevertheless, when *A* is in a restaurant with a group of friends, he prefers to consume far richer food than he would if he was to have dinner by himself at home. Both types of preferences are within the broad category of what Harsanyi would call subjective preferences. Yet, they are very different, and the implications of judging alternative social states can be very different under utilitarianism, depending on whether such calculations are made on the basis of one type of subjective preference or another. This consideration is particularly problematic for those areas of welfare economics where the preferences of a consumer are directly linked to market choices either by defining preferences in terms of market choices or by making certain behavioral assumptions that rigidly link the preferences of a consumer to her market choices. If it is admitted that the consumer can have more than one type of preference, then it is not clear why moral priority must be given to those preferences that generate the market choices of the consumer, as compared to other preferences, which may not be reflected in the consumer's market choices but which are as much the consumer's preferences as the preferences reflected in the choices she makes in the marketplace.

(ii) *Endogenous desires and preferences*: Utilitarianism faces serious difficulties when individual preferences and desires are not exogenously given but are determined "endogenously" by the objective circumstances of the individual. For example, a person may not have the suitable training to appreciate the value of classical music or great literature. In the absence of such training, the person may not desire these things. One can take either these "untrained" existing preferences as the basis of social action and policy, or one can consider the individual's preferences as they would be after suitable training and cultivation. Alternatively, one can compare the happiness or desire fulfillment that the individual will have, without the goods under consideration, before the training, and the happiness or desire fulfillment that he will have with those goods after the training. One can also resort to the notion of "preference over preferences" or "preferences over alternative selves".[14] These different routes can lead to very different public policy conclusions. Such issues often come up when the policy under consideration happens to be about

[14] See Hahn (1982).

the allocation of resources to education and subsidy to fine arts, music, theatre, etc.[15] Another instance of endogenous preferences arises when an individual herself adapts her preferences in response to the environment. If certain sections of the society (e.g. women, people belonging to lower castes, racial minorities, etc.) have been habitually deprived of so much in their lives that they have learned to avoid disappointments by not desiring much,[16] then it is difficult to accept the fulfillment of their self-circumscribed desires as an index of their welfare.[17] This, in fact, constitutes a central point in Sen's (1985, 1987*b*) arguments for not identifying an individual's well-being or her living standards with either her happiness or the fulfillment of her desires. Yet another instance of endogenous preferences is the role of product advertisements in the formation of consumers' preferences, to which Galbraith (1958, ch. 11) has so pointedly drawn our attention. If the consumers' desires for goods and services are influenced in a systematic fashion by persuasive advertisements of producers, then one may have serious reservations about using the satisfaction of those desires as an accurate index of the consumer's welfare.

(iii) *Urgency of desires*: In general, utilitarianism tends to be "neutral" between different types of desires relating to different aspects of human life. Normally, a utilitarian would not give priority to one type of desire over another type of desire except insofar as one of them may be more intense than the other. A strong case can be made, however, for distinguishing those desires (e.g. the desire for food to escape hunger, the desire for protection from the elements, and the desire for treatment of illness) that are, in an objective sense, basic or urgent and those desires, such as the desire to go on a cruise or the desire to have a better house and car than one's neighbors, which may be very intense but less basic or urgent in some objective sense. When Keynes (1931) distinguished between those needs of human beings "which are absolute in the sense that we feel them whatever the situation of our fellow human beings may be" and those needs "which are relative only in that their satisfaction lifts us above, makes us superior to, our fellows", he probably had a similar consideration in mind. Scanlon (1975, 1977) distinguishes between "urgency" and the intensity of desires and preferences. He convincingly argues for giving priority to those "individual interests" which are more urgent. Scanlon's conception of individual interests is an objective conception, but even within the framework of subjective utilities itself, one can make a case for treating more "urgent" desires, as distinct from more intense desires, differently from less urgent desires (the distinction can, however, be based on some objective criterion).

Before concluding this section, it may be worth noting two points. First, some of the problems discussed above are closely related to the issue of whether the

[15] Such goods have sometimes been called "merit goods" in economics (see Musgrave 1987).

[16] See Elster (1982).

[17] As Goodwin (1995, p. 15) writes: "If you cannot get what you want you should simply revise your preferences so that you will want what you can easily get. . . . few of us would find the satisfaction of preferences chosen on that basis alone all that satisfying."

individual under consideration is the sole judge or even the best judge of his own welfare. In welfare economics there is a long and deep-rooted tradition that assumes that (a) the individual is the best judge (if not the sole judge) of his own welfare;[18] and (b) an individual desires things only to the extent that they promote her own welfare. Of course, given these two assumptions, the statement that individual i desires x more than she desires y would entail that x promotes i's welfare to a greater extent than y does. Many of the arguments (such as those based on the individual's ignorance or the endogenous nature of individual preferences) against identifying desire fulfillment with an individual's welfare reject either assumption (a) or assumption (b). Secondly, it may be argued that many of the problems can be taken care of by laundering or purifying the explicit preferences of individuals so as to move closer to their "true" preferences. There are, however, limits to how far one can go in this process of purification of explicit preferences without losing much of the substantive content of utilitarianism.

13.5 The Problem of Aggregation and Distributive Justice

Critics of utilitarianism often claim that utilitarianism ignores distributive justice and another value, equality, which is frequently associated with the notion of distributive justice. In particular, it has been argued that, even if one agrees to consider utility to be the only human good, the summation rule that utilitarianism uses to aggregate the utilities of different individuals rules out important considerations of distributive justice and equality (of course, if one does not accept the exclusive claim of utility to be the only human good, then one has to go beyond the utility-based framework itself to talk about distributive justice and equality). The criticisms of the summation rule have been directed at its general philosophical underpinning as well as its implications.

13.5.1 Some Philosophical Justifications for the Summation Rule

The utility sum criterion is typically justified by appealing to the notion of treating all individuals in an impartial fashion. Here is how Hare (1976, p. 26) visualizes such impartiality:

[18] See Overvold (1982) for a discussion of some of these issues.

> If I am trying to give equal weight to the equal interests of all the parties in a situation, I must, it seems, regard a benefit or harm done to one party as of equal value or disvalue to an equal benefit or harm done to any other party. This means that I shall promote the interests of the parties most, while giving equal weight to them all, if I maximize the total benefit over the entire population; and this is the classical principle of utility.

This, of course, is only one of many alternative conceptions that one may have about attaching equal weight to equal interests of all. If, in a two-person society, we start with a social state that gives us the vector (10, 0) of total utilities and compare it with an alternative social state that gives the vector (9, 1), then Hare's conception of equal weight for equal benefit or harm seems to imply that 1's loss of 1 unit of utility in passing from the first to the second social state should have exactly the same moral weight as 2's gain of 1 unit of utility, so that the harm and the benefit should just cancel each other, and (10, 0) should be just as good as (9, 1). One can object to this specific concept of equal weight for equal interests, however, on the ground that the positions of 1 and 2 with respect to their total utilities in the initial situation are very far from being symmetric, and therefore there is no compelling reason why the gain of one unit of utility for 2 should be considered to have the same weight as the loss of one unit of utility for 1. Rawls (1971) argues that the notion of an ethical evaluator, who views all the individuals' utilities impartially in this sense and maximizes the sum of all those utilities, is an exact analog of the notion of an individual maximizing his own utility subject to constraints imposed by his scarce resources. A utility-maximizing individual allocates his total resources over different uses so as to maximize the total utility from all the uses taken together. Similarly, the impartial evaluator of utilitarianism allocates resources to the different individuals so as to maximize the sum total of the utilities of all these individuals; the individuals have no significance other than how much they contribute to the aggregate utility. Rawls finds this objectionable because he feels that it ignores the separateness of individuals and thus represents an illegitimate extension, to the society, of the idea of decision-making by a utility-maximizing individual:

> On this conception of society separate individuals are thought of as so many different lines along which rights and duties are to be assigned and scarce means of satisfaction allocated . . . so as to give the greatest fulfillment of wants. . . . Utilitarianism does not take seriously the distinction between persons. (Rawls 1971, p. 27)

Following a route rather different from that of Hare (1976), Harsanyi (1953; 1977*a*, ch. 4; 1977*b*) derives a version of utilitarianism (actually, Harsanyi derives a version of utilitarianism where the average of all the individuals' utilities is maximized, but, given a fixed population, the maximization of average utility coincides with the maximization of the sum of utilities of all individuals). Harsanyi starts with the concept of an individual's ethical preferences over alternative social states. The ethical preferences of an individual are assumed to be his impartial evaluation of

alternative social states. The distinguishing feature of Harsanyi's contribution lies in the way in which he characterizes such impartiality. On Harsanyi's conception of impartial preference, an individual impartially prefers social state x to social state y if and only if he prefers the social state x, when there is an equal chance of his being any of the individuals in x, to the social state y, when there is an equal chance of his being any of the individuals in y (being individual j in some social state, z, denotes the prospect of having that individual's subjective features together with his objective economic and social position in social state z). For every individual j and every social state x, let (j, x) denote being individual j (with j's subjective features) in social state x, and let $L(x)$ denote the lottery which gives the same chance of being any of the n individuals (with that individual's subjective features) in social state x. Thus, $L(x)$ attaches probability $1/n$ to each of the "prizes" $(1, x), \ldots, (n, x)$. Assuming that each individual i satisfies the von Neumann–Morgenstern axioms for choice under risk with respect to such uncertain prospects, i's expected utility from $L(x)$ is $\frac{1}{n}\sum_{j=1}^{n} V_i(j, x)$, where V_i is the von Neumann–Morgenstern utility function of i. Harsanyi identifies $V_i(j, x)$ with $V_j(j, x)$, since to be individual j (with j's personality in social state x), i has to experience $V_j(x)$, which constitutes j's utility in social state x (see Harsanyi 1977*b*, p. 50). Thus, i's expected utility from $L(x)$ turns out to be $\frac{1}{n}\sum_{j=1}^{n} V_j(j, x)$, and since Harsanyi starts with the position that, for all social states x and y, individual i ethically considers x to be at least as good as y if and only if i considers $L(x)$ to be at least as good as $L(y)$, it follows that, for all social states x and y, individual i ethically considers x to be at least as good as y if and only if $\frac{1}{n}\sum_{j=1}^{n} V_j(j, x) \geq \frac{1}{n}\sum_{j=1}^{n} V_j(j, y)$; i.e. if and only if $\sum_{j=1}^{n} V_j(j, x) \geq \sum_{j=1}^{n} V_j(j, y)$.[19] Thus, the concept of a lottery that gives equal chance of being any of the individuals in a social state, which is reminiscent of, but different from, Rawls's original position, leads Harsanyi to the criterion based on utility sums, where the utilities are von Neumann–Morgenstern utilities.[20]

While the important result of Harsanyi establishes a criterion for assessing social welfare that is very similar to the traditional utilitarian criterion, the utilities of different individuals figuring in Harsanyi's criterion are von Neumann–Morgenstern utilities which provide representations of the individuals' preferences over risky prospects. Several writers (see e.g. Luce and Raiffa 1957, p. 32, and Arrow 1963, p. 10) have expressed doubts about whether the preference intensities implied by von Neumann utility numbers can be identified with the preference intensities that traditional utilitarians had in mind. There are also difficulties involving the choice of a specific von Neumann–Morgenstern utility function for each person from the class of her von Neumann–Morgenstern utility functions.[21]

[19] Recall that we are assuming the population remains the same.

[20] Following a different route, Harsanyi (1955, 1977*a*, 1977*b*) arrives at the same conclusion on the basis of certain axioms.

[21] See Sen (1977, 1980).

13.5.2 Some Implications of the Utilitarian Rule of Aggregation

To see the implications of going by the sum of utilities, consider first the simple distribution problem where we have a fixed quantity, a, of a single perfectly divisible commodity to be allocated among n individuals. Assume that the utility of each individual i depends only on the amount, a_i, of the commodity that i receives and that all individuals have identical utility functions (so that $u_1 = \ldots = u_n$) with positive and diminishing marginal utility for the commodity. Then, for $u_1(a_1) + \ldots + u_n(a_n)$ to be maximized, the marginal utilities, $u'_1(.), \ldots, u'_n(.)$, must be equal to each other, and the total amount of the commodity must be divided equally among the individuals. Thus, maximization of the sum of individual utilities in this case does imply (1) equality of the (total) utilities of different individuals, (2) equality of the marginal utilities of different individuals, and (3) an equal division of the fixed physical amount of the commodity.[22] If one believes that the total utility of a person is what constitutes her welfare/interest, then the equalization of total utilities of individuals would capture an important feature of our notion of equality. Moreover, it will not matter whether one thinks of equality in terms of equal total utilities or equal marginal utilities, though, intuitively, it is not very clear why one should think of equality in terms of marginal utilities in the first place. Equal division of the total quantity of the commodity, which is implied by the maximization of the sum of individual utilities in this case, would have its own appeal to anyone who tends to think of distributive justice in terms of the individuals' command over quantities of the commodity under consideration.

Next consider the case where we retain all the assumptions made above, but relax the assumption of identical utility functions. It is clear that, once we relax the assumption of identical utility functions, equality of total utilities no longer follows as a consequence of the maximization of the sum of individual utilities, though such maximization will still need the equalization of marginal utilities if we rule out the possibility of "corner solutions". In fact, depending on the utility functions, the total utilities of different individuals can be very different when we maximize the sum of utilities. Thus, even in the context of the very simple problem of distributing a fixed quantity of a perfectly divisible commodity, maximization of aggregate utility can lead to much inequality in the distribution of utilities unless the utility functions are assumed to be identical—an assumption which does not seem to be particularly realistic.

Sen (1973) gives an interesting example to highlight how the maximization of the sum of utilities may run into conflict with our intuition about distributive justice. Sen considers a society of two individuals, 1 and 2, where a fixed amount of income

[22] The conclusion remains basically the same if we complicate the picture by introducing several commodities, with a fixed quantity of each commodity.

is to be divided between them. Suppose that 2 has a physical handicap but 1 is able-bodied, and that, because of this, for every given level of income, 1's utility is twice that of individual 2; i.e. $u_1(w) = 2u_2(w)$ for every income level w. Then an equal division of income between 1 and 2 will give 2 exactly half of 1's total utility. Sen argues that egalitarianism would recommend that 2 be given more income than 1 to compensate 2, at least partly, for his physical handicap. However, utilitarianism will recommend exactly the opposite course of action: since the marginal utility of 2 will be half that of 1 when income is divided equally between the two individuals, the goal of maximizing the sum of the two individual's utilities will require that 1's share be more than half of the total income.

13.6 Concluding Remarks

In this chapter, I have outlined some of the main problems that utilitarianism faces as a philosophy of public action. I would like to conclude by commenting on two different responses to many of these problems.

One response is to extend the notion of utility to include the various factors that have been inappropriately excluded from the traditional utilitarian framework. For example, if people prefer to have rights and liberties at least partly because rights and liberties let them develop their own objectives and interests in life, or because the exercise of rights and liberties develops their human faculties, then one may ask why all these aspects, which I have included in my description of consequences, should not be introduced as arguments in the utility functions of the individuals. In fact, if it is claimed that rights and liberties have an intrinsic value, then why not incorporate rights and liberties directly as arguments of the utility functions of individuals? Consider an example.[23] Suppose that people do not like being ordered to work at a particular job, though they may choose to work at the same job when they are free to choose their job. Thus autonomy in one's job choice has an independent value for an individual apart from the value of the job itself at which he works. Let the variable e denote the job at which the individual i actually works. Assume that e can take any one of three possible values, a, b, and c. Let the variable θ denote the process through which individual i gets to work at her job. θ can be either the process where by i is ordered to work at job e, or θ may be the process where by i freely chooses to take up a job from the set $\{a, b, c\}$. If i cares for the process itself, then we can write i's utility function as $u_i(\theta, e)$. Similarly, it can be argued, any nonutilitarian concern can be accommodated by introducing into the individual utility functions suitably specified arguments to capture such concerns.

[23] See Hahn (1982, pp. 188–90).

One can then proceed to take, if one likes, the sum of these redefined individual utilities. There are several points that may be worth noting about this proposal. First, if we follow this suggestion, we would retain the terminology of utilitarians, but the content of the concept of utility would be vastly different from what they had in mind. Secondly, it is necessary to distinguish between different possible interpretations of the re-specified utility functions. For example, if freedom enters as an argument of the utility function in its own right, then does an individual's utility function reflect her valuation of freedom, or does it reflect simply her desire for freedom? If it reflects just her desire for freedom, then many of the concerns that we considered earlier will appear again. Finally, if the re-specified utility function reflects the individual's valuation, then the problem of interpersonal comparisons of such re-specified utility will be very different from the problem of interpersonal comparisons of happiness or desire satisfaction.

An alternative way of overcoming many of the restrictive features of utilitarianism may be to introduce a more "objective" notion of individual well-being and to use it as the basis for assessing the consequences of public policies. One important example of such an objective approach is the functioning-based approach to human well-being originating in the important contributions of Sen and Nussbaum (see e.g. Sen 1985, 1987*b*, and Nussbaum 1988, 2000). In Sen's (1987*b*, p. 29) terminology, functionings are the "doings" and "beings" that people value in their lives. Thus, being well nourished, being free from morbidity, and interacting with one's friends and family are all examples of functionings. In Sen's framework, not only does an individual's well-being depend on the "functioning bundle" achieved by her, but it also depends on her capability set or the set of all functioning bundles from which she makes her choice. The capability set is intended to reflect the individual's freedom or opportunity to choose the objectives of her life. The functioning approach of Sen and Nussbaum is still in an early phase of its development, but it promises to be an important alternative approach to the notion of human well-being that avoids many of the problems of the traditional utility-based framework of welfare economics.

References

Anand, P. (2003). The Integration of Claims to Health-Care: A Programming Approach. *Journal of Health Economics*, 22, 731–45.

Arrow, A. (1963). *Social Choice and Individual Values*, 2nd edn. New York: Wiley.

d'Aspremont, C., and Gevers, L. (2002). Social Welfare Functionals and Interpersonal Comparability. In K. J. Arrow, A. K. Sen, and K. Suzumura (eds.), *Handbook of Social Choice*, i. 459–541. Amsterdam: North-Holland.

Dworkin, R. (1978). *Taking Rights Seriously*, rev. edn. London: Duckworth.

—— (1984). "Rights as Trumps". In J. Waldron (ed.), *Theories of Rights*, 153–67. Oxford: Oxford University Press.

Elster, J. (1982). Sour Grapes—Utilitarianism and the Genesis of Wants. In Sen and Williams (1982), 219–38.

Gaertner, W., Pattanaik, P. K., and Suzumura, K. (1992). Individual Rights Revisited. *Economica*, 59, 161–77.

Galbraith, J. K. (1958). *The Affluent Society*. Boston: Houghton Mifflin Company.

Goodwin, R. E. (1995). *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Hahn, F. (1982). On Some Difficulties of the Utilitarian Economist. In Sen and Williams (1982), 187–98.

Hammond, P. (1980). Some Uncomfortable Options in Welfare Economics under Uncertainty. Mimeo, Stanford University.

—— (1981). Ex-Post Optimality as a Consistent Objective for Collective Choice under Uncertainty. Economics Technical Report, Institute for Mathematical Studies in the Social Sciences, Stanford University.

—— (1982). Utilitarianism, Uncertainty and Information. In Sen and Williams (1982), 85–102.

Hare, R. M. (1976). Ethical Theory and Utilitarianism. In H. D. Lewis (ed.), *Contemporary British Philosophy*, iv. 113–31. London: Allen and Unwin. Repr. in Sen and Williams (1982), 23–38.

Harsanyi, J. (1953). Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking. *Journal of Political Economy*, 61, 434–5.

—— (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, 63, 309–21.

—— (1977*a*). Morality and the Theory of Rational Behaviour. *Social Research*, 44, 623–56. Repr. in Sen and Williams (1982), 39–62.

—— (1977*b*). *Rational Behavior and Bargaining in Games and Social Situations.* Cambridge: Cambridge University Press.

—— (1977*c*). Rule Utilitarianism and Decision Theory. *Erkenntnis*, 11, 25–33.

Kaplow, L., and Shavell, S. (2001). Any Non-Welfarist Method of Policy Assessment Violates the Pareto Principle. *Journal of Political Economy*, 109, 281–6.

Keynes, J. M. (1931). Economic Possibilities for Our Grandchildren. In *idem*, *Essays in Persuasion*, 358–73. Repr. London: Macmillan. New York: W. W. Norton, 1963.

Luce, R. D., and Raiffa, H. (1957). *Games and Decisions*. New York: Wiley.

Mill, J. S. (1859). *On Liberty*. Repr. in *Utilitarianism and Other Writings*. New York, Penguin Books, 1962.

—— (1861). *Utilitarianism.* Repr. in *Utilitarianism and Other Writings*. New York: Penguin Books, 1962.

Musgrave, R. A. (1987). "Merit Goods". In John Eatwell, Murray Milgate, and Peter Newman (eds.), *The New Palgrave: A Dictionary of Economics*, iii. 452–3. Houndmills, Basingstoke: Palgrave Macmillan.

Nozick, R. (1974). *Anarchy, State and Utopia*. Oxford: Blackwell.

Nussbaum, M. (1988). Nature, Function and Capability: Aristotle on Political Distribution. *Oxford Studies in Ancient Philosophy*, suppl. vol. I, 145–84.

—— (2000). *Women and Human Development*. Cambridge: Cambridge University Press.

Overvold, M. C. (1982). Self-Interest and Getting What You Want. In H. B. Miller and W. H. Williams (eds.), *The Limits of Utilitarianism*, 185–94. Minneapolis: University of Minnesota Press.

Pattanaik, P. K. (1996). On Modelling Individual Rights: Some Conceptual Issues. In K. J. Arrow, A. Sen, and K. Suzumura (eds.), *Social Choice Re-examined*, ii. 100–28. London: Macmillan.

Rawls, J. (1971). *A Theory of Justice*. Oxford: Oxford University Press.

Scanlon, T. M. (1975). Preference and Urgency. *Journal of Philosophy*, 72, 665–9.

—— (1977). Rights, Goals, and Fairness. *Erkenntnis*, 2, 81–94. Repr. in J. Waldron (ed.), *Theories of Rights*, 137–52. Oxford: Oxford University Press, 1984.

Sen, A. (1970*a*). *Collective Choice and Social Welfare*. San Francisco: Holden Day.

—— (1970*b*). The Impossibility of a Paretian Liberal. *Journal of Political Economy*, 78, 152–7.

—— (1973). *On Economic Inequality.* Oxford: Oxford University Press.

—— (1977). Non-Linear Social Welfare Functions: A Reply to Professor Harsanyi. In R. Butts and J. Hintikka (eds.), *Foundational Problems in the Special Sciences*, 279–302. Dordrecht: Reidel.

—— (1979). Interpersonal Comparisons of Welfare. In M. Boskin (ed.), *Economics and Human Welfare*, 183–201. New York: Academic Press. Repr. in *idem*, *Choice Welfare and Measurement*, 264–82. Cambridge, MA: Harvard University Press.

—— (1980). Equality of What?. In S. McMurrin (ed.), *Tanner Lectures on Human Values*, i. 195–220. Cambridge: Cambridge University Press.

—— (1982). Rights and Agency. *Philosophy and Public Affairs*, 11, 3–38.

—— (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.

—— (1987*a*). *On Ethics and Economics*. Oxford: Blackwell.

—— (1987*b*). *The Standard of Living*. Cambridge: Cambridge University Press.

—— (1992). Minimal Liberty. *Economica*, 59, 139–59.

—— and Williams, B. (eds.) (1982). *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Singer, P. (1975). *Animal Liberation: A New Ethics for our Treatment of Animals*. New York: Random House.

—— (1979). *Practical Ethics*. Cambridge: Cambridge University Press.

Sugden, R. (1985). Liberty, Preference, and Choice. *Economics and Philosophy*, 1, 213–29.

Wailoo, A., and Anand, P. (2005). The Nature of Procedural Preferences for Health-Care Rationing Decisions. *Social Science and Medicine*, 60, 223–36.

Williams, B. (1973). A Critique of Utilitarianism. In J. J. C. Smart and B. Williams, *Utilitarianism For and Against*, 75–150. Cambridge: Cambridge University Press.

CHAPTER 14

CONSEQUENTIALISM AND NON-CONSEQUENTIALISM

THE AXIOMATIC APPROACH

KOTARO SUZUMURA
YONGSHENG XU

14.1 Introduction

It is undeniable that most, if not all, welfare economists and social choice theorists are *welfaristic* in their conviction, in the sense that they regard an economic policy and/or economic system to be satisfactory if and only if it is warranted to generate culmination outcomes which score high on the measuring rod of social welfare. It is equally undeniable that there exist people who care not only about welfaristic

We are grateful to Kenneth Arrow, Walter Bossert, Wulf Gaertner, Prasanta Pattanaik, and Amartya Sen, with whom we had several occasions to discuss the subject matter of this chapter. Thanks are also due to a reviewer for the Press, and to Y. Iwata and T. Sakai for several suggestions while we were preparing the chapter. Needless to say, they should not be held responsible for any defects that may remain. Financial support through a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan is gratefully acknowledged.

features of the consequences, but also about non-welfaristic features of the consequences, or even non-consequential features such as the *procedures* through which these consequences are brought about, or the *opportunity sets* from which these consequences are chosen.[1] Since welfare economics and social choice theory are concerned with the design and implementation of an economic policy or economic system from the viewpoint of persons constituting the society, even those welfare economists and social choice theorists with strong welfaristic convictions should be ready to take the judgments of people with non-welfaristic or non-consequentialist convictions into account in order not to be paternalistic in their social welfare analysis. The necessity to develop an analytical framework which enables us to examine the preferences of non-welfaristic and/or non-consequentialist people is all too clear. In a sequence of recent articles, Suzumura and Xu (2001, 2003, 2004) developed such an analytical framework and explored its implications in the context of Arrovian social choice theory.

The starting point of this analysis is to allow an individual to express his preferences of the following type: it is better for him that an outcome x is realized from the opportunity set A than that another outcome y is realized from the opportunity set B.[2] Note, in particular, that he is expressing his intrinsic valuation of the opportunity for choice if he prefers choosing an outcome x from an opportunity set A, where $\{x\}$ is a proper subset of A, rather than choosing the same outcome x from the singleton opportunity set $\{x\}$.

To a certain degree, a situation involving an individual's preference for an outcome x being realized from an opportunity set A against another outcome y being realized from another opportunity set B can be viewed as an example of the principle of *non-consequentialism*. Non-consequentialism emerges as a response to several unsatisfactory implications of *consequentialism*, which is a moral principle requiring that the right action in a given situation be the one that produces the best culmination outcome, "as judged from an impersonal standpoint which gives equal weight to the interests of everyone" (Scheffler 1988, p. 1). The gist of non-consequentialism is that, unlike consequentialism, it is an *agent-relative* morality, in maintaining both "agent-relative constraints, which sometimes prohibits the performance of acts that would have optimal results, and agent-relative permissions,

[1] The relevance of procedures or processes in decision-making has been argued by various writers such as Simon (1976, 1978), Sen (1995, 1996) and Suzumura (1999, 2000), and it has been empirically observed in several settings in economics, including experimental games (see e.g. Rabin (1993, 2002) and Bolton, Brandts and Ockenfels (2005), where it is found that players care not only about culmination outcomes, but also about the processes/procedures through which those culmination outcomes are brought about), and in empirical measurement of happiness (see e.g. Frey and Stutzer (2004), where they find that people care about culmination outcomes as well as procedures, especially participation rights in decision-making).

[2] Much attention has focused on the opportunity set evaluation, beginning with Sen (1985, 1988). See, among many others, Bossert, Pattanaik, and Xu (1994), Gravel (1994, 1998), Jones and Sugden (1982), Pattanaik and Xu (1990, 2000), Sen (1993, 1996), and Suzumura (1999, 2000).

which sometimes makes the performance of such acts optional" (Scheffler 1988, pp. 4–5). In our framework, when an individual prefers a situation involving an outcome x chosen from an opportunity set A containing several other outcomes to another situation involving an outcome y chosen from the opportunity set $\{y\}$ even though the outcome y is regarded better than the outcome x, the individual may be regarded as a non-consequentialist in that the individual attaches some significance to the "agent-relative" constraint, the opportunity aspect of choice, in making choices. This chapter will not go into detailed philosophical discussions about consequentialism and non-consequentialism. Readers interested in reading the relevant philosophical debate between consequentialists and non-consequentialists are advised to refer to Scheffler (1988). Instead, we will make use of our simple analytical framework to put forward concise definitions of various notions of consequentialism and non-consequentialism, and characterize these concepts in terms of a few simple axioms.

The structure of this chapter is as follows. In Section 14.2, we present the basic notations and definitions. Section 14.3 discusses the basic axioms which are assumed throughout this chapter. Some simple implications of these axioms are also identified in this section. In Section 14.4, we define and characterize axiomatically the concept of *extreme consequentialism* and *strong consequentialism*. We then turn in Section 14.5 to the concepts of *extreme non-consequentialism* and *strong non-consequentialism* and their axiomatic characterizations. Observe that these axiomatizations of consequentialism and non-consequentialism are concerned only with rather extreme cases where unequivocal priority is given to eventual consequences (resp. opportunities) not only in the case of extreme consequentialism (resp. extreme non-consequentialism) but also in the case of strong consequentialism (resp. strong non-consequentialism). Sections 14.6 and 14.7 introduce a more general framework, so that active interactions between consequential considerations and procedural considerations are allowed to play an essential role. In other words, it develops a framework which allows tradeoffs between the value of culmination outcomes and the richness of background opportunities. Section 14.6 (resp. 14.7) is devoted to the case where the universe of discourse is a finite (resp. an infinite) set. In Section 14.8, two simple applications of our analytical framework are briefly pursued. The first application is to the Arrovian social choice theory. Observe that Arrow's theoretical framework hinges squarely on the implicit assumption that all individuals in the society are consequentialist in nature. To gauge the extent to which Arrow's impossibility theorem and the resolution thereof hinge on this basic feature of his framework, two extended frameworks, in which individuals are supposed to express their preferences not only regarding culmination outcomes, but also regarding opportunity sets from which outcomes are chosen, are explored. The second application is to the ultimatum games. Capitalizing on the observations based on the experimental ultimatum games, we define and characterize extreme consequentialism, consequentialism, and fairness-conscious non-consequentialism

in this context. Section 14.9 concludes this chapter with several qualifications and observations. Proofs are contained in the Appendix.

14.2 Basic Notations and Definitions

Let X be the universal set which contains all mutually exclusive and jointly exhaustive *social states*. The elements of X will be denoted by $x, y, z, \ldots$, and they are interpreted as *culmination outcomes*. K denotes the set of all nonempty and finite subsets of X. The elements in K will be denoted by $A, B, C, \ldots$, and they are interpreted as *opportunity sets*. Let $\Omega = \{(x, A) : x \in A, A \in K\}$. Elements of Ω will be denoted by $(x, A), (y, B), (z, C), \ldots$, and they are called *extended alternatives* with the intended interpretation that the culmination outcome x is chosen from the opportunity set A. Throughout this chapter, we use the simplest possible measure of the richness of opportunity contained in each opportunity set $A \in K$, namely the cardinality $|A|$ of A. The use of this measure is not unproblematic, which we will point out as we go along.

Let $\succcurlyeq$ be a reflexive, complete and transitive binary relation over Ω. The asymmetric and symmetric parts of $\succcurlyeq$ will be denoted by $\succ$ and $\sim$, respectively. For any $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B)$ is interpreted as "choosing x from the opportunity set A is at least as good as choosing y from the opportunity set B". The binary relation $\succcurlyeq$ may be called an *extended preference ordering*. In this extended framework, when the decision-maker's extended preference is such that $(x, A) \succ (x, \{x\})$ for some $A \in \Omega$, it may be construed that he attaches intrinsic value as well as instrumental value to the opportunity set A. The following definitions are meant to capture the attitudes of the decision-making agent towards extended alternatives.

Definition 1. $\succcurlyeq$ *is said to be*

(1.1) extremely consequential if, for all $(x, A), (x, B) \in \Omega$, $(x, A) \sim (x, B)$;

(1.2) strongly consequential if, for all $x, y \in X$ *and all* $(x, A), (y, B) \in \Omega$, $(x, \{x\}) \sim (y, \{y\})$ *implies* $[(x, A) \succcurlyeq (y, B) \Leftrightarrow |A| \geq |B|]$, *and* $(x, \{x\}) \succ (y, \{y\})$ *implies* $(x, A) \succ (y, B)$;

(1.3) extremely non-consequential if, for all $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow |A| \geq |B|$;

(1.4) strongly non-consequential if, for all $x, y \in X$, *and all* $(x, A), (y, B) \in \Omega$, $|A| > |B| \Rightarrow (x, A) \succ (y, B)$, *and* $|A| = |B| \Rightarrow [(x, \{x\}) \succcurlyeq (y, \{y\}) \Leftrightarrow (x, A) \succcurlyeq (y, B)]$.

Thus, according to extreme consequentialism, two extended alternatives (x, A) and (y, B) in Ω are judged to be indifferent as long as $x = y$, no matter how

the opportunity sets A and B differ from each other. In other words, an extreme consequentialist cares only about culmination outcomes and pays no attention to the background opportunity sets. Strong consequentialism, on the other hand, stipulates that, in evaluating two extended alternatives (x, A) and (y, B) in Ω, the opportunity sets A and B do not matter when the decision-making agent has a strict extended preference for $(x, \{x\})$ against $(y, \{y\})$, and it is only when the decision-making agent is indifferent between $(x, \{x\})$ and $(y, \{y\})$ that the opportunity sets A and B matter in ranking (x, A) *vis-à-vis* (y, B) in terms of the richness of respective opportunities.

Extreme non-consequentialism may be regarded as the polar extreme case of consequentialism in that, in evaluating two extended alternatives (x, A) and (y, B) in Ω, the outcomes x and y are not valued at all, and the richness of opportunities reflected by the opportunity sets A and B exhausts everything that matters. In its complete neglect of culmination outcomes, extreme non-consequentialism is indeed extreme, but it captures the sense in which people may say: "Give me liberty, or give me death." It is in a similar vein that, in evaluating two extended alternatives (x, A) and (y, B) in Ω, strong non-consequentialism ignores the culmination outcomes x and y when the two opportunity sets A and B have different cardinality. It is only when the two opportunity sets A and B have identical cardinality that the culmination outcomes x and y have something to say in ranking (x, A) *vis-à-vis* (y, B).

14.3 Basic Axioms and their Implications

In this section, we introduce three basic axioms for the extended preference ordering $\succcurlyeq$, which are proposed in Suzumura and Xu (2001, 2003), and present their implications.

Independence (IND). For all $(x, A), (y, B) \in \Omega$, and all $z \in X \setminus A \cup B$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow (x, A \cup \{z\}) \succcurlyeq (y, B \cup \{z\})$.

Simple Indifference (SI). For all $x \in X$, and all $y, z \in X \setminus \{x\}$, $(x, \{x, y\}) \sim (x, \{x, z\})$.

Simple Monotonicity (SM). For all $(x, A), (x, B) \in \Omega$, if $B \subseteq A$, then $(x, A) \succcurlyeq (x, B)$.

The axiom (IND) can be regarded as the counterpart of an independence property used in the literature on ranking opportunity sets in terms of the freedom of choice; see, for example, Pattanaik and Xu (1990). It requires that, for all extended alternatives (x, A) and (y, B) in Ω, if an alternative z is not in both A and B, then

the extended preference ranking over $(x, A \cup \{z\})$ and $(y, B \cup \{z\})$ corresponds to that over (x, A) and (y, B), regardless of the nature of the added alternative $z \in X \setminus A \cup B$. This axiom may be criticized along several lines. For example, when freedom of choice is viewed as offering the decision-making agent a certain degree of diversity, (IND) may be problematic. It may be the case that the added alternative z is very *similar* to some existing alternatives in A, but is very *dissimilar* to all the alternatives in B. In such a case, the addition of z to A may not increase the degree of freedom already offered by A, while adding z to B may increase the degree of freedom offered by B substantially (see Bossert, Pattanaik, and Xu 2003, and Pattanaik and Xu 2000, 2006 for some formal analysis of diversity). As a consequence, the decision-making agent may rank $(y, B \cup \{z\})$ strictly above $(x, A \cup \{z\})$, even though he ranks (x, A) at least as high as (y, B). It may also be argued that the added alternative may have "epistemic value" in that it tells us something important about the nature of the choice situation which prompts a rejection of (IND). Consider the following example, which is due to Sen (1996, p. 753): "If invited to tea (t) by an acquaintance you might accept the invitation rather than going home (O), that is, pick t from the choice over $\{t, O\}$, and yet turn the invitation down if the acquaintance, whom you do not know very well, offers you a *wider* menu of having either tea with him or some heroin and cocain (h); that is, you may pick O, rejecting t, from the larger set $\{t, h, O\}$. The expansion of the menu offered by this acquaintance may tell you something about the kind of person he is, and this could affect your decision even to have tea with him." This constitutes a clear violation of (IND) when $A = B$.

The axiom (SI) requires that choosing x from "simple" cases, each involving two alternatives, is regarded as indifferent to each other. It should be noted that (SI) is subject to similar criticisms to (IND).

Finally, the axiom (SM) is a monotonicity property requiring that choosing an alternative x from the set A cannot be worse than choosing the same alternative x from the subset B of A. Various counterparts of (SM) in the literature on ranking opportunity sets in terms of freedom of choice have been proposed and studied (see e.g. Bossert, Pattanaik, and Xu 1994; Gravel 1994, 1998; Pattanaik and Xu 1990, 2000). It basically reflects the conviction that the decision-making agent is not averse to richer opportunities. In some cases, as argued in Dworkin (1982), richer opportunities can be a liability rather than an asset. In such cases, the decision-making agent may prefer choosing x from a smaller set to choosing the same x from a larger set.

The following results, Propositions 1, 2, and 3, summarize the implications of the above three axioms.

proposition 1 (Suzumura and Xu 2001, thm. 3.1). *If $\succcurlyeq$ satisfies (IND) and (SI), then for all $(x, A), (x, B) \in \Omega$, $|A| = |B| \Rightarrow (x, A) \sim (x, B)$.*

Proposition 2. *If $\succcurlyeq$ satisfies (IND) and (SI), then*

(2.1) For all $x \in X$, if there exists $y \in X \setminus \{x\}$ such that $(x, \{x, y\}) \succ (x, \{x\})$, then for all $(x, A), (x, B) \in \Omega$, $|A| \geq |B| \Leftrightarrow (x, A) \succcurlyeq (x, B)$;

(2.2) For all $x \in X$, if there exists $y \in X \setminus \{x\}$ such that $(x, \{x, y\}) \sim (x, \{x\})$, then for all $(x, A), (x, B) \in \Omega$, $(x, A) \sim (x, B)$;

(2.3) For all $x \in X$, if there exists $y \in X \setminus \{x\}$ such that $(x, \{x\}) \succ (x, \{x, y\})$, then for all $(x, A), (x, B) \in \Omega$, $|A| \leq |B| \Leftrightarrow (x, A) \succcurlyeq (x, B)$.

proposition 3 (Suzumura and Xu 2003, lemma 3.1). *Let $\succcurlyeq$ be an ordering over Ω satisfying (IND), (SI), and (SM). Then, for all $(a, A), (b, B) \in \Omega$, and all $x \in X \setminus A$, $y \in X \setminus B$, $(a, A) \succcurlyeq (b, B) \Leftrightarrow (a, A \cup \{x\}) \succcurlyeq (b, B \cup \{y\})$.*

14.4 Consequentialism

In this section, we present axiomatic characterizations of extreme consequentialism and strong consequentialism. To characterize these two versions of consequentialism, we consider the following three axioms, which are proposed in Suzumura and Xu (2001).

Local Indifference (LI): For all $x \in X$, there exists $(x, A) \in \Omega \setminus \{(x, \{x\})\}$ such that $(x, \{x\}) \sim (x, A)$.

Local Strict Monotonicity (LSM): For all $x \in X$, there exists $(x, A) \in \Omega \setminus \{(x, \{x\})\}$ such that $(x, A) \succ (x, \{x\})$.

Robustness (ROB): For all $x, y, z \in X$, all $(x, A), (y, B) \in \Omega$, if $(x, \{x\}) \succ (y, \{y\})$ and $(x, A) \succ (y, B)$, then $(x, A) \succ (y, B \cup \{z\})$.

The axiom (LI) is a mild requirement of extreme consequentialism: for each $x \in X$, there exists an opportunity set A in K, which is distinct from $\{x\}$, such that choosing the alternative x from A is regarded as indifferent to choosing x from the singleton set $\{x\}$. It may be regarded as a local property of extreme consequentialism. The axiom (LSM), on the other hand, requires that, for each $x \in X$, there exists an opportunity set A, which is distinct from $\{x\}$, such that choosing x from the opportunity set A is valued strictly higher than choosing x from the singleton opportunity set $\{x\}$. It reflects the decision-maker's desire to value opportunities at least in this very limited sense. The axiom (ROB) requires that, for all $x, y, z \in X$, all $(x, A), (y, B) \in \Omega$, if the decision-maker values $(x, \{x\})$ higher than $(y, \{y\})$, and (x, A) higher than (y, B), then the addition of z to B while maintaining y being chosen from $B \cup \{z\}$ will not affect the decision-making agent's value-ranking: (x, A) is still valued higher than $(y, B \cup \{z\})$.

The characterizations of extreme consequentialism and strong consequentialism are given in the following two theorems.

Theorem 1 (Suzumura and Xu 2001, thm. 4.1). $\succcurlyeq$ satisfies (IND), (SI), and (LI) if and only if it is extremely consequential.

Theorem 2 (Suzumura and Xu 2001, thm. 4.2). $\succcurlyeq$ satisfies (IND), (SI), (LSM), and (ROB) if and only if it is strongly consequential.

To conclude this section, we note that it is easily checked that the characterization theorems we obtained, namely Theorem 1 for extreme consequentialism and Theorem 2 for strong consequentialism, do not contain any redundancy.

14.5 Non-Consequentialism

To give characterizations of extreme non-consequentialism and strong non-consequentialism, the following axioms will be used.

Indifference of No-Choice Situations (INS): For all $x, y \in X$, $(x, \{x\}) \sim (y, \{y\})$.

Simple Preference for Opportunities (SPO): For all distinct $x, y \in X$, $(x, \{x, y\}) \succ \{y, \{y\})$.

The axiom (INS) requires that, in facing two choice situations in which each choice situation is restricted to a choice from a singleton set, the decision-making agent is indifferent between them. It thus conveys the idea that, in these simple cases, the decision-making agent feels that there is no real freedom of choice in each choice situation, so that he is ready to express his indifference between these simple choice situations *regardless of the nature of the culmination outcomes.* In a sense, it is the lack of freedom of choice that "forces" the decision-making agent to be indifferent between these situations. The underlying idea of (INS) is therefore similar to an axiom proposed by Pattanaik and Xu (1990) for ranking opportunity sets in terms of the freedom of choice, which requires that all singleton sets offer the decision-making agent the same amount of freedom of choice. The axiom (SPO) stipulates that it is always better for the agent to choose an outcome from the set containing two elements (one of which being the chosen culmination outcome) than to choose a culmination outcome from the singleton set. (SPO) therefore displays the decision-making agent's desire to have some genuine opportunities for choice. In this sense, (SPO) is in the same spirit as (LSM). However, as the following result shows, (SPO) is a stronger requirement than (LSM) in the presence of (IND) and (SI).

Proposition 4. *Suppose $\succcurlyeq$ satisfies (IND) and (SI). Then (SPO) implies (LSM).*

The following two results give the characterizations of extreme non-consequentialism and strong non-consequentialism.

Theorem 3. $\succcurlyeq$ satisfies (IND), (SI), (LSM) and (INS) if and only if it is extremely non-consequential.

Theorem 4. $\succcurlyeq$ satisfies (IND), (SI), and (SPO) if and only if it is strongly non-consequential.

We may note that the independence of the axioms used in Theorems 3 and 4 can be checked easily.

14.6 Active Interactions between Outcomes and Opportunities: The Case of Finite X

So far, we have focused exclusively on simple special cases where no tradeoff exists between consequential considerations, which reflect the decision-making agent's concern about culmination outcomes, and non-consequential considerations, which reflect his concern about richness of opportunities from which culmination outcomes are chosen. For these simple special cases, we have characterized the concepts of consequentialism and non-consequentialism. In this section, we generalize our previous framework by accommodating situations where consequential considerations and non-consequential considerations are allowed to interact actively.

Let $\mathcal{Z}$ and $\mathcal{R}$ denote the set of all positive integers and the set of all real numbers, respectively. We first state the following result.

Theorem 5 (Suzumura and Xu 2003, thm. 3.3). Suppose X is finite. $\succcurlyeq$ satisfies (IND), (SI), and (SM) if and only if there exist a function $u : X \to \mathcal{R}$ and a function $f : \mathcal{R} \times \mathcal{Z} \to \mathcal{R}$ such that

(T5.1) For all $x, y \in X$, $u(x) \geq u(y) \Leftrightarrow (x, \{x\}) \succcurlyeq (y, \{y\})$;
(T5.2) For all $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow f(u(x), |A|) \geq f(u(y), |B|)$;
(T5.3) f is non-decreasing in each of its arguments and has the following property:
For all integers $i, j, k \geq 1$ and all $x, y \in X$, if $i + k, j + k \leq |X|$, then
(T5.3.1) $f(u(x), i) \geq f(u(y), j) \Leftrightarrow f(u(x), i + k) \geq f(u(y), j + k)$.

The function u in Theorem 5 can be regarded as the usual utility function defined on the set of (conventional) social states, whereas the cardinality of opportunity sets

may be regarded as an index of the richness of opportunities offered by opportunity sets. The function f thus weighs the utility of consequential outcomes against the value of richness of opportunities. The active interactions between the utility of consequential outcomes and the value of richness of opportunities are therefore captured by Theorem 5. It is clear that the concepts of consequentialsm and non-consequentialism can be obtained as special cases of Theorem 5 by defining the appropriate f functions.

14.7 Active Interactions between Outcomes and Opportunities: The Case of Infinite X

A limitation of Theorem 5 is that it assumes X to be finite. In many contexts in economics, the universal set of social states is typically infinite. The following two results deal with this case: Theorem 6 presents a full characterization of all the orderings satisfying (IND), (SI), and (SM), while Theorem 7 gives a representation of any ordering characterized in Theorem 6.

Theorem 6 (Suzumura and Xu 2003, thm. 4.1). $\succcurlyeq$ satisfies (IND), (SI), and (SM) if and only if there exists an ordering $\succcurlyeq^{\#}$ on $X \times \mathcal{Z}$ such that

(T6.1) For all $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow (x, |A|) \succcurlyeq^{\#} (y, |B|)$;
(T6.2) For all integers $i, j, k \geq 1$ and all $x, y \in X$, $(x, i) \succcurlyeq^{\#} (y, j) \Leftrightarrow (x, i + k) \succcurlyeq^{\#} (y, j + k)$, and $(x, i + k) \succcurlyeq^{\#} (x, i)$.

To present our next theorem, we need the following continuity property, which was introduced in Suzumura and Xu (2003). Suppose that $X = \mathcal{R}_{+}^{n}$ for some natural number n.

Continuity (CON): For all $(x, A) \in \Omega$, all $y, y^i \in X$ $(i = 1, 2, \dots)$, and all $B \in K \cup \{\emptyset\}$, if $B \cap \{y^i\} = B \cap \{y\} = \emptyset$ for all $i = 1, 2, \dots$, and $\lim_{i \to \infty} y^i = y$, then $[(y^i, B \cup \{y^i\}) \succcurlyeq (x, A)$ for $i = 1, 2, \dots] \Rightarrow (y, B \cup \{y\}) \succcurlyeq (x, A)$, and $[(x, A) \succcurlyeq (y^i, B \cup \{y^i\})$ for $i = 1, 2, \dots] \Rightarrow (x, A) \succcurlyeq (y, B \cup \{y\})$.

Theorem 7 (Suzumura and Xu 2003, thm. 4.5). Suppose that $X = \mathcal{R}_{+}^{n}$ and that $\succcurlyeq$ satisfies (IND), (SI), (SM), and (CON). Then, there exists a function $v : X \times \mathcal{Z} \to \mathcal{R}$, which is continuous in its first argument, such that

(T7.1) For all $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow v(x, |A|) \geq v(y, |B|)$,
(T7.2) For all $i, j, k \in \mathcal{Z}$ and all $x, y \in X$, $v(x, i) \geq v(y, j) \Leftrightarrow v(x, i + k) \geq v(y, j + k)$ and $v(x, i + k) \geq v(x, i)$.

14.8 Applications

14.8.1 Arrovian Social Choice

In this subsection, we discuss how our notions of consequentialism and non-consequentialism can affect the fate of Arrow's impossibility theorem in social choice theory. For this purpose, let X consist of at least three, but finite, social alternatives. Each alternative in X is assumed to be a public alternative, such as a list of public goods to be provided in the society, or a description of a candidate in a public election. The set of all individuals in the society is denoted by $N = \{1, 2, \ldots, n\}$, where $+\infty > n \geq 2$. Each individual $i \in N$ is assumed to have an extended preference ordering R_i over Ω, which is *reflexive, complete*, and *transitive*. For any $(x, A), (y, B) \in \Omega$, $(x, A)R_i(y, B)$ is interpreted as follows: i feels at least as good when choosing x from A as when choosing y from B. The asymmetric part and the symmetric part of R_i are denoted by $P(R_i)$ and $I(R_i)$, respectively, which denote the strict preference relation and the indifference relation of $i \in N$.

The set of all logically possible orderings over Ω is denoted by $\mathcal{R}$. Then, a *profile* $\mathbf{R} = (R_1, R_2, \ldots, R_n)$ of extended individual preference orderings, one extended ordering for each individual, is an element of $\mathcal{R}^n$. An *extended social welfare function* (ESWF) is a function f which maps each and every profile in some subset D_f of $\mathcal{R}^n$ into $\mathcal{R}$. When $R = f(\mathbf{R})$ holds for some $\mathbf{R} \in D_f$, $I(R)$ and $P(R)$ stand, respectively, for the social indifference relation and the social strict preference relation corresponding to R.

We assume that each and every profile $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f$ is such that R_i satisfies the properties (IND), (SI), and (SM) for all $i \in N$.

In addition to the domain restriction on D_f introduced above, we first introduce two conditions corresponding to Arrow's (1963) Pareto principle and non-dictatorship to be imposed on f. They are well known, and require no further explanation.

Strong Pareto Principle (SP): For all $(x, A), (y, B) \in \Omega$, and for all $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f$, if $(x, A)P(R_i)(y, B)$ holds for all $i \in N$, then we have $(x, A)P(R)(y, B)$, and if $(x, A)I(R_i)(y, B)$ holds for all $i \in N$, then we have $(x, A)I(R)(y, B)$, where $R = f(\mathbf{R})$.

Non-Dictatorship (ND): There exists no $i \in N$ such that $[(x, A)P(R_i)(y, B) \Rightarrow (x, A)P(R)(y, B)$ for all $(x, A), (y, B) \in \Omega]$ holds for all $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f$, where $R = f(\mathbf{R})$.

There are various ways of formulating Arrow's IIA in our context. Consider the following:

Independence of Irrelevant Alternatives (i) (IIA(i)): For all $\mathbf{R^1} = (R_1^1, R_2^1, \ldots, R_n^1), \mathbf{R^2} = (R_1^2, R_2^2, \ldots, R_n^2) \in D_f$, and $(x, A), (y, B) \in \Omega$, if $[(x, A)R_i^1(y, B)$

$\Leftrightarrow (x, A)R_i^2(y, B)$ and $(x, \{x\})R_i^1(y, \{y\}) \Leftrightarrow (x, \{x\})R_i^2(y, \{y\})]$ for all $i \in N$, then $[(x, A)R^1(y, B) \Leftrightarrow (x, A)R^2(y, B)]$ where $R^1 = f(\mathbf{R^1})$ and $R^2 = f(\mathbf{R^2})$.

Independence of Irrelevant Alternatives (ii) (IIA(ii)): For all $\mathbf{R^1} = (R_1^1, R_2^1, \ldots, R_n^1)$, $\mathbf{R^2} = (R_1^2, R_2^2, \ldots, R_n^2) \in D_f$, and $(x, A), (y, B) \in \Omega$ with $|A| = |B|$, if $[(x, A)R_i^1(y, B) \Leftrightarrow (x, A)R_i^2(y, B)]$ for all $i \in N$, then $[(x, A)R^1(y, B) \Leftrightarrow (x, A)R^2(y, B)]$, where $R^1 = f(\mathbf{R^1})$ and $R^2 = f(\mathbf{R^2})$.

Full Independence of Irrelevant Alternatives (FIIA): For all $\mathbf{R^1} = (R_1^1, R_2^1, \ldots, R_n^1)$, $\mathbf{R^2} = (R_1^2, R_2^2, \ldots, R_n^2) \in D_f$, and $(x, A), (y, B) \in \Omega$, if $[(x, A)R_i^1(y, B) \Leftrightarrow (x, A)R_i^2(y, B)]$ for all $i \in N$, then $[(x, A)R^1(y, B) \Leftrightarrow (x, A)R^2(y, B)]$, where $R^1 = f(\mathbf{R^1})$ and $R^2 = f(\mathbf{R^2})$.

(IIA(i)) says that the extended social preference between any two extended alternatives (x, A) and (y, B) depends on each individual's extended preference between them, as well as each individual's extended preference between $(x, \{x\})$ and $(y, \{y\})$: for all profiles $\mathbf{R^1}$ and $\mathbf{R^2}$, if $[(x, A)R_i^1(y, B)$ if and only if $(x, A)R_i^2(y, B)$, and $(x, \{x\})R_i^1(y, \{y\})$ if and only if $(x, \{x\})R_i^2(y, \{y\})]$ for all $i \in N$, then $(x, A)R^1(y, B)$ if and only if $(x, A)R^2(y, B)$, where $R^1 = f(\mathbf{R^1})$ and $R^2 = f(\mathbf{R^2})$. (IIA(ii)), on the other hand, says that the extended social preference between any two extended alternatives (x, A) and (y, B) with $|A| = |B|$ depends on each individual's extended preference between them. Finally, (FIIA) says that the extended social preference between *any* two extended alternatives (x, A) and (y, B) depends on each individual's extended preference between them. It is clear that (IIA(i)) is logically independent of (IIA(ii)), and both (IIA(i)) and (IIA(ii)) are logically weaker than (FIIA).

Let us observe that each and every individual in the original Arrow framework can be regarded as an extreme consequentialist. Thus, Arrow's impossibility theorem can be viewed as an impossibility result in the framework of extreme consequentialism. What will happen to the impossibility theorem in a framework which is broader than extreme consequentialism? For the purpose of answering this question, let us now introduce three domain restrictions on f by specifying some appropriate subsets of D_f. In the first place, let $D_f(E)$ be the set of all profiles in D_f such that all individuals are extreme consequentialists. Secondly, let $D_f(E \cup S)$ be the set of all profiles in D_f such that at least one individual is an extreme consequentialist *uniformly* for all profiles in $D_f(E \cup S)$ and at least one individual is a strong consequentialist *uniformly* for all profiles in $D_f(E \cup S)$. Finally, let $D_f(N)$ be the set of all profiles in D_f such that at least one individual is a strong non-consequentialist *uniformly* for all profiles in $D_f(N)$.

Our first result in this subsection is nothing but a restatement of Arrow's original impossibility theorem in the framework of extreme consequentialism.

Theorem 8. Suppose that all individuals are extreme consequentialists. Then, there exists no extended social welfare function f with the domain $D_f(E)$ which satisfies (SP), (ND), and either (IIA(i)) or (IIA(ii)).

However, once we go beyond the framework of extreme consequentialism, as shown by the following results, a new scope for resolving the impossibility result is opened.

Theorem 9. Suppose that there exist at least one uniform extreme consequentialist over $D_f(E \cup S)$ and at least one uniform strong consequentialist over $D_f(E \cup S)$ in the society. Then, there exists an extended social welfare function f with the domain $D_f(E \cup S)$ satisfying (SP), (IIA(i)), (IIA(ii)), and (ND).

Theorem 10 (Suzumura and Xu 2004, thm. 4). Suppose that there exists at least one person who is a uniform strong non-consequentialist over $D_f(N)$. Then, there exists an extended social welfare function f with the domain $D_f(N)$ that satisfies (SP), (FIIA), and (ND).

To conclude this subsection, the following observations may be in order. To begin with, as shown by Iwata (2006), the possibility result obtained in Theorem 9 no longer holds if (IIA(i)) or (IIA(ii)) is replaced by (FIIA) while retaining (SP) and (ND) intact. On the other hand, as reported in Iwata (2006), there *exists* an ESWF over the domain $D_f(E \cup S)$ that satisfies (FIIA), (ND), and (WP): for all $(x, A), (y, B) \in \Omega$, and all $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f(E \cup S)$, if $(x, A)P(R_i)(y, B)$ for all $i \in N$, then $(x, A)P(R)(y, B)$, where $R = F(\mathbf{R})$. The proof of this result is quite involved, and interested readers are referred to Iwata (2006). Secondly, given that (FIIA) is stronger than (IIA(i)) or (IIA(ii)), the impossibility result of Theorem 8 still holds if (IIA(i)) or (IIA(ii)) is replaced by (FIIA) while retaining (SP) and (ND) intact. Thirdly, since the ESWF constructed in the proof of Theorem 10 satisfies (FIIA), it is clear that there exists an ESWF on $D_f(N)$ that satisfies (SP), (ND), and both (IIA(i)) and (IIA(ii)).

14.8.2 Ultimatum Games

In experimental studies of two-player extensive form games with complete information, it is observed that the second mover is not only concerned about his own monetary payoff, but cares also about the feasible set that is generated by the first mover's choice, from which he must make his choice (see e.g. Cox, Friedman, and Gjerstad 2007, and Cox, Friedman, and Sadiraj 2008). For the sake of easy presentation, we shall focus on ultimatum games where two players, the Proposer and the Responder, are to divide a certain amount of money between them, and see what is the framework which naturally suggests itself in this context.

Formally, an ultimatum game consists of two players, the Proposer and the Responder. The sequence of the game is as follows. The Proposer moves first, and he is presented a set X of feasible division rules by the experimenter. A division rule is chosen by the Proposer from the set X, which consists of the division rules in the pattern of (50, 50), (80, 20), (60, 40), (70, 30), and the like. The Proposer chooses a division rule $(x, 1-x) \in X$, where $0 \leq x \leq 1$. The intended interpretation is that the Proposer gets x percent and the Responder gets $(1-x)$ percent of the money to be divided. Upon seeing a division rule chosen by the Proposer from the given set X, the Responder then chooses an amount $m \geq 0$ of money to be divided between them. As a consequence, the Proposer's monetary payoff is xm, and the Responder's monetary payoff is $(1-x)m$. Consider the same payoff 8 for the Proposer and 2 for the Responder derived from two different situations, one involving the Proposer's choice of the (80, 20) division rule from the set {(80, 20)} and the other involving the Proposer's choice of the (80, 20) division rule from the set {(80, 20), (70, 30), (60, 40), (50, 50), (40, 60), (30, 70), (20, 80)}, the Responder's choice of money to be divided remaining the same at 10. Though the two situations yield the same payoff vector, the Responder's behavior has been observed to be very different. Though there are several possible explanations for such different behaviors on the Responder's side, we can explain the difference in the Responder's behavior via our notions of consequentialism and non-consequentialism.

Let $(x, 1-x)$ be the division rule chosen by the Proposer from the given set A of feasible division rules. The associated payoff vector with the division rule $(x, 1-x) \in A$ is denoted by $m(x) = (m_P(x), m_R(x))$, where $m_P(x)$ is the Proposer's payoff and $m_R(x)$ is the Responder's payoff. In our extended framework, we may describe the situation by the triple $(m(x), x, A)$, with the interpretation that the payoff vector is $m(x)$ for the chosen division rule $(x, 1-x)$ from the feasible set A. Let X be the finite set of all possible division rules, and Ω be the set of all possible triples $(m(x), x, A)$, where $A \subseteq X$ and $(x, 1-x) \in A$. Let $\succcurlyeq$ be the Responder's preference relation (reflexive and transitive, but not necessarily complete) over Ω, with its symmetric and asymmetric parts denoted, respectively, by $\sim$ and $\succ$. Then, we may define several notions of consequentialism and non-consequentialism. For example, we may say that the Responder is

(i) an *extreme consequentialist* if, for all $(m(x), x, A), (m(y), y, B) \in \Omega$, $m(x) = m(y) \Rightarrow (m(x), x, A) \sim (m(y), y, B)$;
(ii) a *consequentialist* if, for all $(m(x), x, A), (m(y), y, B) \in \Omega$, $[m(x) = m(y), x = y] \Rightarrow (m(x), x, A) \sim (m(y), y, B)$;
(iii) a *non-consequentialist* if, for some $(m(x), x, A), (m(y), y, B) \in \Omega$, we have $m(x) = m(y)$ but $(m(x), x, A) \succ (m(y), y, B)$.

Let us begin by providing a simple axiomatic characterization of the two notions of consequentialism. For this purpose, consider the following axioms.

Local Indifference* **(LI*):** For all $(m(x), x, X), (m(x), x, \{(x, 1-x)\}) \in \Omega$, $(m(x), x, X) \sim (m(x), x, \{(x, 1-x)\})$.

Monotonicity* **(M*):** For all $(m(x), x, A), (m(x), x, B) \in \Omega$, if $A \subseteq B$, then $(m(x), x, B) \succcurlyeq (m(x), x, A)$.

Conditional Indifference between No-choice Situations* **(CINS*):** For all $(m(x), x, \{(x, 1-x)\}), (m(y), y, \{(y, 1-y)\}) \in \Omega$, if $m(x) = m(y)$ then $(m(x), x, \{(x, 1-x)\}) \sim (m(y), y, \{(y, 1-y)\})$.

We may now assert the following:

Theorem 11. $\succcurlyeq$ over Ω satisfies (LI*) and (M*) if and only if it is consequential.

Theorem 12. $\succcurlyeq$ over Ω satisfies (LI*), (M*), and (CINS*) if and only if it is extremely consequential.

Turn, now, to the concept of non-consequentialism. Recollect that the experimental studies revealed that there is a situation, where $(x, 1-x) \in X$ and $\{(x, 1-x)\}$ is a proper subset of A which in turn is a subset of X, such that the Responder's preferences exhibit the following: $(m(x), x, \{(x, 1-x)\}) \succ (m(x), x, A)$. This is precisely a situation where the Responder is disgusted by the fact that the Provider has chosen an outrageously unequal method of division $(x, 1-x)$, not only from the no-choice situation $\{(x, 1-x)\}$, but also from the opportunity set which contains a conspicuously egalitarian method of division. Since our definition of non-consequentialism is so widely embracing, this revealed preference of the Responder can thereby be accommodated. Consider now a subclass of non-consequentialism which reads as follows: the Responder is a *fairness-conscious non-consequentialist* if, for all $(m(x), x, A), (m(y), y, B) \in \Omega$, if $m(x) = m(y)$, $x = y$ and $[z \geq x$ for all $(z, 1-z) \in A \cup B]$, then $|A| \geq |B| \Leftrightarrow (m(x), x, A) \succcurlyeq (m(y), y, B)$. This subclass of non-consequentialism may be characterized by introducing the following axioms:

Conditional Simple Preference for Opportunities* **(CSPO*):** For all $(m(x), x, \{(x, 1-x), (y, 1-y)\})$ and $(y, 1-y) \in X$, if $y > x$, then $(m(x), x, \{(x, 1-x), (y, 1-y)\}) \succ (m(x), x, \{(x, 1-x)\})$.

Conditional Independence* **(CIND*):** For all $(m(x), x, A), (m(x), x, B) \in \Omega$ and $(z, 1-z) \in X \setminus (A \cup B)$, if $m(x) = m(y)$, then $z \geq x$, $(m(x), x, A) \succcurlyeq (m(x), x, B) \Leftrightarrow (m(x), x, A \cup \{(z, 1-z)\}) \succcurlyeq (m(x), x, B \cup \{(z, 1-z)\})$.

We are now ready to assert the following:

Theorem 13. $\succcurlyeq$ over Ω satisfies (CINS*), (CSPO*), and (CIND*) if and only if it is a fairness-conscious non-consequentialist.

14.9 Concluding Remarks

In view of the undeniable dominance of consequentialism in the whole spectrum of modern welfare economics and social choice theory, it goes without saying that the clarification of what we mean by consequentialism and non-consequentialism, what role, if any, consequentialism *vis-à-vis* non-consequentialism plays in some of the fundamental propositions in normative economics, and what basic axioms, which are mutually exclusive and jointly sufficient to characterize consequentialism and non-consequentialism, are of fundamental importance. Capitalizing on some recent work, including our own, we have tried in this chapter to present a coherent account of what we know about these basic questions. In concluding, two qualifying and clarifying remarks are in order.

In the first place, we have assumed throughout the chapter that the universal set of alternatives, or at least each and every opportunity set which may be presented to the decision-making agent, is a finite set. It is this assumption that allows us to use a simple measure of the richness of opportunities, namely the number of alternatives in the opportunity set under scrutiny. Needless to say, this is a simplifying assumption which may well be crucially restrictive. This is well known in the related but distinct literature on the measurement of freedom of choice. Suffice it to note that choosing an outcome x from the singleton set $\{x\}$ may be judged to be inferior to choosing the same outcome x from the larger opportunity set A, where $\{x\}$ is a proper subset of A, if the decision-maker is a non-consequentialist who cares not just about culmination outcomes but also about opportunity sets which stand behind the choice of culmination outcomes. However, his preference for (x, A) over $(x, \{x\})$ may well be challenged if $A = \{x, y\}$, where $x =$ "a blue car" and $y =$ "a car exactly the same as x, except for its color, which is only slightly darker than that of x". In the literature on freedom of choice, there are several attempts to cope effectively with this problem. We have chosen to stick to the simplest possible treatment in order not to blur the crucial features of our novel problem by being fussy about less than central features such as the measurement of opportunity.

In the second place, there is a well-known alternative to our definition of consequentialism and non-consequentialism. Unlike our definition in terms of extended preference ordering over the pairs of culmination outcomes and background opportunity sets, this alternative definition makes use of extended preference ordering defined over the pairs of culmination outcomes and social decision-making procedures through which these outcomes are brought about. Due recognition of the importance of procedural considerations *vis-à-vis* consequential considerations abound in the literature. Suffice it to refer to Schumpeter (1942), Arrow (1951), and Lindbeck (1988) as a small sample list of economists who, in their respective ways, recognized the need for including social decision-making procedures or

mechanisms within the extended evaluative framework of normative economics. This extended framework provides us with an alternative method for articulating consequentialism and non-consequentialism. See, for example, Hansson (1992, 1996), who explored the possibility of resolving Arrow's impossibility result in the extended framework, Gaertner and Xu (2004), who investigated the effects of procedures on decision-makers' choices, Pattanaik and Suzumura (1994, 1996), and Suzumura and Yoshihara (2007), who explored the problem of initial conferment of libertarian rights.

The gate is wide open for further exploration of normative economics which goes beyond the traditional confinement of welfarist consequentialism.

Appendix

Proof of Proposition 1. Let $\succcurlyeq$ satisfy (IND) and (SI). Let $(x, A), (x, B) \in \Omega$ be such that $|A| = |B|$.

If $|A| = |B| = 1$, then $A = B = \{x\}$. By reflexivity of $\succcurlyeq$, $(x, A) \sim (x, B)$ follows immediately. If $|A| = |B| = 2$, then $(x, A) \sim (x, B)$ follows from (SI) directly. Consider now that $|A| = |B| = m + 1$, where $\infty > m \geq 2$.

Suppose first that $A \cap B = \{x\}$. Let $A = \{x, a_1, \ldots, a_m\}$ and $B = \{x, b_1, \ldots, b_m\}$. From (SI), we must have $(x, \{x, a_i\}) \sim (x, \{x, b_j\})$ for all $i, j = 1, \ldots, m$. By (IND), from $(x, \{x, a_2\}) \sim (x, \{x, b_1\})$, we obtain $(x, \{x, a_1, a_2\}) \sim (x, \{x, a_1, b_1\})$, and from $(x, \{x, a_1\}) \sim (x, \{x, b_2\})$, we obtain $(x, \{x, a_1, b_1\}) \sim (x, \{x, b_1, b_2\})$. By the transitivity of $\succcurlyeq$, it follows that $(x, \{x, a_1, a_2\}) \sim (x, \{x, b_1, b_2\})$. By using similar arguments, from (IND) and by the transitivity of $\succcurlyeq$, we can obtain $(x, A) \sim (x, B)$.

Next, suppose that $A \cap B = \{x\} \cup C$ where $C \neq \emptyset$. When $A \setminus C = B \setminus C = \emptyset$, it must be the case that $A = B$. By reflexivity of $\succcurlyeq$, $(x, A) \sim (x, B)$ follows easily. Suppose therefore that $A \setminus C \neq \emptyset$. Note that $B \setminus C \neq \emptyset$ and $|A \setminus C| = |B \setminus C|$. From above, we must then have $(x, (A \setminus C) \cup \{x\}) \sim (x, (B \setminus C) \cup \{x\})$. By the repeated use of (IND), $(x, A) \sim (x, B)$ can be obtained. □

Proof of Proposition 2. Let $\succcurlyeq$ satisfy (IND) and (SI). We will give a proof for the case (2.1). The proofs for the cases (2.2) and (2.3) are similar, and we omit them. Let $(x, A), (x, B) \in \Omega$. If $|A| = |B|$, then, by Proposition 1, $(x, A) \sim (x, B)$. Suppose now that $|A| \neq |B|$. Without loss of generality, let $|A| > |B|$. Consider $G \subset A$ such that $|G| = |B|$ and $x \in G$. Then, by Proposition 1, $(x, G) \sim (x, B)$. Let $A = G \cup H$ where $H = \{h_1, \ldots, h_t\}$. Let $G = \{x, g_1, \ldots, g_r\}$. Note that if there exists $y \in X \setminus \{x\}$ such that $(x, \{x, y\}) \succ (x, \{x\})$, then, from Proposition 1 and by the transitivity of $\succcurlyeq$, it must be true that $(x, \{x, z\}) \succ (x, \{x\})$ for all $z \in X \setminus \{x\}$. In particular, $(x, \{x, h_1\}) \succ (x, \{x\})$. Therefore, by the repeated use of (IND), we have $(x, G \cup \{h_1\}) \succ (x, G)$. Similarly, $(x, G \cup \{h_1, h_2\}) \succ (x, G \cup \{h_1\})$, and $(x, G \cup \{h_1, h_2, h_3\}) \succ (x, G \cup \{h_1, h_2\})$, and …, and $(x, G \cup H) \succ (x, G \cup H \setminus \{h_t\})$. By the transitivity of $\succcurlyeq$, it follows that $(x, A) \succ (x, G)$. Then, noting that $|G| = |B|$, from Proposition 1 and the transitivity of $\succcurlyeq$, we have $(x, A) \succ (x, B)$. □

Proof of Proposition 3. Let $\succcurlyeq$ satisfy (IND), (SI), and (SM). Let $(a, A), (b, B) \in \Omega$, $x \in X \setminus A$, $y \in X \setminus B$, and $(a, A) \succcurlyeq (b, B)$. Because $\succcurlyeq$ is an ordering, we have only to show that $(a, A) \sim (b, B) \Rightarrow (a, A \cup \{x\}) \sim (b, B \cup \{y\})$ and $(a, A) \succ (b, B) \Rightarrow (a, A \cup \{x\}) \succ (b, B \cup \{y\})$.

First, we show that

$$(a, A) \sim (b, B) \Rightarrow (a, A \cup \{x\}) \sim (b, B \cup \{y\}). \qquad (*)$$

Since $x \in X \setminus A$ and $y \in X \setminus B$, it is clear that $A \neq X$ and $B \neq X$. We consider three sub-cases: (i) $A = \{a\}$; (ii) $B = \{b\}$; and (iii) $|A| > 1$ and $|B| > 1$.

(i) $A = \{a\}$. In this case, we distinguish two sub-cases: (i.1) $x \notin B$ and (i.2) $x \in B$. Consider (i.1). Since $x \notin B$, it follows from $(a, \{a\}) \sim (b, B)$ and (IND) that $(a, \{a, x\}) \sim (b, B \cup \{x\})$. By Proposition 1, $(b, B \cup \{x\}) \sim (b, B \cup \{y\})$. Transitivity of $\succcurlyeq$ then implies that $(a, \{a, x\}) \sim (b, B \cup \{y\})$. Since $(a, \{a, x\}) = (a, A \cup \{x\})$, we obtain (*) in this sub-case. Consider now (i.2), where $x \in B$. To begin with, consider the sub-case where $B \cup \{y\} = \{a, b\}$. Given that $x \in X \setminus A$ and $y \in X \setminus B$, we have $x = b$ and $y = a$, hence $B = \{b\}$. Since $|X| \geq 3$, there exists $c \in X$ such that $c \notin \{a, b\}$. It follows from $(a, \{a\}) \sim (b, \{b\}) = (b, B)$ and (IND) that $(a, \{a, c\}) \sim (b, \{b, c\})$. From Proposition 1, $(a, \{a, b\}) \sim (a, \{a, c\})$, and $(b, \{b, c\}) \sim (b, \{a, b\})$. Then, transitivity of $\succcurlyeq$ implies $(a, \{a, b\}) \sim (b, \{a, b\})$; that is, $(a, \{a, x\}) \sim (b, B \cup \{y\})$. Turn now to the sub-case where $B \cup \{y\} \neq \{a, b\}$. If $y \neq a$, starting with $(a, \{a\}) \sim (b, B)$ and invoking (IND), $(a, \{a, y\}) \sim (b, B \cup \{y\})$. By Proposition 1, $(a, \{a, x\}) \sim (a, \{a, y\})$. Transitivity of $\succcurlyeq$ implies that $(a, \{a, x\}) \sim (b, B \cup \{y\})$. If $y = a$, given that $|X| \geq 3$ and $B \cup \{y\} \neq \{a, b\}$, there exists $z \in B$ such that $z \notin \{a, b\}$. By Proposition 1, $(b, B) \sim (b, (B \cup \{y\}) \setminus \{z\})$. From $(a, \{a\}) \sim (b, B)$, transitivity of $\succcurlyeq$ implies $(a, \{a\}) \sim (b, (B \cup \{y\}) \setminus \{z\})$. Now, noting that $z \neq a$, by (IND), $(a, \{a, z\}) \sim (b, B \cup \{y\})$ holds. From Proposition 1, $(a, \{a, x\}) \sim (a, \{a, z\})$. Transitivity of $\succcurlyeq$ now implies $(a, \{a, x\}) \sim (b, B \cup \{y\})$, which establishes (*) in this sub-case.

(ii) $B = \{b\}$. This case can be treated similarly to case (i).

(iii) $|A| > 1$ and $|B| > 1$. Consider $A', A'' \in K$ such that $\{a, b\} \subset A'' \subset A'$, $|A''| = \min\{|A|, |B|\} > 1$, $|A'| = \max\{|A|, |B|\} > 1$. Since $A \neq X$ and $B \neq X$, the existence of such A' and A'' is guaranteed. It should be clear that there exists $z \in X$ such that $z \notin A'$. If $|A| \geq |B|$, consider (a, A') and (b, A''). From Proposition 1, $(a, A') \sim (b, A'')$ follows from the construction of A' and A'', the assumption that $(a, A) \sim (b, B)$, and transitivity of $\succcurlyeq$. Note that there exists $z \in X \setminus A'$. By (IND), $(a, A' \cup \{z\}) \sim (b, A'' \cup \{z\})$. By virtue of Proposition 1, noting that $|A \cup \{x\}| = |A' \cup \{z\}|$ and $|B \cup \{y\}| = |A'' \cup \{z\}|$, $(a, A \cup \{x\}) \sim (b, B \cup \{y\})$ follows easily from transitivity of $\succcurlyeq$. If $|A| < |B|$, consider (a, A'') and (b, A'). Following a similar argument as above, we can show that $(a, A \cup \{x\}) \sim (b, B \cup \{y\})$. Thus, (*) is proved. The next order of our business is to show that

$$(a, A) \succ (b, B) \Rightarrow (a, A \cup \{x\}) \succ (b, B \cup \{y\}). \qquad (**)$$

As in the proof of (*), we distinguish three cases: (a) $A = \{a\}$; (b) $B = \{b\}$; and (c) $|A| > 1$ and $|B| > 1$.

(a) $A = \{a\}$. (a.1) $x \notin B$. In this sub-case, from $(a, \{a\}) \succ (b, B)$, by (IND), we obtain $(a, \{a, x\}) \succ (b, B \cup \{x\})$. By Proposition 1, $(b, B \cup \{x\}) \sim (b, B \cup \{y\})$. Transitivity of $\succcurlyeq$ implies that $(a, \{a, x\}) = (a, A \cup \{x\}) \succ (b, B \cup \{y\})$. (a.2) $x \in B$. If $B \cup \{y\} =$

$\{a, b\}$, then, given that $x \notin A$ and $y \notin B$, we have $x = b$ and $y = a$. Since $|X| \geq 3$, there exists $c \in X$ such that $c \notin \{a, b\}$. It follows from $(a, \{a\}) \succ (b, \{b\}) = (b, B)$ and (IND) that $(a, \{a, c\}) \succ (b, \{b, c\})$. From Proposition 1, $(a, \{a, b\}) \sim (a, \{a, c\})$ and $(b, \{b, c\}) \sim (b, \{a, b\})$. Transitivity of $\succcurlyeq$ implies $(a, \{a, b\}) \succ (b, \{a, b\})$, viz., $(a, \{a, x\}) = (a, A \cup \{x\}) \succ (b, B \cup \{y\})$. If $B \cup \{y\} \neq \{a, b\}$, we consider (a.2.i) $y \neq a$ and (a.2.ii) $y = a$. Suppose that (a.2.i) $y \neq a$. From $(a, \{a\}) \succ (b, B)$, by (IND), $(a, \{a, y\}) \succ (b, B \cup \{y\})$. By Proposition 1, $(a, \{a, x\}) \sim (a, \{a, y\})$. Transitivity of $\succcurlyeq$ now implies $(a, \{a, x\}) = (a, A \cup \{x\}) \succ (b, B \cup \{y\})$. Suppose next that (a.2.ii) $y = a$. Since $|X| \geq 3$ and $B \cup \{y\} \neq \{a, b\}$, there exists $c \in B$ such that $c \notin \{a, b\}$. By Proposition 1, $(b, B) \sim (b, (B \cup \{y\})\backslash\{c\})$. From $(a, \{a\}) \succ (b, B)$, by transitivity of $\succcurlyeq$, $(a, \{a\}) \succ (b, (B \cup \{y\})\backslash\{c\})$. Now, noting that $c \neq a$, by (IND), $(a, \{a, c\}) \succ (b, B \cup \{y\})$. From Proposition 1, $(a, \{a, c\}) \sim (a, \{a, x\})$. Transitivity of $\succcurlyeq$ implies $(a, \{a, x\}) = (a, A \cup \{x\}) \succ (b, B \cup \{y\})$.

(b) $B = \{b\}$. If $x \notin B$, it follows from $(a, A) \succ (b, B)$ and (IND) that $(a, A \cup \{x\}) \succ (b, \{b, x\})$. By Proposition 1, $(b, \{b, x\}) \sim (b, \{b, y\}) = (b, B \cup \{y\})$. Transitivity of $\succcurlyeq$ now implies $(a, A \cup \{x\}) \succ (b, \{b, y\}) = (b, B \cup \{y\})$. If $x \in B$, then $x = b$. Consider first the case where $y = a$. If $A = \{a\}$, it follows from (a) that $(a, \{a, x\}) = (a, A \cup \{x\}) \succ (b, \{b, y\}) = (b, B \cup \{y\})$. Suppose $A \neq \{a\}$. Given that $x = b$, $y = a$, $x \notin A$, and $y \notin B$, and noting that $|X| \geq 3$, there exists $c \in A\backslash\{a, b\}$. From Proposition 1, $(a, (A \cup \{x\})\backslash\{c\}) \sim (a, A)$. From transitivity of $\succcurlyeq$ and noting that $(a, A) \succ (b, \{b\})$, $(a, (A \cup \{x\})\backslash\{c\}) \succ (b, \{b\})$ holds. By (IND), $(a, A \cup \{x\}) \succ (b, \{b, c\})$. From Proposition 1, $(b, \{b, y\}) \sim (b, \{b, c\})$. Therefore, $(a, A \cup \{x\}) \succ (b, \{b, y\}) = (b, B \cup \{y\})$ follows easily from transitivity of $\succcurlyeq$. Consider next that $y \neq a$. If $y \notin A$, then, by (IND) and $(a, A) \succ (b, \{b\})$, we obtain $(a, A \cup \{y\}) \succ (b, \{b, y\})$ immediately. By Proposition 1, $(a, A \cup \{y\}) \sim (a, A \cup \{x\})$. Transitivity of $\succcurlyeq$ implies $(a, A \cup \{x\}) \succ (b, \{b, y\}) = (b, B \cup \{y\})$. If $y \in A$, noting that $y \neq a$, $y \notin B$, and $x = b$, we have $|(A \cup \{x\})\backslash\{y\}| = |A|$. By Proposition 1, $(a, A) \sim (a, (A \cup \{x\})\backslash\{y\})$. Transitivity of $\succcurlyeq$ implies $(a, (A \cup \{x\})\backslash\{y\}) \succ (b, \{b\})$. By (IND), it then follows that $(a, A \cup \{x\}) \succ (b, \{b, y\}) = (b, B \cup \{y\})$.

(c) $|A| > 1$ and $|B| > 1$. This case is similar to case (iii) above, and we may safely omit it.

Thus, (**) is proved. (*) together with (**) completes the proof of Proposition 3. □

Proof of Theorem 1. It can be easily shown that if $\succcurlyeq$ is extremely consequential, then it satisfies (IND), (SI), and (LI). Therefore, we have only to prove that, if $\succcurlyeq$ satisfies (IND), (SI), and (LI), then, for all $(x, A), (x, B) \in \Omega$, $(x, A) \sim (x, B)$ holds.

Let $\succcurlyeq$ satisfy (IND), (SI), and (LI). First, observe that from Proposition 1, we have the following:

$$\text{For all } (x, A), (x, B) \in \Omega, |A| = |B| \Rightarrow (x, A) \sim (x, B). \tag{T1.1}$$

Thus, we have only to show that

$$\text{For all } (x, A), (x, B) \in \Omega, |A| > |B| \Rightarrow (x, A) \sim (x, B). \tag{T1.2}$$

From Proposition 2, and by (LI) and the completeness of $\succcurlyeq$, it must be true that

$$\text{For all distinct } x, y \in X, (x, \{x, y\}) \sim (x, \{x\}). \tag{T1.3}$$

From (T1.3), by the repeated use of (IND), (T1.1), and the transitivity of $\succcurlyeq$, (T1.2) can be established. □

Proof of Theorem 2. Again, it can be shown that if $\succcurlyeq$ is strongly consequential, then it satisfies (IND), (SI), (LSM), and (ROB). Therefore, we have only to prove that, if $\succcurlyeq$ satisfies (IND), (SI), (LSM), and (ROB), then, for all $(x, A), (y, B) \in \Omega$, $(x, \{x\}) \sim (y, \{y\})$ implies $[(x, A) \succcurlyeq (y, B) \Leftrightarrow |A| \geq |B|]$, and $(x, \{x\}) \succ (y, \{y\})$ implies $(x, A) \succ (y, B)$.

Let $\succcurlyeq$ satisfy (IND), (SI), (LSM), and (ROB). Note that, from Proposition 1, we have the following:

$$\text{For all } x \in X \text{ and all } (x, A), (x, B) \in \Omega, |A| = |B| \Rightarrow (x, A) \sim (x, B). \tag{T2.1}$$

Next, from Proposition 2, and by (LSM) and the completeness of $\succcurlyeq$, it must be true that

$$\text{For all distinct } x, y \in X, (x, \{x, y\}) \succ (x, \{x\}). \tag{T2.2}$$

From (T2.2) and by the repeated use of (IND), we can derive the following:

$$\text{For all } x \in X \text{ and all } (x, A), (x, B) \in \Omega, |A| > |B| \Rightarrow (x, A) \succ (x, B). \tag{T2.3}$$

Now, for all $x, y \in X$, consider $(x, \{x\})$ and $(y, \{y\})$. If $(x, \{x\}) \sim (y, \{y\})$, then, since X contains at least three alternatives, by IND, for all $z \in X \backslash \{x, y\}$, we must have $(x, \{x, z\}) \sim (y, \{y, z\})$. From (T2.1) and by the transitivity of $\succcurlyeq$, we then have $(x, \{x, y\}) \sim (y, \{x, y\})$. Then, by IND, we have $(x, \{x, y, z\}) \sim (y, \{x, y, z\})$. Since the opportunity sets are finite, by the repeated application of (T2.1) and (T2.3), the transitivity of $\succcurlyeq$, and (IND), we then obtain

$$\begin{aligned}&\text{For all } x, y \in X \text{ and all } (x, A), (y, B) \in \Omega, \text{ if } (x, \{x\}) \sim (y, \{y\}),\\ &\text{then } |A| \geq |B| \Leftrightarrow (x, A) \succcurlyeq (y, B).\end{aligned} \tag{T2.4}$$

If, on the other hand, $(x, \{x\}) \succ (y, \{y\})$, then, for all $z \in X$, (ROB) implies $(x, \{x\}) \succ (y, \{y, z\})$. Since opportunity sets are finite, by repeated use of (ROB), we then obtain $(x, \{x\}) \succ (y, A)$ for all $(y, A) \in \Omega$. Therefore, from (T2.1) and (T2.3), and by the transitivity of $\succcurlyeq$, we obtain

$$\text{For all } x, y \in X \text{ and all } (x, A), (y, B) \in \Omega, \text{ if } (x, \{x\}) \succ (y, \{y\}), \text{ then } (x, A) \succ (y, B). \tag{T2.5}$$

(T2.5), together with (T2.1), (T2.3), and (T2.4), completes the proof. □

Proof of Proposition 4. Let $\succcurlyeq$ satisfy (IND), (SI), and (SPO). Let $x \in X$. For all $y \in X \setminus \{x\}$, by (SPO), $(x, \{x, y\}) \succ (y, \{y\})$. Then, (IND) implies $(x, \{x, y, z\}) \succ (y, \{y, z\})$ for all $z \in X \setminus \{x, y\}$. By (SI), $(y, \{y, z\}) \sim (y, \{x, y\})$. It follows from the transitivity of $\succcurlyeq$ that $(x, \{x, y, z\}) \succ (y, \{x, y\})$. By (SPO), $(y, \{x, y\}) \succ (x, \{x\})$. Then, $(x, \{x, y, z\}) \succ (x, \{x\})$ follows from the transitivity of $\succcurlyeq$. Therefore, for $A = \{x, y, z\}$, $(x, A) \succ (x, \{x\})$ holds. Thus, $\succcurlyeq$ satisfies (LSM). □

Proof of Theorem 3. It can be checked that if $\succcurlyeq$ is extremely non-consequential, then it satisfies (IND), (SI), (LSM), and (INS). Therefore, we have only to prove that if $\succcurlyeq$ satisfies (IND), (SI), (LSM), and (INS), then, for all $(x, A), (y, B) \in \Omega$, $|A| \geq |B| \Leftrightarrow (x, A) \succcurlyeq (y, B)$.

Let $\succcurlyeq$ satisfy (IND), (SI), (LSM), and (INS). First, we note that, following a similar method to that used for proving (T2.3), the following can be established:

$$\text{For all } (x, A), (x, B) \in \Omega, |A| > |B| \Rightarrow (x, A) \succ (x, B). \tag{T3.1}$$

Together with Proposition 1 and recollecting that $\succcurlyeq$ is complete, we must have the following:

$$\text{For all } (x, A), (x, B) \in \Omega, |A| > |B| \Leftrightarrow (x, A) \succ (x, B). \tag{T3.2}$$

Now, for all $x, y \in X$, it follows from (INS) that $(x, \{x\}) \sim \{y, \{y\})$. For all $z \in X \setminus \{x, y\}$, by (IND), $(x, \{x, z\}) \sim (y, \{y, z\})$. It follows from (T3.2) and the transitivity of $\succcurlyeq$ that $(x, \{x, y\}) \sim (y, \{x, y\})$. By the repeated use of (T3.2), (IND), and the transitivity of $\succcurlyeq$ and noting that opportunity sets are finite, we can show that

$$\text{For all } (x, A), (y, B) \in \Omega, (x, A) \succcurlyeq (y, B) \Leftrightarrow |A| \geq |B|. \tag{T3.3}$$

Theorem 3 is thus proved. □

Proof of Theorem 4. It can be checked easily that if $\succcurlyeq$ is strongly non-consequential, then it satisfies (IND), (SI), and (SPO). Therefore, we have only to prove that if $\succcurlyeq$ satisfies (IND), (SI), and (SPO), then, for all $(x, A), (y, B) \in \Omega, |A| > |B| \Rightarrow (x, A) \succ (y, B)$ and $|A| = |B| \Rightarrow [(x, \{x\}) \succcurlyeq (y, \{y\}) \Leftrightarrow (x, A) \succcurlyeq (y, B)]$.

Let $\succcurlyeq$ satisfy (IND), (SI), and (SPO). By Proposition 4 and following a similar proof method, we can establish that

$$\text{For all } (x, A), (x, B) \in \Omega, |A| > |B| \Leftrightarrow (x, A) \succ (x, B). \tag{T4.1}$$

For all distinct $x, y \in X$, it follows from (SPO) that $(x, \{x, y\}) \succ (y, \{y\})$. Then, from (T4.1) and by the transitivity of $\succcurlyeq$, $(x, \{x, z\}) \succ (y, \{y\})$ holds for all $z \in X \setminus \{x, y\}$. By virtue of (IND), from $(x, \{x, y\}) \succ (y, \{y\})$, $(x, \{x, y, z\}) \succ (y, \{y, z\})$ holds for all $z \in X \setminus \{x, y\}$. From (T4.1) and by the transitivity of $\succcurlyeq$, we obtain the following:

$$\text{For all } (x, A), (y, B) \in \Omega, \text{ if } |A| = |B| + 1 \text{ and } |B| \leq 2, \text{ then } (x, A) \succ (y, B). \tag{T4.2}$$

From (T4.2), by the repeated use of (IND), (T4.1), and the transitivity of $\succcurlyeq$, coupled with the finiteness of opportunity sets, the following can be established:

$$\text{For all } (x, A), (y, B) \in \Omega, \text{if } |A| = |B| + 1, \text{ then } (x, A) \succ (y, B). \tag{T4.3}$$

From (T4.3), by the transitivity of $\succcurlyeq$ and (T4.1), we have

$$\text{For all } (x, A), (y, B) \in \Omega, \text{if } |A| > |B|, \text{ then } (x, A) \succ (y, B). \tag{T4.4}$$

Consider now $(x, \{x\})$ and $(y, \{y\})$. If $(x, \{x\}) \sim \{y, \{y\})$, following a similar argument as in the proof of Theorem 4, we obtain

$$\text{For all } (x, A), (y, B) \in \Omega, \text{ if } (x, \{x\}) \sim (y, \{y\}) \text{ and } |A| = |B|, \text{ then } (x, A) \sim (y, B). \tag{T4.5}$$

If, on the other hand, $(x, \{x\}) \succ \{y, \{y\})$, we can then follow a similar argument as in the proof of Theorem 2 to obtain

$$\text{For all } (x, A), (y, B) \in \Omega, \text{ if } (x, \{x\}) \succ (y, \{y\}) \text{ and } |A| = |B|, \text{ then } (x, A) \succ (y, B). \tag{T4.6}$$

(T4.6), together with (T4.4) and (T4.5), completes the proof. □

Proof of Theorem 5. We first check the necessity part of the theorem. Suppose $u : X \to \mathcal{R}$ and $f : \mathcal{R} \times \mathcal{Z} \to \mathcal{R}$ are such that (T5.1), (T5.2), and (T5.3) are satisfied.

(SI): Let $x \in X$ and $y, z \in X \backslash \{x\}$. Note that $|\{x, y\}| = |\{x, z\}|$. Therefore, $f(u(x), |\{x, y\}|) = f(u(x), |\{x, z\}|)$, which implies that $(x, \{x, y\}) \sim (x, \{x, z\})$ is true.

(SM): Let $(x, A), (x, B) \in \Omega$ be such that $B \subset A$. Then, $f(u(x), |A|) \geq f(u(x), |B|)$ holds, since f is non-decreasing in each of its arguments and $|A| \geq |B|$. Therefore, $(x, A) \succcurlyeq (x, B)$.

(IND): Let $(x, A), (y, B) \in \Omega$, and $z \in X \backslash A \cup B$. From (T5.3.1), we have

$$f(u(x), |A|) \geq f(u(y), |B|) \Leftrightarrow f(u(x), |A| + 1) \geq f(u(y), |B| + 1)$$
$$\Leftrightarrow f(u(x), |(A \cup \{z\})|) \geq f(u(y), |(B \cup \{z\})|).$$

Therefore, $(x, A) \succcurlyeq (y, B) \Leftrightarrow (x, A \cup \{z\}) \succcurlyeq (y, B \cup \{z\})$.

Next, we show that, if $\succcurlyeq$ satisfies (IND), (SI), and (SM), then there exist a function $f : \mathcal{R} \times \mathcal{Z} \to \mathcal{R}$ and a function $u : X \to \mathcal{R}$ such that (T5.1), (T5.2), and (T5.3) hold. Let $\succcurlyeq$ satisfy (IND), (SI), and (SM). Note that X is finite, and so is Ω. The ordering of $\succcurlyeq$ implies that there exist $u : X \to \mathcal{R}$ and $F : \Omega \to \mathcal{R}$ such that

$$\text{For all } x, y \in X, (x, \{x\}) \succcurlyeq (y, \{y\}) \Leftrightarrow u(x) \geq u(y); \tag{T5.4}$$

$$\text{For all } (x, A), (y, B) \in \Omega, (x, A) \succcurlyeq (y, B) \Leftrightarrow F(x, A) \geq F(y, B). \tag{T5.5}$$

(T5.1) then follows immediately. To show that (T5.2) holds, let $(x, A), (y, B) \in \Omega$ be such that $u(x) = u(y)$ and $|A| = |B|$. From $u(x) = u(y)$, we must have $(x, \{x\}) \sim (y, \{y\})$. Then, by making repeated use of Proposition 3, if necessary, and noting that $|A| = |B|$, $(x, A) \sim (y, B)$ can be obtained easily. Define $\Sigma \subset \mathcal{R} \times \mathcal{Z}$ as follows: $\Sigma := \{(t, i) \in \mathcal{R} \times \mathcal{Z} | \exists (x, A) \in \Omega : t = u(x) \text{ and } i = |A|\}$. Next, define a binary relation $\succcurlyeq^*$ on Σ as follows: For all $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow (u(x), |A|) \succcurlyeq^* (u(y), |B|)$. From the above discussion and noting that $\succcurlyeq$ satisfies (SM) and (IND), the binary relation $\succcurlyeq^*$ defined on Σ is an ordering, and it has the following properties:

(SM′): For all $(t, i), (t, j) \in \Sigma$, if $j \geq i$ then $(t, j) \succcurlyeq^* (t, i)$;

(IND′): For all $(s, i), (t, j) \in \Sigma$, and all integer k, if $i + k \leq |X|$ and $j + k \leq |X|$, then $(s, i) \succcurlyeq^* (t, j) \Leftrightarrow (s, i + k) \succcurlyeq^* (t, j + k)$.

Since Σ is finite and $\succcurlyeq^*$ is an ordering on Σ, there exists a function $f : \mathcal{R} \times \mathcal{Z} \to \mathcal{R}$ such that, for all $(s, i), (t, j) \in \Sigma$, $(s, i) \succcurlyeq^* (t, j)$ iff $f(s, i) \geq f(t, j)$. From the definition of $\succcurlyeq^*$ and Σ, we must have the following: For all $(x, A), (y, B) \in \Omega$, $(x, A) \succcurlyeq (y, B) \Leftrightarrow (u(x), |A|) \succcurlyeq^* (u(y), |B|) \Leftrightarrow f(u(x), |A|) \geq f(u(y), |B|)$. To prove that f is nondecreasing in each of its arguments, we first consider the case in which $u(x) \geq u(y)$ and $|A| = |B|$. Given $u(x) \geq u(y)$, it follows from the definition of u that $(x, \{x\}) \succcurlyeq (y, \{y\})$. Noting that $|A| = |B|$, by the repeated use of Proposition 3, if necessary, we must have $(x, A) \succcurlyeq (y, B)$. Thus, f is nondecreasing in its first argument. To show that f is nondecreasing in its second argument as well, we consider the case in which $u(x) = u(y)$ and $|A| \geq |B|$. From $u(x) = u(y)$, we must have $(x, \{x\}) \sim (y, \{y\})$. Then, from the earlier argument, $(x, A') \sim (y, B)$ for some $A' \subset A$ such that $|A'| = |B|$. Now, by (SM), $(x, A) \succcurlyeq (x, A')$.

Then, $(x, A) \succcurlyeq (y, B)$ follows from the transitivity of $\succcurlyeq$. Therefore, f is nondecreasing in each of its arguments. Finally, (T5.3.1) follows clearly from (IND′). □

Proof of Theorem 6. Let $\succcurlyeq$ satisfy (IND), (SI), and (SM). By Proposition 1, we have,

$$\text{For all } (a, A), (a, B) \in \Omega, \text{ if } |A| = |B|, \text{ then } (a, A) \sim (a, B). \tag{T6.3}$$

By Proposition 3, the following can be shown to be true:

$$\text{For all } x, y \in X \text{ and all } (x, A), (y, A) \in \Omega, (x, \{x\}) \succcurlyeq (y, \{y\}) \Leftrightarrow (x, A) \succcurlyeq (y, A). \tag{T6.4}$$

We now show that, for all $(x, A), (y, B) \in \Omega$, if $(x, \{x\}) \sim (y, \{y\})$ and $|A| = |B|$, then $(x, A) \sim (y, B)$. Let $C \in K$ be such that $|C| = |A| = |B|$ and $\{x, y\} \subset C$. From (T6.4), we have $(x, C) \sim (y, C)$. Note that $(x, C) \sim (x, A)$ and $(y, C) \sim (y, B)$ follow from (T6.3). By transitivity of $\sim$, we have $(x, A) \sim (y, B)$. Define a binary relation $\succcurlyeq^{\#}$ on $X \times \mathcal{Z}$ as follows: For all $x, y \in X$ and all positive integers i, j, $(x, i) \succcurlyeq^{\#} (y, j) \Leftrightarrow [(x, A) \succcurlyeq (y, B)$ for some $A, B \in K$ such that $x \in A$, $y \in B$, $i = |A|$, $j = |B|]$. From the above discussion, $\succcurlyeq^{\#}$ is well-defined and is an ordering. A similar method of proving (T5.3) can be invoked to prove that (T6.2) holds. □

Proof of Theorem 7. From Theorem 6, we know that there exists an ordering $\succcurlyeq^{\#}$ on $X \times \mathcal{Z}$ such that

$$\text{For all } (x, A), (y, B) \in \Omega, (x, A) \succcurlyeq (y, B) \Leftrightarrow (x, |A|) \succcurlyeq^{\#} (y, |B|); \tag{T7.3}$$

$$\text{For all integers } i, j, k \geq 1 \text{ and all } x, y \in X, (x, i) \succcurlyeq^{\#} (y, j) \Leftrightarrow (x, i + k) \succcurlyeq^{\#} (y, j + k);$$
$$\text{and } (x, i + k) \succcurlyeq^{\#} (x, i). \tag{T7.4}$$

For all $(x, i) \in X \times \mathcal{Z}$ and all $j \in \mathcal{Z}$, consider the sets $U(x; i, j)$ and $L(x; i, j)$ defined as follows:

$$U(x; i, j) = \{y \in X | (y, j) \succcurlyeq^{\#} (x, i)\}, L(x; i, j) = \{y \in X | (x, i) \succcurlyeq^{\#} (y, j)\}.$$

From (T7.3), by (CON), it is clear that both $U(x; i, j)$ and $L(x; i, j)$ are closed. Note that $X = \mathcal{R}_+^n$. By Broome (2003), there is a function $v : X \times \mathcal{Z} \to \mathcal{R}$, which is continuous in its first argument, that represents $\succcurlyeq^{\#}$. Therefore, (T7.1) obtains. (T7.2) follows from (T7.1), (T7.3), and (T7.4). □

Proof of Theorem 8. Suppose that there exists an ESWF f on $D_f(E)$ which satisfies (SP) and (IIA(i)). Since all individuals are extreme consequentialists,

$$\forall i \in N : (x, A) R_i (y, B) \Leftrightarrow (x, X) R_i (y, X) \tag{T8.1}$$

holds for all $(x, A), (y, B) \in \Omega$ and for all $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f(E)$. Note that the conditions (IND), (SI), and (SM) impose no restriction whatsoever on the profile $\mathbf{R} = (R_1, R_2, \ldots, R_n)$ even when, for each and every $i \in N$, R_i is restricted on $\Omega_X := \{(x, X) \in X \times K | x \in X\}$. Note also that (SP) and (IIA(i)) imposed on f imply that the same conditions must be satisfied on the restricted space Ω_X. By virtue of the Arrow impossibility theorem, therefore, there exists a dictator, say $d \in N$, for f on the restricted space Ω_X. That is, for all $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f(E)$ and all $(x, X), (y, X) \in \Omega_X$, $(x, X) P(R_d) (y, X) \Rightarrow (x, X) P(R) (y, X)$, where $R = f(\mathbf{R})$. We now show that for all $(x, A), (y, B) \in \Omega$, $(x, A) P(R_d) (y, B) \Rightarrow (x, A) P(R) (y, B)$; viz. d is a dictator for f on

the full space Ω. Note that since d is an extreme consequentialist, $(x, A)P(R_d)(y, B)$ if and only if $(x, X)P(R_d)(y, X)$. Since all individuals are extreme consequentialists, it must be true that $(x, A)I(R_i)(x, X)$ and $(y, B)I(R_i)(y, X)$ for all $i \in N$. Therefore, by (SP), $(x, A)I(R)(x, X)$ and $(y, B)I(R)(y, X)$. By virtue of the transitivity of R, it then follows that $(x, X)P(R)(y, X) \Rightarrow (x, A)P(R)(y, B)$. That is, we have shown that $(x, A)P(R_d)(y, B) \Rightarrow (x, A)P(R)(y, B)$. In other words, d is a dictator for f on the full space Ω. Therefore, there exists no ESWF that satisfies (SP), (IIA(i)), and (ND).

A similar argument can be used to show that there exists no ESWF that satisfies (SP), (IIA(ii)), and (ND). □

Proof of Theorem 9. Let $e \in N$ be a uniform extreme consequentialist and $s \in N$ be a uniform strong consequentialist. By definition,

$$\forall(x, A), (x, B) \in \Omega : (x, A)I(R_e)(x, B), \tag{T9.1}$$

$$\forall(x, A), (y, B) \in \Omega : (x, \{x\})I(R_s)(y, \{y\}) \Rightarrow [(x, A)R_s(y, B) \Leftrightarrow |A| \geq |B|], \tag{T9.2}$$

and

$$\forall(x, A), (y, B) \in \Omega : (x, \{x\})P(R_s)(y, \{y\}) \Rightarrow (x, A)P(R_s)(y, B) \tag{T9.3}$$

hold. Now consider the following ESWF:

$$\begin{aligned}
&\forall(x, A), (y, B) \in \Omega : \\
&(x, \{x\})P(R_s)(y, \{y\}) \Rightarrow [(x, A)R(y, B) \Leftrightarrow (x, A)R_s(y, B)]; \\
&(x, \{x\})I(R_s)(y, \{y\}) \Rightarrow [(x, A)R(y, B) \Leftrightarrow (x, A)R_e(y, B)], \\
&\text{where } R = f(\mathbf{R}).
\end{aligned}$$

It may easily be verified that the above ESWF satisfies (SP) and (ND). To verify that it satisfies both (IIA(i)) and (IIA(ii)), we consider $(x, A), (y, B) \in \Omega$, and $\mathbf{R} = (R_1, R_2, \ldots, R_n)$, $\mathbf{R}' = (R'_1, R'_2, \ldots, R'_n) \in D_f(E \cup S)$. Let $R = f(\mathbf{R})$ and $R' = f(\mathbf{R}')$.

To begin with, suppose that we have $(x, A)R_i(y, B) \Leftrightarrow (x, A)R'_i(y, B)$ as well as $(x, \{x\})R_i(y, \{y\}) \Leftrightarrow (x, \{x\})R'_i(y, \{y\})$ for all $i \in N$. If $(x, \{x\})P(R_s)(y, \{y\})$, then $(x, \{x\})P(R'_s)(y, \{y\})$, $(x, A)P(R_s)(y, B)$, as well as $(x, A)P(R'_s)(y, B)$. Thus, the ESWF gives us $(x, A)P(R)(y, B)$ and $(x, A)P(R')(y, B)$. Secondly, if $(y, \{y\})P(R_s)(x, \{x\})$, then $(y, \{y\})P(R'_s)(x, \{x\})$, $(y, B)P(R_s)(x, A)$, and $(y, B)P(R'_s)(x, A)$. Thus, the ESWF gives us $(y, B)P(R)(x, A)$ and $(y, B)P(R')(x, A)$. Thirdly, if $(x, \{x\})I(R_s)(y, \{y\})$, then $(x, \{x\})I(R'_s)(y, \{y\})$. Thus, the ESWF implies that $(x, A)R(y, B) \Leftrightarrow (x, A)R_e(y, B)$ and $(x, A)R'(y, B) \Leftrightarrow (x, A)R'_e(y, B)$. Note that individual e is an extreme consequentialist. It is therefore clear that, in this case, if $(x, A)R_e(y, B) \Leftrightarrow (x, A)R'_e(y, B)$, then $(x, A)R(y, B) \Leftrightarrow (x, A)R'(y, B)$. Therefore, (IIA(i)) is satisfied.

Next, suppose that $|A| = |B|$ and that $[(x, A)R_i(y, B) \Leftrightarrow (x, A)R'_i(y, B)]$ for all $i \in N$. To show that $(x, A)R(y, B) \Leftrightarrow (x, A)R'(y, B)$ in this case, we observe that, when $|A| = |B|$, $(x, A)R_s(y, B) \Leftrightarrow (x, \{x\})R_s(y, \{y\})$ and $(x, A)R'_s(y, B) \Leftrightarrow (x, \{x\})R'_s(y, \{y\})$. Then the proof that the above ESWF satisfies (IIA(ii)) is similar to the proof showing that the ESWF satisfies (IIA(i)). We have only to note that the individual e is an extreme consequentialist.

The binary relation R generated by this ESWF is clearly reflexive and complete. We now show that R is transitive. Let $(x, A), (y, B)$ and $(z, C) \in \Omega$ be such that $(x, A)R(y, B)$ and $(y, B)R(z, C)$. Note that, since $(x, A)R(y, B)$, by the ESWF constructed above, we cannot have $(y, \{y\})P(R_s)(x, \{x\})$. Then, by the completeness of R_s, there are only two cases to be distinguished, and considered separately: (a) $(x, \{x\})I(R_s)(y, \{y\})$; (b) $(x, \{x\})P(R_s)(y, \{y\})$.

Case (a): In this case, we must have $(x, A)R_e(y, B)$. If $(y, \{y\})I(R_s)(z, \{z\})$, then it follows from $(y, B)R(z, C)$ that $(y, B)R_e(z, C)$. Then, the transitivity of R_e implies $(x, A)R_e(z, C)$. By the transitivity of R_s, $(x, \{x\})I(R_s)(z, \{z\})$. Therefore, $(x, A)R(z, C)$ if and only if $(x, A)R_e(z, C)$. Hence, $(x, A)R(z, C)$ follows from $(x, A)R_e(z, C)$. If $(y, \{y\})P(R_s)(z, \{z\})$, then, by the transitivity of R_s, it follows that $(x, \{x\})P(R_s)(z, \{z\})$. Therefore, $(x, A)R(z, C)$ if and only if $(x, A)R_s(z, C)$. Since s is a strong consequentialist, given that $(x, \{x\})P(R_s)(z, \{z\})$, we must have $(x, A)P(R_s)(z, C)$. Therefore, $(x, A)P(R)(z, C)$. Hence, $(x, A)R(z, C)$ holds. Note that, given $(y, B)R(z, C)$, we cannot have $(z, \{z\})P(R_s)(y, \{y\})$. Therefore, the transitivity of R holds in this case.

Case (b): In this case, we must have $(x, A)P(R_s)(y, B)$, hence $(x, A)P(R)(y, B)$. Since $(y, B)R(z, C)$, we must then have $(y, \{y\})R_s(z, \{z\})$. By the transitivity of R_s, it follows that $(x, \{x\})P(R_s)(z, \{z\})$. Thus, $(x, A)P(R_s)(z, C)$ follows from s being a strong consequentialist. By construction, in this case, $(x, A)R(z, C)$ if and only if $(x, A)R_s(z, C)$. Hence, $(x, A)P(R)(z, C)$. Therefore, the transitivity of R holds in this case.

Combining the cases (a) and (b), the transitivity of R is proved. □

Proof of Theorem 10. Let $n^* \in N$ be a uniform strong non-consequentialist over $D_f(N)$. Then, for all $\mathbf{R} = (R_1, R_2, \ldots, R_n) \in D_f(N)$ and all $(x, A), (y, B) \in \Omega$, it follows from $|A| > |B|$ that $(x, A)P(R_{n^*})(y, B)$. Consider now the following ESWF f: For all $(x, A), (y, B) \in \Omega$,

$$\text{if } |A| > |B|, \text{ then } (x, A)P(R)(y, B);$$

$$\text{if } |A| = |B| = 1, \text{ then } (x, \{x\})R(y, \{y\}) \text{ if and only if } (x, \{x\})R_1(y, \{y\});$$

$$\text{if } |A| = |B| = 2, \text{ then } (x, A)R(y, B) \text{ if and only if } (x, A)R_2(y, B);$$

$$\vdots$$

$$\text{if } A = B = X, \text{ then } (x, A)R(y, B) \text{ if and only if } (x, A)R_k(y, B), \text{ where } k = \min\{|N|, |X|\},$$

where $R = f(\mathbf{R})$. It is easy to verify that this f satisfies (SP), (FIIA), and (ND). It is also clear that R generated by this ESWF is reflexive and complete. We now show that R is transitive as well. Let $(x, A), (y, B), (z, C) \in \Omega$ be such that $(x, A)R(y, B)$ and $(y, B)R(z, C)$. Then, clearly, $|A| \geq |B|$ and $|B| \geq |C|$. If $|A| > |B|$ or $|B| > |C|$, then $|A| > |C|$. By the constructed ESWF, $(x, A)P(R)(z, C)$ follows easily. Thus the transitivity of R holds for this case. Now, suppose $|A| = |B| = |C|$. Note that in this case, for all $(a, G), (b, H) \in \Omega$ such that $|G| = |H| = |A|$, $(a, G)R(b, H)$ if and only if $(a, G)R_k(b, H)$, where $k \in N$ and $k = \min\{|N|, |A|\}$. Therefore, the transitivity of R follows from the transitivity of R_k. The above two cases exhaust all the possibilities. Therefore R is transitive. □

Proof of Theorem 11. It can be verified that if $\succcurlyeq$ is a consequentialist in nature, then it satisfies both (LI*) and (M*). Suppose now that $\succcurlyeq$ satisfies (LI*) and (M*). We need

to show that $\succcurlyeq$ must be a consequentialist; i.e. for all $(m(x), x, A), (m(y), y, B) \in \Omega$, $[m(x) = m(y), x = y] \Rightarrow (m(x), x, A) \sim (m(y), y, B)$. Let $(m(x), x, A), (m(y), y, B) \in \Omega$ be such that $[m(x) = m(y), x = y]$. By (LI*), $(m(x), x, \{(x, 1-x)\}) \sim (m(x), x, X)$. Note that A and B must be such that $\{(x, 1-x)\} \subseteq A \subseteq X$ and $\{(x, 1-x)\} \subseteq B \subseteq X$. By (M*), it then follows that $(m(x), x, X) \succcurlyeq (m(x), x, A) \succcurlyeq (m(x), x, \{(x, 1-x)\})$ and $(m(x), x, X) \succcurlyeq (m(x), x, B) \succcurlyeq (m(x), x, \{(x, 1-x)\})$. Noting that $(m(x), x, \{(x, 1-x)\}) \sim (m(x), x, X)$, it then follows easily that $(m(x), x, \{(x, 1-x)\}) \sim (m(x), x, A)$ and $(m(x), x, \{(x, 1-x)\}) \sim (m(x), x, B)$. The transitivity of $\succcurlyeq$ now implies that $(m(x), x, A) \sim (m(x), x, B)$. □

Proof of Theorem 12. It can be verified that if $\succcurlyeq$ is an extreme consequentialist in nature, then it satisfies (LI*), (M*), and (CINS*). Suppose now that $\succcurlyeq$ satisfies (LI*), (M*), and (CINS*). We need to show that $\succcurlyeq$ is an extreme consequentialist; i.e., for all $(m(x), x, A), (m(y), y, B) \in \Omega$, $m(x) = m(y) \Rightarrow (m(x), x, A) \sim (m(y), y, B)$. Let $\succcurlyeq$ over Ω satisfy (LI*), (M*), and (CINS*), and $(m(x), x, A), (m(y), y, B) \in \Omega$ be such that $m(x) = m(y)$. From Theorem 8, we have $(m(x), x, A) \sim (m(x), x, \{(x, 1-x)\})$ and $(m(y), y, B) \sim (m(y), y, \{(y, 1-y)\})$. By (CINS*) and noting that $m(x) = m(y)$, we have $(m(x), x, \{(x, 1-x)\}) \sim (m(y), y, \{(y, 1-y)\})$. The transitivity of $\succcurlyeq$ now implies that $(m(x), x, A) \sim (m(y), y, B)$, as defined. □

Proof of Theorem 13. The proof is similar to that of Theorem 3, and therefore we omit it. □

References

Arrow, K. J. (1951). *Social Choice and Individual Values*. New York: Wiley. 2nd edn., 1963.

Bolton, G. E., Brandts, J., and Ockenfels, A. (2005). Fair Procedures: Evidence from Games Involving Lotteries. *Economic Journal*, 117, 1054–76.

Bossert, W., Pattanaik, P. K., and Xu, Y. (1994). Ranking Opportunity Sets: An Axiomatic Approach. *Journal of Economic Theory*, 63, 326–45.

——— ——— ——— (2003). Similarity of Options and the Measurement of Diversity. *Journal of Theoretical Politics*, 15, 405–21.

Broome, J. (2003). Representing an Ordering When the Population Varies. *Social Choice and Welfare*, 20, 243–6.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A Tractable Model of Reciprocity and Fairness. *Games and Economic Behavior*, 99/1, 17–45.

——— ——— and Sadiraj, V. (2008). Revealed Altruism. *Econometrica*, 76/1, 31–69.

Dworkin, G. (1982). Is More Choice Better Than Less? In P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds.). *Midwest Studies in Philosophy, VII: Social and Political Philosophy*, 47–61, Minneapolis: University of Minnesota Press.

Frey, B. S., and Stutzer, A. (2004). Beyond Outcomes: Measuring Procedural Utility. *Oxford Economic Papers*, 57, 90–111.

Gaertner, W., and Xu, Y. (2004). Procedural Choice. *Economic Theory*, 24, 335–49.

GRAVEL, N. (1994). Can a Ranking of Opportunity Sets Attach an Intrinsic Importance to Freedom of Choice? *American Economic Review: Papers and Proceedings*, 84, 454–8.

—— (1998). Ranking Opportunity Sets on the Basis of their Freedom of Choice and their Ability to Satisfy Preferences: A Difficulty. *Social Choice and Welfare*, 15, 371–82.

HANSSON, S. O. (1992). A Procedural Model of Voting. *Theory and Decision*, 32, 269–301.

—— —— (1996). Social Choice with Procedural Preferences. *Social Choice and Welfare*, 13, 215–30.

IWATA, Y. (2006). Consequences, Opportunities, and Arrovian Theorems with Consequentialist Domains, Paper presented at the Spring Meeting of the Japanese Economic Association.

JONES, P., and SUGDEN, R. (1982). Evaluating Choices. *International Review of Law and Economics*, 2, 47–65.

LINDBECK, A. (1988). Individual Welfare and Welfare State Policy. *European Economic Review*, 32, 295–318.

PATTANAIK, P. K., and SUZUMURA, K. (1994). Rights, Welfarism and Social Choice. *American Economic Review: Papers and Proceedings*, 84, 435–439.

—— (1996). Individual Rights and Social Evaluation. *Oxford Economic Papers*, 48, 194–212.

—— and XU, Y. (1990). On Ranking Opportunity Sets in Terms of Freedom of Choice. *Recherches Economiques de Louvain*, 56, 383–90.

—— —— (2000). On Diversity and Freedom of Choice. *Mathematical Social Sciences*, 40, 123–30.

—— —— (2006). Ordinal Distance, Dominance, and the Measurement of Diversity, Working Paper, 06–57, Andrew Young School of Policy Studies, Georgia State University.

RABIN, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83, 1281–302.

—— (2002). A Perspective on Psychology and Economics. *European Economic Review*, 46, 657–85.

SCHEFFLER, S. (1988). *Consequentialism and its Critics*. Oxford: Oxford University Press.

SCHUMPETER, J. A. (1942). *Capitalism, Socialism, and Democracy*. New York: Harper & Brothers.

SEN, A. K. (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.

—— (1988). Freedom of Choice: Concept and Content. *European Economic Review*, 32, 269–94.

—— (1993). Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedom. *Oxford Economic Papers*, 45, 519–41.

—— (1995). Rationality and Social Choice. *American Economic Review*, 85, 1–24.

—— (1996). Maximization and the Act of Choice. *Econometrica*, 65, 745–79.

SIMON, H. A. (1976). From Substantive to Procedural Rationality. In S. J. Latsis (ed.). *Methods and Appraisal in Economics*. New York: Cambridge University Press.

—— (1978). Rationality as a Process and Product of Thought. *American Economic Review*, 68, 1–16.

SUZUMURA, K. (1999). Consequences, Opportunities, and Procedures. *Social Choice and Welfare*, 16, 17–40.

——(2000). Welfare Economics beyond Welfarist-Consequentialism. *Japanese Economic Review*, 51, 1–32.
—— and Xu, Y. (2001). Characterizations of Consequentialism and Non-Consequentialism. *Journal of Economic Theory*, 101, 423–36.
————(2003). Consequences, Opportunities, and Generalized Consequentialism and Non-Consequentialism. *Journal of Economic Theory*, 111, 293–304.
————(2004). Welfarist-Consequentialism, Similarity of Attitudes, and Arrow's General Impossibility Theorem. *Social Choice and Welfare*, 22, 237–51.
—— and Yoshihara, N. (2007). On Initial Conferment of Individual Rights, Working Paper, Institute of Economic Research, Hitotsubashi University.

CHAPTER 15

FREEDOM OF CHOICE

KEITH DOWDING
MARTIN VAN HEES

15.1 Introduction

There are many reasons why one might be interested in human freedom. One argument, persuasively made by Amartya Sen, is that a person's well-being is partly dependent on the freedom the person enjoys. In order to assess the well-being of human beings, we need information about how free they are. Another consideration arises in a political context. Freedom of choice is generally considered to be a good thing, with greater choice better than less. Any theory of social justice claiming such freedom is important, and that individuals should be as free as possible, requires some idea of how it can be measured. Naturally, then, any problems encountered in measuring freedom in general, or freedom of choice in particular, reverberates throughout any libertarian claims.

Fitting the importance of the subject, there is by now an extensive literature using an axiomatic-deductive approach to the measurement of freedom of choice. This chapter aims to provide an introduction to this literature, to point out some problems with it, and to discuss avenues for further research. In Section 15.2 we first present a result established by Pattanaik and Xu (1990), which gives an axiomatic characterization of an extremely simple and counter-intuitive measurement: to wit, the cardinality rule which says that the more options a person has, the more

freedom of choice he possesses. We distinguish two main responses to this rule, which focus on what we label as the *diversity* and *opportunity* issue, respectively. The analysis of the diversity issue is based on the idea that the cardinality rule is flawed for failing to incorporate information about the differences between alternatives. The second line, focusing on the opportunity issue, assumes that any convincing measurement of freedom of choice should refer to the preferences that individuals have over the various options. Since the diversity issue is usually addressed without recourse to preferences, we can also describe the two lines in the literature as *the non-preference-based* and the *preference-based* approaches to the measurement of freedom, respectively.[1] After presenting the outlines of these two approaches, and some of the alternative measurements arising from them, we argue (in Section 15.5) that both approaches neglect information that might be relevant to the measurement of freedom of choice, i.e. information about the things that individuals are *not* free to do. In Section 15.6 we query what is being attempted in the literature. Is it trying to measure the *extent* of a person's freedom of choice or the *value* of it? We argue that, if we take it to be measuring the extent of freedom, the differences between the two types of approach can be explained in terms of a difference in their underlying assumptions concerning the definition of freedom. We argue subsequently that if the proposed rankings concern the value of freedom, there are important elements in the overall assessment of the value of freedom that are not captured by any of the axiomatic formulations: viz. the costs of freedom. More choice need not be undeniably superior to less; a nonlinear relationship may exist, with maximal freedom of choice not necessarily being optimal. We conclude the chapter by suggesting some new lines of inquiry.

15.2 The Cardinality Ranking

The axiomatic-deductive approach adopted to address the question of how much freedom of choice an individual enjoys begins by assuming that an individual is confronted with an opportunity set consisting of mutually exclusive alternatives from which she might choose exactly one. The alternatives are usually taken to be commodity bundles, but they may, for instance, also stand for actions.

If S denotes the set of all possible alternatives, an opportunity set is a nonempty subset of this set S (unless stated otherwise, it is here taken to be finitely large).

[1] Exceptions to the separate treatment of the two issues are Gravel and Bervoets (2004) and Peragine and Romero-Medina (2006). In these contributions, rankings are characterized on the basis of both information about the (dis)similarity between alternatives and preferences.

Each opportunity set describes a possible choice situation, and the question is how to compare these choice situations in terms of the degree of freedom of choice they offer the individual. Stated more formally, the question of the measurement of freedom of choice concerns the derivation of an individual freedom ranking $\succcurlyeq$ (to be interpreted as "gives at least as much freedom of choice as") over the set of all possible nonempty subsets of S.

In a seminal paper, Pattanaik and Xu (1990) presented three conditions—in the form of axioms—that a freedom measurement should satisfy. They then showed that there is only one measurement that satisfies all three.[2] Their first axiom states the idea that opportunity sets consisting of one alternative only all yield the same amount of freedom.

Axiom 1 (Indifference between No-Choice Situations (INS)). For all $x, y \in S$, $\{x\} \sim \{y\}$.

The idea underlying this axiom is that singleton sets do not offer any freedom of choice at all, since, by assumption, an individual always has to choose exactly one alternative from an opportunity set. The next axiom expresses the fact that situations that offer at least some choice give more freedom of choice.

Axiom 2 (Strict Monotonicity (SM)). For all distinct alternatives $x, y \in S$ $(x \neq y)$, $\{x, y\} \succ \{x\}$.

Pattanaik's and Xu's third axiom states that adding or subtracting the same element from any two opportunity sets should not affect the freedom ranking of the two sets with respect to each other.

Axiom 3 (Independence (IND)). For all opportunity sets A and B and all $x \notin A \cup B$, $A \succcurlyeq B$ iff $A \cup \{x\} \succcurlyeq B \cup \{x\}$.

Pattanaik and Xu showed that these three axioms yield a rule, the so-called cardinality rule, according to which the freedom of choice of an opportunity set is given by the number of items in the set: the more there are, the more freedom it provides. Letting $\#A$ denote the cardinality of A, that is, the number of elements in A; this cardinality rule $\succcurlyeq_{\#}$ is defined as follows: $A \succcurlyeq B$ iff $\#A \geq \#B$.[3]

Theorem 1 (Pattanaik and Xu 1990). Let $\succcurlyeq$ be a transitive and reflexive relation over the set of all finite subsets of S. The ranking $\succcurlyeq$ satisfies (INS), (SM), and (IND) iff $\succcurlyeq = \succcurlyeq_{\#}$.

[2] For earlier axiomatic analyses of the question as to how to measure freedom of choice, see Sen (1985) and Suppes (1987). See also Kreps (1979).

[3] For extensions of the cardinal approach to a setting in which opportunity sets can be infinitely large, see Pattanaik and Xu (2000*b*), Xu (2004), and Savaglio and Vannucci (2006).

Pattanaik and Xu suggest that the result has the "flavor" of an impossibility theorem, as the cardinality rule is deeply unattractive. Is the choice between two matches in a matchbox really equivalent to the choice between a ski-ing holiday and a state-of-the-art sound system? If not, then at least one of the axioms has to give way.

Pattanaik and Xu suggest that the axiom of independence is problematic, for it fails to take account of the extent to which alternatives might differ from each other. They illustrate this with an example in which an individual's freedom to choose between different modes of transportation is compared. In their example an individual is forced to travel either by train or in a blue car. According to (INS), the two sets {train} and {blue car}, yield the same degree of freedom of choice, namely none. Now suppose we add the alternative "red car" to both sets. Independence implies that the addition of the new alternative does not affect the ranking of each opportunity set with respect to each other. Hence, {train, red car} yields equal freedom to {red car, blue car}. But surely a choice between two altogether different types of transportation (train or car) yields greater freedom than merely having a choice between two differently colored cars. In other words, independence ignores the degree of dissimilarity between various alternatives. Adding an alternative that is substantially different from those already available should provide greater freedom than adding an alternative barely distinguishable from one in the original opportunity set.

If this were all that is wrong with the cardinality rule, it could perhaps still be used when the alternatives are different enough from each other. But others suggest that the approach is misfounded from the start, by ignoring the "opportunity aspect" of freedom (Sen 1990, 1991, 1993). The idea is that freedom is not simply a choice between alternatives but is about the opportunities it provides; that is, it concerns the ability to live as one would like and to achieve things one prefers to achieve (Sen 1990, p. 471). Hence, we cannot assess the degree of freedom of individuals if we do not take account of the value of their options. In particular, since our preferences give value to freedom, we cannot derive a freedom ranking without any reference to preferences. Consider, for instance, the axioms INS and SM. Sen (1990) criticizes the axiom of INS for ignoring the fact that there is an important difference between being forced to do something that we do in fact want to do and being forced to do something that we do not want to do. According to Sen, the person who is obliged to hop home from work is less free than someone obliged to walk home, since it is obvious that anyone would prefer to walk home. The axiom of monotonicity similarly ignores the value of the options. Does adding alternatives to an opportunity set always increase freedom of choice? Does adding "being beheaded at dawn" (Sen 1991, p. 24) or "getting a terrible disease" (Puppe 1996, p. 176) to my opportunity set really add to my freedom?

15.3 Freedom and Diversity

Though Pattanaik and Xu later propose incorporating preferences into their framework, their original paper suggests that the problem with the cardinality rule occurs with the third axiom: independence (IND). Pattanaik and Xu (1990) argue that the framework should be expanded in such a way that information about the diversity of the alternatives be included, or that its use should be restricted to alternatives equally similar or close to each other. The axiom of independence has to be redefined, perhaps together with the monotonicity axiom, to arrive at a measurement of freedom which also takes into account the degree of similarity or dissimilarity between alternatives.

Now we might note here that the diversity issue might be conjoined with the opportunity issue. To say that two items in A are more alike than two items in B *is* to say that a person is more likely to be indifferent over the two items in A than over the two in B. In fact, if we truly could not distinguish between two alternatives x and y, we could hardly have a strict preference for one over the other. Furthermore, any description of the world presupposes particular criteria for establishing which entities are similar and which are not. It cannot be precluded that these criteria can be described in the same terms as the ones in which we try to capture the opportunity issue.[4] Despite this likely relationship between diversity and opportunity, the diversity issue is usually distinguished from the opportunity one, and here we follow that line.

Clearly, incorporating diversity within the framework requires some information about the (dis)similarity between the alternatives. One way is to assume that the elements of an opportunity set can be described as points in n-dimensional real space $\Re^n$. Within such a framework, Marlies Ahlert (Klemisch-Ahlert 1993) proposes to let the freedom of choice of a set of elements depend on the convex hull of that set: the larger the convex hull the more freedom of choice the set offers. Similarly, Rosenbaum (2000) takes the (normalized) maximum distance in $\Re^n$ between a pair of alternatives in an opportunity set as indicating the freedom of choice the opportunity set provides. Another proposal is to take the entropy of a set as indicative of its freedom (Suppes 1996). However, these rankings can be criticized for the fact that they take the degree of diversity within a set to be *identical* with the degree of freedom of choice offered by the set. To see why this is problematic, assume, for example, that the alternatives represent opinions that one might situate on some left–right scale and take the ranking based on maximum distance or the

[4] We might try to keep preferences out of a measurement of freedom of choice to as large an extent as possible, but the individuation of alternatives is itself a form of valuation (Dowding 1992, pp. 308–12). People value alternatives under different descriptions, and so the value of any given opportunity set to an individual depends at least in part upon the descriptions of the alternatives contained within it. See also Sugden (2003).

convex hull. A society in which one can express only the two radical views then provides equal freedom to one where all shades of opinion might be expressed. Indeed, it follows that societies in which one of the extreme opinions cannot be expressed, but all of the others can (such as in Germany, where the expression of extreme-right views is forbidden), provide less freedom than those in which *only* the extreme views can be expressed (Van Hees 2004, p. 255). Diversity may be relevant for measuring freedom of choice but the two notions should not be taken to be equivalent: having a few, but very dissimilar, elements may give less freedom of choice than having a set containing many elements even when the diversity of the latter set is less.

An alternative approach was proposed by Bavetta and Del Seta (2001). They assume that the universal set S can be partitioned into elementary subsets. The partition can be interpreted in terms of a similarity relation: elements that belong to the same equivalence class are similar. Bavetta and Del Seta then give characterizations of two rankings. One of these rankings, the *outer approximation ranking*, is especially relevant for the diversity issue. The ranking counts for each opportunity set how many equivalence classes of similar elements it contains. The higher the number, the more freedom of choice the opportunity set provides. The ranking thus forms a refinement of the cardinality ranking: the freedom of choice of a set is now given by the cardinality of the number of *dissimilar* elements in it, rather than by the total number of elements of the set. Pattanaik and Xu (2000*a*) adopt a similar approach, but they do not assume that the relation of similarity between alternatives always induces a partition. In their view, it may fail to be transitive: an alternative x may be similar to y, and y similar to z, without x also being similar to z. The ranking they propose cannot be based, therefore, on the number of equivalence classes. Instead, it is based on the cardinality of the so-called smallest similarity-based partition of the opportunity set. Their ranking coincides with the outer approximation ranking in the special case that the similarity relation is transitive, and can therefore be seen as a generalization of it.

Though these approaches do not suffer from the fact that freedom is reduced to diversity, it is a shortcoming that they do not take account of the different *degree* to which alternatives differ from each other. A blue car differs less from a red bus than a red bus does from a glass of red wine. The same problem as with the cardinality rule therefore arises: the opportunity set {blue car, red bus} gives the same degree of freedom as {blue car, red wine}. In fact, the original problem persists if a blue car is not taken to be similar to a red car. In that case the set {red car, blue car} has the same number of equivalence classes (and thus also the same number of sets in the smallest-based partition) as {train, blue car}.[5]

[5] In a recent paper, Gravel and Bervoets (2004) present an ordinal notion of similarity which enables one to take account of the degrees in which alternatives may differ from each other. However, the diversity rankings they axiomatize focus on the maximal dissimilarity within a set, and

To summarize, existing proposals to solve the diversity problem do not yet provide the definitive answer. They either suffer from the fact that freedom of choice is reduced to diversity, thereby ignoring the importance of the cardinality of a set, or they do take account of cardinality considerations—by focusing on the number of classes containing similar alternatives—but thereby ignore the degree to which alternatives differ.

The diversity issue is about finding a measurement that gets the right balance between cardinality and diversity considerations. The existing proposals overemphasize either the diversity or the cardinality aspect. Van Hees (2004) examines possible ways of incorporating information about both the degree to which alternatives are dissimilar to each other and the cardinality of a set that contains them. He shows, however, that an adaptation of the axioms of Pattanaik and Xu along those lines leads to impossibility results. Though this gives a reason for some pessimism about the prospect of solving the diversity issue, it is too early to conclude that the problem of balancing diversity and cardinality in a (non-preference-based) measurement of freedom cannot be solved. A growing recent literature on the measurement of diversity may be helpful (Weitzman 1998; Bossert *et al.* 2001, 2003; Nehring and Puppe 2002). This literature is not aimed at measuring freedom of choice, but it would be interesting to examine how these measurements of diversity might be applied to the diversity issue in the freedom of choice literature.

15.4 Freedom and Preferences

The second line we have distinguished focuses on the opportunity aspect of freedom. The idea here is that we deem freedom to be important because it provides us with the opportunity to do certain things; that is, freedom is "to live the way we would like, do the things we would choose to do, achieve the things we would prefer to achieve" (Sen 1990, p. 471). To take account of this aspect in the measurement of freedom, preferences are introduced.

One way of doing so adjusts the axioms by taking account of the *actual* preferences of the person whose freedom we are assessing. Sen (1990) rejects the axiom of INS for ignoring the difference between being forced to do something that we want to do and being forced to do something that we do not want to do. According to Sen, if a person walks home anyway, then being forced to do so reduces freedom less than if the person is forced to hop home on one

as a measure of freedom therefore suffer from the same problems as the rankings of Ahlert and of Rosenbaum.

leg. Similarly, the axiom of monotonicity is adapted: obtaining a second but very unattractive option ("being beheaded at dawn") does not increase one's freedom.

As Sen (1993, p. 529) observes, however, measuring freedom in terms of a person's actual preferences ignores the importance of our future preferences, preferences about which we are often uncertain. Since these preferences also determine the opportunity aspect of one's freedom, one should make one's ranking of opportunity sets dependent on these future preferences as well. Indeed, opportunity is not only about the things that we can do now, but also about what we can do in the future. We often cherish our freedom to choose not because of what it enables us to do now, but rather because of the flexibility it provides in minimizing the risk that one's future preferences will not be satisfiable (Kreps 1979).[6]

So we might reformulate the axioms in terms of multiple preferences, where these preferences stand for the preferences that the person could have in the future (Arrow 1995). We can also take a more sociological line, and say that the set of multiple preferences consists of the actual preference orderings of all those persons who live in the same circumstances as the person whose freedom we are assessing (Sugden 1998). Pattanaik and Xu (1998) also defend the use of multiple preferences to the measurement of freedom, but following Jones and Sugden (1982), the preferences that they take to be relevant are not potential preferences but "reasonable" ones. They give the example of a woman who is absolutely convinced that she will never want to join the army. Even without such preferences, it seems strange to say that a ban on women joining the army does not reduce the woman's freedom (Pattanaik and Xu 1998, p. 177). The reason is that "given the woman's situation, she could have reasonably chosen to join the army . . . even though she actually does not do so and even though she attaches zero probability to her wanting to do so" (Pattanaik and Xu 1998, p. 179). Pattanaik and Xu claim that freedom has an intrinsic value, and that this value lies not in having options that one might prefer, but rather in having choices between meaningful alternatives. Thus, not every addition to an opportunity set should count as an increase in freedom of choice; nor should every reduction count as a decrease. What matters are the preferences a *reasonable* person might have.

Given suitable assumptions about an individual's set of potential (reasonable) preferences $\mathcal{R}$, the three axioms can be reformulated in different ways, and thus yield different rankings. To give an indication of how potential or reasonable

[6] A formal argument for a shift to a multiple preference approach is provided by Puppe (1995), who examines freedom rankings of opportunity sets that form finite subsets of $\Re^n$ ($n \geq 2$). Puppe shows that if such a ranking is preference-based, continuous, and satisfies a "Minimal Preference for Freedom" condition, which states that for any non-singleton set A there is subset B of A such that A provides strictly more freedom than B, then the freedom ranking *cannot* be based on a single preference ordering over the basic alternatives. See also Nehring and Puppe (1996).

preferences can affect the axioms, consider the following adaptation of the axiom of strict monotonicity (where $\mathcal{R}$ denotes either the set of potential preference orderings of the agent in question or the set of preferences that are reasonable given the agent's situation):

Axiom 4. For all $A \subseteq S$ and all $x \in S - A$, if for some $R \in \mathcal{R}$ and all $y \in A$, xRy, then $A \cup \{x\} \succ A$.

Given a reference set $\mathcal{R}$ of potential (reasonable) preference orderings, define for any A the set of maximal elements of A, $\max(A)$, as the set of elements in A that are considered to be at least as good as any other element in A according to some potential (reasonable) preference ordering R: $\max(A) = \{x \mid$ there is an $R \in \mathcal{R}$ such that for all $y \in A, xRy\}$. Pattanaik and Xu (1998) characterize the ranking $\succcurlyeq_{\max}$ according to which a set A offers at least as much freedom as B if, and only if, the number of maximal elements in A is at least as large as that in B:

$$A \succcurlyeq_{\max} B \quad \text{iff} \quad \#\max(A) \geq \#\max(B).$$

If we take preferences to be relevant, however, then it may be thought to be somewhat counter-intuitive to focus only on the number of maximal elements of a set. After all, it may be the case that two sets have the same number of maximal elements even though the maximal elements of the first set are clearly superior (in terms of the potential or reasonable preferences) to those of the second. Pattanaik and Xu therefore characterize a second ranking. It is the ranking that results if one first eliminates from the sets A and B any alternative x such that for all $R \in \mathcal{R}$, x is either not maximal in $A \cup \{x\}$ or not maximal in $B \cup \{x\}$, and if one then compares the number of remaining maximal elements in A and B, respectively. A ranking that also avoids the problems of $\succcurlyeq_{\max}$ is the following one (cf. Puppe and Xu 2007):

$$A \succcurlyeq_{u} B \quad \text{iff} \quad \#(A \cap \max(A \cup B)) \geq \#(B \cap \max(A \cup B)).$$

There are by now numerous approaches to the measurement of freedom of choice through the use of preferences, and we cannot discuss them all here.[7] One line of research that we would like to point out, though, is that developed by Puppe (1996, 1998; see also Nehring and Puppe 1999, and Xu 2003) and which further develops the approach of Kreps (1979). This line in the literature can be seen as reversing the order of inquiry. The contributions that we have discussed thus far start with axioms that refer to a given reference set of preferences, and then examine which freedom rankings are induced by them. Here it is examined which kind of

[7] See, among others, Arlegi and Nieto (2001); Bavetta and Gualla (2003); Bavetta and Peragine (2006); Gravel (1998); Pattanaik and Xu (2000*b*), and Romero-Medina (2001).

freedom rankings can be represented as rankings that are based on ("revealed by") a reference set of preferences.

15.5 Unfreedoms

Should freedom rankings satisfy the axiom of *strong monotonicity*, which states that an increase of options always entails a strict increase of one's freedom? Clearly, the cardinality rule has that property. A preference-based approach entails only *weak monotonicity*: an increase of one's options will not lead to a decrease of one's freedom. In the philosophical literature there is, however, a strong tradition deriving from the account of negative liberty (MacCallum 1967; Berlin 1969), according to which adding alternatives to an opportunity set need not only fail to increase individual freedom but may accompany a *decrease* in freedom (Steiner 1974–5, 1994; Carter 1999; Kramer 2003). Freedom here is not concerned exclusively with what a person could do, but also with the things the person cannot do: to wit, those things unavailable due to constraints imposed by others. In such negative accounts of liberty, unfreedoms are at least as important as freedoms in the overall measurement. If the degree of one's overall freedom is taken to be a function of both one's freedoms and one's unfreedoms, then one might obtain extra freedoms and yet experience a decrease in overall freedom because of accompanied increases in unfreedoms.

Thus far it has been assumed that the set of all feasible alternatives, S, remains constant. Now suppose that we adopt a dynamic view, and that the set of feasible alternatives might change; for example, technological innovation might yield new feasible alternatives. Suppose, for instance, that a new medicine cures a large number of diseases for which no cure existed hitherto. Say there are n distinct diseases which can be cured, and assume that the set of relevant alternatives has thus expanded by n alternatives (each of the n alternatives describing the action of curing one of those diseases). By law, however, the medicine may be used to treat only one particular disease; moreover, this happens to be one of the less serious ones. Although the range of possible choices has expanded, one could plausibly argue that a person's overall freedom has decreased because of the concomitant prohibition against using the medicine to cure one of the $n - 1$ other, more serious diseases.

Technological innovations are not the only examples that suggest we might miss information if we focus only on those things that a person is free to do and ignore unfreedoms. Consider again the freedom to choose modes of transportation, and two situations, where a person can go to work either by car or by bus. The situations

differ only with respect to the possibility of taking a train. In the first, there is an operating railway system, but the person is not allowed to travel by train because of his skin color. In the second situation there is no railway system—say, because developing it in that remote area would be prohibitively expensive—but if there were one, he would be allowed to take it. In both situations the same options are available to the person, yet we may feel reluctant to claim that he enjoys the same degree of freedom. Note, furthermore, that such reluctance is compatible with both a preference-based and a non-preference-based view on the measurement of freedom.

One way of incorporating information about unfreedoms in the measurement of freedom is to take the ratio between the (weighted) number of things one is free to do and the (weighted) number of things one is unfree to do. Such an approach has been defended by several philosophers (Steiner 1983, 1994; Carter 1999; Kramer 2003; Van Hees 1998 offers an axiomatization of two such rankings), though their specific proposals are somewhat arbitrary and suffer logical problems (see Van Hees 2000, pp. 126–35, for a detailed critique of Carter's measure, for example). These proposals are non-preference-based; but, as indicated above, the need to take account of unfreedoms can also be said to apply to a preference-based view on the measurement of freedom.

We should note here, however, that it can be argued that a focus on unfreedoms means a shift from the analysis of "freedom of choice" to that of "freedom *simpliciter*". In an important article, Carter (2004) argues that we should keep separate our accounts of "freedom" and "freedom of choice". It may well be reasonable to suggest, for example, that (weak or strong) monotonicity is not an important criterion for the measurement of "freedom", but is it so reasonable to claim that it is not important for the measurement of the range of one's "freedom of choice"? Stated differently, unfreedoms may be important for the measurement of "freedom" but need not be so for "freedom of choice". How convincing such an argument is depends, of course, on how one understands the concept of "freedom of choice". If one emphasizes the "choice" aspect of it, and interprets freedom of choice as the range or extent of one's choices, then unfreedoms do indeed seem not to be relevant. After all, the example discussed above may show that one's freedom may be reduced even though one obtains extra options, but it seems strange to say that it shows that the range of a person's choice has decreased as well. However, if the emphasis is put on establishing *freedom* of choice, then the argument for incorporating information about unfreedoms does apply. Moreover, as we made clear at the beginning, the interest in the measurement of freedom of choice originates from a general interest in the relation between freedom, on the one hand, and notions such as well-being and justice, on the other; it does not arise from a particular interest in the semantics of the notion of freedom of choice. Clearly, given such a motivation, information about unfreedoms *is* important, since they affect our well-being and are relevant to theories of social justice.

15.6 The Extent of Freedom versus the Value of Freedom

In both the philosophical and the formal literature there has been extensive debate about whether or not a measurement of freedom should be based on information about preferences. The controversy between proponents and opponents of a preference-based measurement can be understood in two different ways. The first way is the one that we have implicitly adopted thus far and is based on a distinction between *particular freedoms* (the various particular things that a person is free to do) and a person's *overall freedom* (the set of all particular freedoms that a person enjoys). In this perspective, both approaches can be interpreted as aiming at measuring the extent of a person's overall freedom; but they are based on different assumptions about what it means to say that a person has a particular freedom.

Just as we can distinguish a preference-neutral from a preference-based approach in the context of the measurement of *overall* freedom, so we can make such a distinction with respect to the definition of *particular* freedoms. A preference-neutral definition of particular freedoms, say the freedom to do x, does not refer to the (actual, potential, reasonable) preferences concerning x, whereas in a preference-based definition such a reference is present. A cardinal approach to measuring overall freedom makes sense if we adopt such a preference-neutral definition of particular freedom. With a preference-neutral definition, *any* element of an opportunity set can be seen to represent a particular freedom. Getting an extra option, then, always entails that one acquires an extra particular freedom, and, ignoring unfreedoms and the issue of diversity, one can thus be said to have enlarged one's overall freedom. On the other hand, if we say that one can only be said to be free to do some x if x is valuable (where the value of the options is determined by actual, potential, or reasonable preferences), then an element of one's opportunity set may fail to constitute a particular freedom. A cardinal approach then fails: being able to choose an extra alternative means only that one has acquired an extra option. Since the extra option need not form a particular freedom—it may be very unattractive—the overall freedom of a person may remain unaffected. On such an account of what it means to be free to do a particular thing, a preference-based measurement does indeed make more sense.

The second way of interpreting the controversy is to say that the measurements try to capture not the extent of one's overall freedom—as we have been assuming thus far—but rather *the value* of one's overall freedom. On this view, the differences between the two types of measurements can be explained by the fact that they focus on different aspects of the value of freedom. The preference-based approaches focus on the so-called *specific value* of the freedom to choose from an opportunity set: that is, the value of being able to choose from an opportunity set insofar as that value is

reducible to the value of its elements. The cardinal approaches, on the other hand, aim at capturing the *nonspecific value* of freedom, which is the value of being able to choose from an opportunity set insofar as that value *cannot* be reduced to the value of the elements.[8]

An example of an argument for freedom's nonspecific view can be taken from John Stuart Mill (1859). Mill emphasized the causal link between exercising freedom of choice and developing one's capacity as a decision-maker. The more options an agent has to choose between, the more complex are the processes of individual decision-making. Because of the complexity of such processes, individuals further develop and cultivate their decision-making capacities. If such effects occur regardless of the nature of the options that one has to choose between, and if those effects are valuable, then they form an instance of freedom's nonspecific value. The value arises from having choices as such, rather than from the nature of the particular options that constitute that freedom of choice. Another argument for freedom's nonspecific value can be couched in terms of autonomy and responsibility. To take a very rudimentary version of the argument, to be able to choose between different courses of action creates at least some form of responsibility on the agent's side. If one refrains from choosing particular alternatives, then one bears (at least prima facie) some responsibility for not having done so. If the bearing of this kind of responsibility is taken to be valuable, then, again, freedom has a value that is not reducible to the value of the particular things that one might choose.

There are at least two difficulties with a defense of a cardinal approach to measuring freedom formulated in terms of freedom's nonspecific value. The first difficulty has to do with the description of the alternatives that goes into an opportunity set. It has to be shown that freedom's nonspecific value cannot be reduced to its specific value if the alternatives are redescribed appropriately. Suppose, for instance, that we say that having a choice between twenty mediocre but very different types of red wines yields more non-specific value than a choice between three excellent Chianti's. Furthermore, suppose that having to choose on the first occasion somehow enhances the person's wine-choosing skills—say, because the person feels the need to obtain some information about the different kinds of wine. If we describe the first twenty choices, however, not as "wine x" but as "wine x + enhancement of wine-choosing skills", then the value of the opportunity set seems to be reducible to the value of its elements.

Yet, irrespective of the feasibility of such a reduction of freedom's nonspecific value to its specific value, there remains a fundamental problem with any defense of a cardinal approach in terms of freedom's nonspecific value. After all, it is not at all clear why a cardinal measurement should be taken to capture adequately the

[8] The distinction between freedom's nonspecific and specific value originates in Carter (1999). Its application here to opportunity sets follows Van Hees (2000).

nonspecific value of a set. In fact, there are arguments for believing that this is not generally true. Consider the argument in terms of developing one's decision-making skills. To assess such an argument, we should also focus on information costs and psychological costs. Opportunity sets may be so large that they confuse people, and thereby have no positive affect on people's decision skills. This is underscored by empirical evidence. When faced with large opportunity sets, people often shy away from making choices at all in situations where they are happy to do so when faced with smaller sets (see Schwartz 2004 for a brief description of some of the evidence).

Thus, if we define *particular* freedoms without reference to preferences, cardinal measurements are more convincing if they are taken to establish *the extent* of one's freedom of choice than if one claims that they aim to measure the (*nonspecific*) *value* of one's freedom of choice. Conversely, given a preference-neutral definition of particular freedoms, a preference-based ranking makes sense as a measurement of the (specific) value of one's freedom but less so as a measurement of the extent of one's freedom. Indeed, there seems to be a clear link between preferences and the specific value of a person's freedom of choice. If a person strictly prefers alternative x to all other available alternatives, then any opportunity set containing x can be said to have greater instrumental value for him than those not containing x; they bring greater indirect utility. Or, if one adopts a wider perspective, any opportunity set containing alternatives which he finds potentially attractive (that is, there is a positive probability that he might prefer them to all other alternatives at some future point in time) gives him greater expected indirect utility than opportunity sets not containing such alternatives. Furthermore, not only do preferences help to bring out this instrumental value of freedom, but—as we saw—part of freedom's intrinsic value may be captured by using reasonable rather than potential preferences.

However, information costs and psychological costs may also affect the specific value of freedom. In order to be able to choose rather than merely "pick" (Ullmann-Margalit and Morgenbesser 1977), one must be able to distinguish the alternatives. Imagine someone facing an enormous opportunity set with a restricted time period for choosing one of the alternatives. How could one choose one's most preferred alternative if one does not have time to find out which alternative that is? All one might be able to do in such a situation is randomly to divide the opportunity set into subsets and then consider the alternatives in one of the subsets. Thus, adding alternatives to a very large opportunity set need not add any value, and indeed might subtract such value, since subdivision through picking may remove the best options. Another reason for not valuing an enlargement of one's opportunity set is that the added alternatives may form a *temptation*. That is, the agent may not want to choose those alternatives, but may nevertheless be tempted to do so (Gul and Pesendorfer 2001). The resistance of such temptation may require costly self-control, which reduces the value of the set.

To conclude, if we are interested in ascertaining *the extent* of a person's freedom, then the debate about whether to adopt a preference-based or a preference-neutral measurement is essentially a debate about the nature of particular freedoms: any choice option—attractive or not—is thought to constitute a particular freedom (yielding a cardinal measurement), or a person can be said to be free only to do valuable things (yielding a preference-based measurement). If we take the existing rankings as trying to capture *the value* of freedom, then both approaches focus on different aspects of that value, and thus both can be said to be of importance. However, they also both have their shortcomings. There are costs attached to having more freedom, and any attempt to measure the value of freedom—whether preference-dependent or not—that ignores those considerations yields at best a partial answer to the question of how valuable one's freedom is.

15.7 Some Open Questions

The axiomatic-deductive approach has yielded some important insights about freedom, but its rather abstract setting distances it from standard debates in moral and political philosophy. In this final section we discuss some avenues for future research. Several relevant open questions have already been alluded to: the formulation of a freedom ranking that takes account of diversity issues without thereby reducing freedom to diversity, the possible dependence of the diversity issue and the opportunity issue, and the need to take account of the costs of freedom.

One limitation of the abstract framework is that opportunity sets are given exogenously, and that hardly any reference is made to the institutional setting within which individuals make their choices.[9] An example of the institutional aspect of freedom that has been largely ignored concerns the relation between rights and freedom. Furthermore, as we already mentioned in our discussion of the relevance of unfreedoms, the constraints that others impose on our choices are relevant for assessing our freedom. Strategic considerations, as modeled within game theory, then become important. To distinguish an analysis that takes account of the strategic dimension of our choices from the approach discussed thus far, we may call the notion of freedom that it focuses on *freedom of action* rather than freedom of choice.

[9] An important exception is, of course, formed by the capability approach as developed by Amartya Sen (1985). If we define an opportunity set as a capability set, then a person's opportunity set is explained partly in terms of institutional characteristics (e.g. budget constraints, available commodities). See also Pattanaik and Xu (2000*b*) and Xu (2003, 2004), who present rankings of opportunity sets in a specifically economic context.

In strategic contexts especially, the importance of strategies as opposed to outcomes becomes apparent. Are the important freedoms composed of the strategies or actions we adopt to reach our preferred outcomes, or do the outcomes themselves constitute the subject of our freedom? Gravel *et al.* (1998) focus on the ranking of a person's opportuniy set in a framework in which there are more individuals and in which there are dependencies between the opportunity sets of these individuals. In particular, scarcity considerations entail that goods can belong to the opportunity sets of some but not all individuals in society. Van Hees (1999, 2000) and Vannucci (2003) use effectivity functions to analyze freedom in an interactive setting. To be effective for a set of outcomes means that the agent can insure that the outcome of the game belongs to the set. A person's opportunity set is then taken to be a family of sets, i.e. the sets for which the individual is effective. In Van Hees's approach, the sets are induced by a legal framework, thus establishing a close relation between rights and freedom. Vannucci interprets the opportunity sets as describing a person's *positive freedom* and characterizes a cardinal (partial) ordering of them. Whereas these approaches abstract from the game form underlying the effectivity functions, Bervoets (2007) discusses rankings of *game forms* in terms of the degree of freedom they offer.

An interesting aspect of the shift of the analysis to a strategic context is that the distinction between preference-based and non-preference-based measurements becomes more difficult to sustain, especially if we focus upon games rather than game forms or effectivity functions. If we model freedom of action in terms of games, then preferences determine the scope of one's freedom as one agent's actions might depend upon her knowledge of another agent's preferences. Furthermore, if my freedom depends on the course of action that you are likely to take, and if you take account of my preferences in deciding what to do, then *my* freedom will be partly dependent on *my* preferences (Dowding and Van Hees 2007).

Another issue that has received little attention in the formal literature thus far is the relation between individual freedom and group freedom. In political discussions we are often interested in the extent to which different groups of individuals can be said to have freedom, rather than the specific freedom of individual agents themselves. Can we derive rankings of *collective* freedom, and if so, how are they related to the measurements of *individual* freedom? Can a judgment about the degree of collective freedom enjoyed by a group be derivable from information about the degree of freedom enjoyed by the members of the group (Van Hees 2000)? Or is such a reduction not feasible for at least some types of collective agents (Hindriks 2008)? In addressing these questions, it would be illuminating to examine the differences and similarities between the measurement of freedom and the measurement of power. Within cooperative game theory there is an extensive formal literature on the measurement of power. Given the close relationship between the concepts of freedom and power, it would be interesting

to explore the extent to which the two types of analysis can profit from each other, and in particular insure that they are not measuring the same quantity.

References

Arlegi, R., and Nieto, J. (2001). Ranking Opportunity Sets: An Approach Based on the Preference for Flexibility. *Social Choice and Welfare*, 18, 23–36

Arrow, K. J. (1995). A Note on Freedom and Flexibility. In K. Basu, P. K. Pattanaik, and K. Suzumura (eds.), *Choice, Welfare, and Development: A Festschrift in Honor of Amartya K. Sen*, 7–16. Oxford: Clarendon Press.

Bavetta, S., and Del Seta, M. (2001). Constraints and the Measurement of Freedom of Choice. *Theory and Decision*, 50, 213–38.

—— and Gualla, F. (2003). Autonomy-Freedom and Deliberation. *Journal of Theoretical Politics*, 15, 424–43.

—— and Peragine, V. (2006). Measuring Autonomy Freedom. *Social Choice and Welfare*, 26, 31–45.

Berlin, I. (1969). *Four Essays on Liberty*. Oxford: Oxford University Press.

Bervoets, S. (2007). Freedom of Choice in a Social Context: Comparing Game Forms. *Social Choice and Welfare*, 29/2, 295–315.

Bossert, W., Pattanaik, P. K., and Xu, Y. (2001). The Measurement of Diversity. Université de Montréal, Cahier 17.

—— —— —— (2003). Similarity of Options and the Measurement of Diversity. *Journal of Theoretical Politics*, 15/4, 405–22.

Carter, I. (1999). *A Measure of Freedom*. Oxford: Oxford University Press.

—— (2004). Choice, Freedom, and Freedom of Choice. *Social Choice and Welfare*, 22, 61–81.

Dowding, K. (1992). Choice: Its Increase and Its Value. *British Journal of Political Science*, 22, 301–14.

—— and Van Hees, M. (2007). Counterfactual Success and Negative Freedom. *Economics and Philosophy*, 23, 141–62.

Gravel, N. (1998). Ranking Opportunity Sets on the Basis of their Freedom of Choice and their Ability to Satisfy Preferences: A Difficulty. *Social Choice and Welfare*, 15, 371–82.

—— and Bervoets, S. (2004). Appraising Diversity with an Ordinal Notion of Similarity: An Axiomatic Approach. NRM—Nota di Lavoro, 45.

—— Laslier, J.-F., and Trannoy, A. (1998). Individual Freedom of Choice in a Social Setting. In J. F. Laslier, M. Fleurbaey, N. Gravel, and A. Trannoy (eds.), *Freedom in Economics: New Perspectives in Normative Analysis*, 76–92. London: Routledge.

Gul, F., and Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, 69, 1403–35.

Hindriks, F. (2008). The Freedom of Collective Agents. *Journal of Political Philosophy*, 16, 165–83.

Jones, P., and Sugden, R. (1982). Evaluating Choice. *International Review of Law and Economics*, 2, 47–65.

Klemisch-Ahlert, M. (1993). Freedom of Choice. A Comparison of Different Rankings of Opportunity Sets. *Social Choice and Welfare*, 10, 189–207.

Kramer, M. H. (2003). *The Quality of Freedom*. Oxford: Oxford University Press.
Kreps, D. M. (1979). A Representation Theorem for "Preference for Flexibility". *Econometrica*, 47, 565–77.
MacCallum, G. C. (1967). Negative and Positive Freedom. *The Philosophical Review*, 76, 312–34.
Mill, J. S. (1859). *On Liberty*. Repr. London: Penguin, 1985.
Nehring, K., and Puppe, C. (1996). Continuous Extensions of an Order on a Set to the Power Set. *Journal of Economic Theory*, 68, 456–76.
——— (1999). On the Multi-Preference Approach to Evaluating Opportunities. *Social Choice and Welfare*, 16, 41–63.
——— (2002). A Theory of Diversity. *Econometrica*, 70, 1155–90.
Pattanaik, P. K., and Xu, Y. (1990). On Ranking Opportunity Sets in Terms of Freedom of Choice. *Recherches Economiques de Louvain*, 56, 383–90.
——— (1998). On Preference and Freedom. *Theory and Decision*, 44, 173–98.
——— (2000*a*). On Diversity and Freedom of Choice. *Mathematical Social Sciences*, 40, 123–30.
——— (2000*b*). On Ranking Opportunity Sets in Economic Environments. *Journal of Economic Theory*, 93, 48–71.
Peragine, V., and Romero-Medina, A. (2006). On Preference, Freedom and Diversity. *Social Choice and Welfare*, 27, 29–40.
Puppe, C. (1995). Freedom of Choice and Rational Decisions. *Social Choice and Welfare*, 12, 137–53.
——— (1996). An Axiomatic Approach to "Preference for Freedom of Choice". *Journal of Economic Theory*, 68, 174–99.
——— (1998). Individual Freedom and Social Choice. In J. F. Laslier, M. Fleurbaey, N. Gravel, and A. Trannoy (eds.), *Freedom in Economics: New Perspectives in Normative Analysis*, 49–68. London: Routledge.
——— and Xu, Y. (2007). Revealed Preference for Freedom and Ordinal Rankings of Opportunity Sets. Mimeo.
Romero-Medina, A. (2001). More On Preferences and Freedom. *Social Choice and Welfare*, 18, 179–91.
Rosenbaum, E. F. (2000). On Measuring Freedom. *Journal of Theoretical Politics*, 12, 205–77.
Savaglio, E., and Vannucci, S. (2006). On the Volume-Ranking of Opportunity Sets in Economic Environments. Mimeo.
Schwartz, B. (2004). *The Paradox of Choice: Why More is Less*. New York: HarperCollins.
Sen, A. K. (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.
——— (1990). Welfare, Freedom and Social Choice: A Reply. *Recherches Economiques de Louvain*, 56, 451–85.
——— (1991). Welfare, Preference and Freedom. *Journal of Econometrics*, 50, 15–29.
——— (1993). Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedoms. *Oxford Economic Papers*, 45, 519–41.
Steiner, H. (1974–5). Individual Liberty. *Proceedings of the Aristotelian Society*, 75, 33–50.
——— (1983). How Free: Computing Personal Liberty. In A. Phillips Griffiths (ed.), *Of Liberty*, 73–89. Cambridge: Cambridge University Press.
——— (1994). *An Essay on Rights*. Oxford: Blackwell.
Sugden, R. (1998). The Metric of Opportunity. *Economics and Philosophy*, 14, 307–37.

Sugden, R. (2003). Opportunity as a Space for Individuality: Its Value and the Impossibility of Measuring it. *Ethics*, 113, 783–809.

Suppes, P. (1987). Maximizing Freedom of Choice: An Axiomatic Analysis. In G. R. Feiwel (ed.), *Arrow and the Foundations of the Theory of Economic Policy*, 243–54. Basingstoke: Macmillan.

—— (1996). The Nature and Measurement of Freedom. *Social Choice and Welfare*, 13, 183–200.

Ullmann-Margalit, E., and Morgenbesser, S. (1977). Picking and Choosing. *Social Research*, 44, 757–85.

Van Hees, M. (1998). On the Analysis of Negative Freedom. *Theory and Decision*, 45, 1175–97.

—— (1999). Liberalism, Efficiency and Stability: Some Possibility Results. *Journal of Economic Theory*, 88, 294–309.

—— (2000). *Legal Reductionism and Freedom*. Dordrecht: Kluwer Academic.

—— (2004). Freedom of Choice and Diversity of Options: Some Difficulties. *Social Choice and Welfare*, 22, 253–66.

Vannucci, S. (2003). The Cardinality-Based Ranking of Opportunity Sets in an Interactive Setting. Università degli Studi di Siena, Dipartimento di Economia Politica, Working Paper.

Weitzman, M. L. (1998). The Noah's Ark Problem. *Econometrica*, 66, 1279–98.

Xu, Y. (2003). On Ranking Compact and Comprehensive Opportunity Sets. *Mathematical Social Sciences*, 45, 109–19.

—— (2004). On Ranking Linear Budget Sets in Terms of Freedom of Choice. *Social Choice and Welfare*, 22, 281–9.

CHAPTER 16

RESPONSIBILITY

MARC FLEURBAEY

16.1 Introduction

Imagine a simple world in which individual well-being depends on only two things: money and cheerfulness. Money is received from two sources. At the beginning of their lives, individuals receive a personal bequest and at the same time are submitted to a state tax and transfer. The combination of these two operations determines their disposable wealth, and they then pursue their life plans and obtain a certain level of well-being. For simplicity we suppose that well-being is proportional to disposable wealth and to a certain index of cheerfulness:

$$\text{well-being} = (\text{bequest} \pm \text{transfer}) \times \text{cheerfulness}$$

Imagine that there are four categories of people, an individual falling into one category or another depending on whether his bequest is low or high, and whether his cheerfulness is low or high. In the absence of transfers, the situation is assumed to be as in Table 16.1. The figures in the table cells measure well-being.

Assume that, for whatever reason, we consider that inequalities due to bequests are unjust, but that, in contrast, individuals should be held responsible for their cheerfulness. In other words, individuals are considered partly responsible for their well-being. How can this partial responsibility be translated into appropriate transfer policies?

This chapter has benefited from comments by the editors and a referee.

Table 16.1. ***Laissez-faire*** **situation**

	Low cheerfulness (= 1)	High cheerfulness (= 3)
Low bequest (= 1)	1	3
High bequest (= 3)	3	9

Table 16.2. Natural policy

	Low cheerfulness (= 1)	High cheerfulness (= 3)
Low bequest (= 1)	2 (transfer = +1)	6 (transfer = +1)
High bequest (= 3)	2 (transfer = −1)	6 (transfer = −1)

Table 16.3. Utilitarian policy

	Low cheerfulness (= 1)	High cheerfulness (= 3)
Low bequest (= 1)	0 (transfer = −1)	12 (transfer = +3)
High bequest (= 3)	0 (transfer = −3)	12 (transfer = +1)

In this particular example, there is a "natural policy" that presents itself immediately as an appealing option. It consists in equalizing the wealth that is at the disposal of individuals after transfer. This policy is illustrated in Table 16.2, for the case when the four categories of individuals come in equal numbers.

Over the entire population, this policy equalizes exactly what individuals are not responsible for: namely, their wealth (made of bequest and transfers). In the table every individual has a disposable wealth equal to 2.

One may observe, however, that the Natural policy is not the only one that neutralizes the inequalities due to differential bequest. Here is another example of a policy which performs the same neutralizing operation in a different way (see Table 16.3). This policy is dubbed "utilitarian" because it maximizes total utility with the available resources while compensating bequest inequalities.

These examples show that a restrictive reading of the idea that inequalities for which individuals are not responsible should be suppressed leaves it quite indeterminate what precise redistribution should be made. Compensation for unequal circumstances (like bequests in this example) cannot be the only goal of social policy; it must be supplemented by a reward principle telling us whether and how redistribution should be sensitive to responsibility characteristics as well.

16.2 The Framework

A general framework for the study of responsibility-sensitive redistribution policies can be conceived as an extension of the above example.[1] Suppose that there is a relevant measure of individual outcome, e.g. utility, income, success, for which individuals are partly responsible. Their partial responsibility is embodied in certain responsibility characteristics, e.g. preferences and effort, which have an impact on their outcome. In addition, their outcome also depends on factors for which they are not held responsible, e.g. social background and inherited traits and resources. These are called "circumstances". Finally, transfers and various kinds of public intervention can correct inequalities. We then write individual outcome as a function of these three kinds of factors:

$$\text{Individual outcome} = f\ (\text{transfers, circumstances, responsibility characteristics}).$$

Among the circumstances, one can record social interactions and influences of others' actions. This formula does not require assuming that individuals live in isolation.

This framework is convenient for the analysis of redistributive policies inspired by a variety of ethical approaches. In particular, Rawls (1971) and Dworkin (2000) have proposed to equalize a comprehensive notion of resources across individuals, arguing that individuals should be held responsible for their preferences and goals in life, because autonomous moral agents should assume responsibility for their personal conception of the good life. In this perspective, circumstances are the initial resources available to individuals (including personal parameters like talent), and responsibility characteristics correspond to preferences and utility functions. A related but different approach has been advocated by Arneson (1989), Cohen (1989), Roemer (1998), and Sen (1985). It consists in seeking to equalize opportunities of a certain outcome (subjective welfare for Arneson, a more objective measure of advantage for Cohen and Sen). In this case, responsibility characteristics coincide with the dispositions which govern the choice made by the individual in the set of options which is formed as a combined result of circumstances and transfers. Other approaches can be accommodated. For instance, if one simply wants income not to depend on social background, as in studies of social mobility, one can readily apply this framework, with income as the individual outcome, social background as the circumstances, and other characteristics in the responsibility sphere.

[1] Other surveys relying on this framework can be found in Peragine (1999), Fleurbaey and Maniquet (2009).

16.3 The Liberal Approach

We have seen in Section 16.1 that many policies can neutralize inequalities in circumstances. They all differ by how they apportion the final outcome to responsibility characteristics. There is a distinctive liberal approach to this problem, which is exemplified by Rawls's and Dworkin's theories. It consists in seeking to equalize the (transfers, circumstances) compound across individuals, with the idea that there is no reason to make this compound vary as a function of responsibility characteristics. In their approach, having such-or-such goal in life is not a reason for obtaining a different amount of resources. In more general terms, it would be illiberal to reward or penalize certain responsibility characteristics. This corresponds to a conception of responsibility as a private sphere in which public institutions should not interfere.

The idea of equalizing the (transfers, circumstances) compound raises a difficulty, since this is a multidimensional quantity. One needs a synthetic index or something of that sort in order to compare the value of the compound across individuals. Sometimes this index is already incorporated in the function f, when, as in the example of Section 16.1, the function is separable in the compound. The Natural policy of Section 16.1 is precisely the policy that equalizes (bequest + transfer) across individuals; so it exemplifies the liberal approach in that case. But when this separability does not hold, it is more ambiguous how to equalize the compound.

Rawls has not given precise indications about how to build this index, suggesting only that it should reflect representative preferences of the worst-off group. Dworkin has invoked the envy test—discussed below—before turning to another idea, the hypothetical insurance. The latter consists in imagining the possibility of individuals taking out an insurance against bad circumstances, before knowing their actual circumstances. Although prima facie appealing, this idea has a basic drawback. When individuals think that they have a chance of obtaining good circumstances, they are sometimes willing not to take out an insurance if their marginal utility is greater in the good state. This seems a poor justification for not making transfers in favor of individuals who actually have bad circumstances (and never enjoyed the real prospect of getting good ones). Moreover, such a scheme dramatically departs from the initial idea of equalizing the (transfers, circumstances) compound, since it may involve not making compensatory transfers for unequal circumstances, or even making transfers in the wrong direction![2]

The idea of relying on the envy test is more interesting, since this test has in effect been conceived as a way to obtain a certain equality of multidimensional bundles of

[2] See Roemer (1985, 2002*a*).

resources.[3] The envy test is satisfied when no individual would rather have another's bundle (otherwise he is said to envy the other). Translated into our framework, this means that the situation should be such that no individual would be better off with his own responsibility characteristics and another's (transfers, circumstances) compound.

Apply this test to two individuals with identical responsibility characteristics. This means that neither of them should be better off with their common responsibility characteristics and the other's (transfers, circumstances) compound. In short, they should have equal outcome. This nicely embodies the *compensation principle*, according to which circumstances should not create inequalities: when every pair of individuals who differ only in their circumstances have the same outcome, this is clearly achieved.

Now apply the test to two individuals with identical circumstances. This implies that neither of them should be better off with his own characteristics (responsibility and circumstances) and the other's transfer. In short, they should have equivalent transfers. This embodies the *liberal reward principle*, that responsibility characteristics in themselves justify no transfer: individuals who differ only in their responsibility characteristics should not suffer discriminatory transfers.[4]

16.4 Compensation versus Reward

The envy test is a nice method in principle, but it often fails in practice to deliver precise recommendations, because there are many situations in which no transfer policy can eliminate envy. When, in a pair of individuals, each of them thinks that his circumstances are worse than the other's, any transfer between them will increase the envy of one of them. Even when everyone agrees on how to rank circumstances, if the worst-off in circumstances also thinks this to be a greater handicap than the other estimates, he will require a greater transfer than allowed by the other, and necessarily one of them will be envious.

[3] The seminal works on no-envy are Kolm (1972) and Varian (1974). This equity concept is often criticized, by welfarist authors, for failing to take account of people's unequal abilities to enjoy the resources. The approach presented here eliminates this objection, because such abilities can be put in the circumstance parameters over which the envy test is performed.

[4] Early formulations of these principles appeared in Dworkin (1981) and Barry (1991). There are other formulations of requirements inspired by these general principles of compensation and reward. It is essential to distinguish the general idea of a principle from the multiplicity of precise requirements which can embody it.

Table 16.4. Example

	Productivity ($/hour)	WTP for leisure ($/hour)
Ann	2	5
Barb	20	5
Chris	2	15
Deb	20	15

This difficulty originates in a conflict between the compensation principle and the liberal reward principle. I will illustrate this with a new kind of example.[5] Consider a population with heterogeneous productivities and heterogeneous preferences regarding leisure, such that individuals are held responsible for their preferences but not for their productivity. Productivity is measured in dollars per hour, and preferences are described in terms of willingness to pay (WTP) for one extra hour of leisure. We assume linear preferences, which means that this WTP, for a given individual, is the same no matter how much he works and consumes. The population is again assumed to have four equal-sized groups, and for simplicity's sake, let us imagine that each group has only one individual (see Table 16.4).

Pareto efficiency requires Barb and Deb to work full-time (eight hours per day, say), because their productivity exceeds the compensation they request in order to work an additional hour. It also requires that Ann and Chris should not work, unless they are forced to work for the benefit of someone else. Let us focus first on situations in which Ann and Chris do not work. The *compensation principle* imposes that Ann and Barb should have equivalent situations according to their common preferences, as well as Chris and Deb. In view of the fact that in each of these pairs only one of them works, this means that the consumption gap between the working and the idle must equal $5 \times 8 = 40$ dollars for Ann and Barb, and $15 \times 8 = 120$ dollars for Chris and Deb. On the other hand, the *liberal reward principle* says that Ann and Chris, on one side, and Barb and Deb, on the other side, should have equal consumption because in each pair they have the same quantity of labor (and the same productivity). But if this is achieved, necessarily there will be the same consumption gap between Ann and Barb and between Chris and Deb, contradicting the above point that it should be $40 for the former and $120 for the latter. When Ann and Chris are made to work, it is again impossible to give them the same consumption–leisure bundle and do the same for Barb and Deb while obeying the compensation principle.

If it is impossible to satisfy the compensation principle and the liberal reward principle fully, one can try to satisfy them to some extent. For instance, one can seek a policy rule that equalizes outcomes when *all* the individuals have identical

[5] This example is closely connected to the example provided by Pazner and Schmeidler (1974) in order to show that envy-free efficient allocations may fail to exist in a production economy.

responsibility characteristics, or that equalizes transfers when *all* the individuals have identical circumstances.

An example of the former option is the Conditional Equality rule, which picks a reference value for responsibility characteristics, and seeks to equalize the hypothetical outcomes that individuals would have if their responsibility characteristics were at the reference value. That is, it seeks to equalize

$$\text{hypothetical individual outcome} = f\ (\text{transfers, circumstances, reference responsibility characteristics}).$$

This is a very natural formula, since individuals are guaranteed a certain outcome on condition that they have the reference responsibility characteristics. Any deviation from this reference is ignored, and individuals bear its consequences fully. This is well in line with the liberal reward principle, but not so well in line with the compensation principle, since only individuals with the reference responsibility characteristics have their differential circumstances fully neutralized. However, outcomes will be equalized when *all* the individuals have identical responsibility characteristics and the reference value coincides with these common characteristics.

An example of the latter option is the Egalitarian-Equivalent rule, which picks a reference value for the circumstances and seeks to obtain a distribution of outcomes that could also be obtained if all the population had the reference circumstances and received an equal transfer. That is, it seeks to obtain the following equality for each individual, where transfers* denotes a certain value which is the same for all individuals:

$$\text{actual outcome} = f\ (\text{transfers}^*, \text{reference circumstances, responsibility characteristics}).$$

Observe that this equality implies that individuals with identical responsibility characteristics must have the same outcome. The liberal reward principle, in contrast, is less well satisfied, since individuals who differ only in their responsibility characteristics and have circumstances different from the reference may receive substantially different transfers.

In the separable case, as in the example of Section 16.1, the Conditional Equality and Egalitarian-Equivalent rules coincide and advocate the Natural policy which equalizes the (transfers, circumstances) compound.

There are other interesting rules in this vein.[6] In particular, there are rules which are more closely inspired by the envy test, and seek to minimize the occurrence

[6] The economic literature on all of these solutions includes Bossert (1995); Bossert and Fleurbaey (1996); Bossert *et al.* (1999); Cappelen and Tungodden (2002, 2003); Fleurbaey (1994, 1995*b*, 2008); Fleurbaey and Maniquet (1996, 1999, 2005); Gaspart (1998); Iturbe-Ormaetxe (1997); Iturbe-Ormaetxe and Nieto (1996); Kolm (2004); Maniquet (1998, 2004); Moulin (1994); Peragine (1999); Sprumont (1997); and Tungodden (2005).

or the intensity of envy in the population. One is Van Parijs's (1995) Undominated Diversity rule, which seeks a situation such that for no pair of individuals i, j, everyone in the population would rather have i's (transfers, circumstances) compound than j's. This rule is not very demanding, since it suffices to find one dissenting voice in order to consider that i's compound does not dominate j's. A more interesting rule, called Envied Intensity, computes for each individual the tax that should be levied on him so that no one else envies him any more (this tax is null if no one envies him already). Then it seeks to minimize the sum of these amounts over the whole population. This rule satisfies the compensation principle fully, and the liberal reward principle in the case of populations with identical circumstances. Moreover, it selects envy-free allocations whenever such allocations exist.

These various criteria have been applied to various contexts, including the problem of income taxation when one cannot observe individual choices of labor and effort but only gross income, so that redistribution has disincentive effects. In particular, depending on how one chooses the reference value for circumstances, the Egalitarian-Equivalent rule advocates maximizing either the minimum income (which can take the form of a negative income tax or a universal grant) or the net income of the working poor.[7]

16.5 The Utilitarian Approach

In Section 16.1 I gave the example of a "utilitarian" policy which neutralizes the impact of circumstances but also seeks to maximize total well-being. Such an approach can be justified by an understanding of the implications of responsibility which differs from the liberal conception. According to this different interpretation, responsibility means that one can be indifferent to inequalities between individuals when these individuals are responsible for them. Indifference to inequalities is not the same as liberal nonintervention. Let me explain why.

In welfare economics, there are two extreme and salient social welfare functions. One is the utilitarian criterion, which simply computes the sum of well-being levels over the whole population. It corresponds to the case in which one is indifferent about the distribution of well-being and is simply interested in the sum. The other is the maximin criterion, which focuses exclusively on the smallest value of well-being over the whole population. It expresses a strong aversion to inequalities.

[7] See Fleurbaey and Maniquet (2006, 2007) and Van der Veen (2004). Applications to health policy have been made by Schokkaert *et al.* (1998), and Schokkaert and Van de Voorde (2004).

According to the *utilitarian reward principle*, one should maximize the sum of outcomes over subgroups of individuals who differ only in their responsibility characteristics, because the fact that they are responsible for their differences justifies, on this interpretation, indifference about inequalities among them.[8] This is different from the liberal approach, which advocates submitting them to the same transfers. The utilitarian reward principle may, on the contrary, advocate transfers between them in order to increase their total outcome.

The utilitarian reward principle operates only over subgroups of people who differ only in their responsibility characteristics. In contrast, in subgroups of individuals who differ only in circumstances, the compensation principle applies, and one should rather implement the maximin criterion. Since these two kinds of subgroups overlap, it is not obvious how to proceed.

Two simple options immediately come to mind. If one first computes the average outcome in each circumstance class (i.e. subgroup with identical circumstances), one can then apply the maximin criterion to such average figures. Symmetrically, if one first computes the minimum within each responsibility class (i.e. subgroup with identical responsibility characteristics), one can then apply the utilitarian criterion over such minimum numbers. This provides two social welfare functions, respectively called the Min of Means and the Mean of Mins.

The duality of these two social welfare functions is again a token of the conflict between the compensation principle and the reward principle. Consider the following situations, for a population with four categories, as in Section 16.1 (see Tables 16.5 and 16.6).

Situation B is better than situation A according to the compensation principle, because the inequalities within each responsibility class are reduced, and the worst-off is better off. But Situation A is better than situation B according to the utilitarian reward principle, because the sum of outcomes within each circumstance class is greater.

One can check that the Min of Means criterion satisfies the utilitarian reward principle better than the compensation principle (for instance, it ranks A above B), whereas the Mean of Mins criterion is better for the compensation principle (it ranks B above A). However, the two agree when there is an unambiguously worst-circumstance class, which is dominated by the other classes for all values of responsibility characteristics. In this case, both criteria advocate maximizing the average outcome of this circumstance class.

Like the liberal approach, the utilitarian approach has been applied to a variety of policy issues, like income taxation, education policies, and international

[8] Let it be clear that the utilitarian reward principle has little to do with the classical utilitarian doctrine according to which the system of rewards and penalties must be devised so as to maximize the total sum of utilities over the whole population. In particular, it is restricted to subgroups with identical circumstances.

Table 16.5. Situation A

	Responsibility characteristics 1	Responsibility characteristics 2
Circumstances 1	2	22
Circumstances 2	10	15

Table 16.6. Situation B

	Responsibility characteristics 1	Responsibility characteristics 2
Circumstances 1	3	17
Circumstances 2	6	16

aid.[9] It has also inspired studies of social mobility which involve comparing the distributions of outcome (usually income) for the circumstance classes of individuals defined by their social origin.[10] It appears in particular that social mobility indices which are concerned with the distance of individuals from their origin fail to capture the more ethically relevant idea of equal opportunities. A society in which all the children of poor parents become rich and all the children of rich parents become poor has a lot of "mobility", but this is not the kind of mobility one should be interested in. Instead, one can compute an index of inequality of opportunity inspired by the Min of Means criterion: that is, based on inequalities between the average levels of outcomes of different circumstance classes. It is also possible to compute a similar index of inequality inspired by the Mean of Mins. In this case the compensation principle means that individuals who are at the same percentile of the distribution of outcomes in their respective circumstance classes should have the same outcome.[11]

[9] The relevant literature includes Hild and Voorhoeve (2004); Ooghe and Lauwers (2005); Ooghe *et al.* (2007); Peragine (2002, 2004); Roemer (1993, 1998, 2002*b*); Roemer *et al.* (2003); Schokkaert *et al.* (2004); Van de gaer (1993); and Vandenbroucke (2001).

[10] See in particular O'Neill *et al.* (2000); Roemer (2004); Schluter and Van de gaer (2002); and Van de gaer *et al.* (2001).

[11] Roemer (1998) proposes that this is the correct way to measure responsibility characteristics, at least if circumstance classes are identified correctly. The idea is that belonging to a certain circumstance class may affect the possibility of having various responsibility characteristics, so that one should be held responsible not for one's absolute value of responsibility characteristics, but only for one's relative position in one's circumstance class. This statistical approach is convenient, as a way of eliminating any correlation between circumstance and responsibility characteristics, but its holistic underpinnings are obscure. It implies, for instance, that one's position can change as a result of changes in the responsibility characteristics of other members of one's class, even when this has no direct impact on oneself.

16.6 Comparing the Four Approaches

The liberal/utilitarian divide and the compensation/reward divide provide us with four kinds of allocation rules and social orderings.[12] Although there are other solutions, such as Undominated Diversity and Envied Intensity, one can take Conditional Equality, Egalitarian Equivalence, Min of Means, and Mean of Mins as the salient options in this configuration (see Table 16.7).

One can defend a priority of the compensation principle over reward principles by the argument that it is more important to neutralize the effect of circumstances on personal outcomes than to respect a particular reward principle. But there are cases where such a priority is questionable or impractical. For instance, when redistribution must be performed before individual responsibility characteristics are determined, one is forced to adopt a liberal kind of reward, because the transfers are then necessarily independent of responsibility characteristics, and all individuals in any given circumstance class will be treated identically. There are also situations in which the compensation principle is not compelling. Consider a responsibility class in which individuals have very exotic preferences—for which they are responsible—about the various circumstances. Should transfers among this class obey these exotic preferences, or should they rather obey normal preferences as with Conditional Equality? This is debatable.

The liberal/utilitarian divide corresponds to a deep difference in the interpretation of the implications of responsibility. The two approaches require a different informational basis. The utilitarian approach needs to know only the distribution of outcome in the various classes of circumstances and responsibility in order to evaluate a social situation, whereas the liberal criteria require data on the transfers in addition, in order to assess inequalities of the (transfers, circumstances) compounds across individuals. On the other hand, the utilitarian criteria require an interpersonally comparable measure of outcome, in order to compute means or sums over individuals, whereas the liberal criteria need to know only the ordinal and noncomparable rankings of the (transfers, circumstances) compounds for each value of the responsibility characteristics. Take, for instance, Conditional Equality. It simply compares the (transfers, circumstances) compounds for the reference value of the responsibility characteristics. Similarly, Egalitarian Equivalence simply compares the (transfers, circumstances) and (transfers*, reference circumstances) compounds for the concerned individual's value of responsibility characteristics. A transformation of the outcome function f (it can be different for different values of

[12] An allocation rule defines a particular choice for each set of feasible allocations. A social ordering ranks all the options of each feasible set, and therefore provides a finer ordering of the options than an allocation rule. All the general concepts presented in this chapter have been developed for allocation rules as well as social orderings, in the literature.

Table 16.7. Synopsis of solutions

	Liberal reward	Utilitarian reward
Compensation	Egalitarian-Equivalent Envied Intensity, ...	Mean of Mins
Reward	Conditional Equality Undominated Diversity, ...	Min of Means

the responsibility characteristics) that would preserve its ordinal properties would not affect any of these computations.

This can be viewed as a decisive practical advantage of the liberal approach when individuals are held responsible for their utility functions, because utility levels are much are harder to observe (and even to make sense of) than ordinal preferences. Because of this, the utilitarian approach is rather counter-intuitive in many applications, because it implies that one should try to maximize the sum (or mean) of utilities in certain circumstance classes. For instance, with regard to the problem of income taxation the application of Min of Means or Mean of Mins implies that one should try to maximize the average utility of low-skilled individuals, transferring income from individuals with low marginal utility to individuals with high marginal utility. It appears more palatable, and closer to common practice, though, to consider that if marginal utility of income belongs to the responsibility sphere, we should simply ignore it in the design of taxes and transfers.

The liberal approach can also be criticized, however, for its bias toward the *laissez-faire* policies. Even if individuals are held responsible for their preferences, we do not want to ignore them completely if market failures lead to a deficit in satisfaction. It is therefore important to formulate conditions of liberal reward which are moderate enough to be compatible with the Pareto principle.

Moreover, both approaches can be questioned for their total indifference to inequalities due to responsibility characteristics. In principle these criteria are compatible with an arbitrary degree of *ex post* inequalities, including with people falling below the subsistence threshold or becoming the slaves of others. It seems strange to imagine that personal responsibility could justify abject poverty and extreme subordination. Human rights, in particular, are usually thought to be so basic that they cannot be waived, even "responsibly".[13]

[13] Criticisms of the equal-opportunity theories of justice have been voiced e.g. by Anderson (1999); Fleurbaey (1995*a*); Hurley (2003), Scheffler (2003), and Wolff (1998). Arneson (2000) proposes a revised theory that takes account of some of them.

16.7 Responsibility versus Freedom

The economic analysis of responsibility-sensitive social criteria simply explores the distributive *implications* of holding individuals partly responsible for their own fate. The difference between liberal and utilitarian reward, and the conflicts between reward principles and the compensation principle, can be presented independently of how one decides to draw the boundary between responsibility and circumstances and of how one wants to justify the role of responsibility in social evaluation. These are, however, important issues to be addressed in order to determine how relevant and appealing the whole enterprise is.

What should individuals be held responsible for? The egalitarian literature in philosophy contains two main views on the topic. One, defended by Rawls (1982) and Dworkin (2000), is that individuals should be held responsible for their preferences and goals in life. It has been attacked by Arneson (1989) and Cohen (1989) for failing to address the plight of individuals who suffer disadvantages because of preferences which they have inherited. According to these authors, individuals should be held responsible only for what they genuinely choose or control. This approach is closer to the commonsense notion of responsibility, and it directly connects responsibility to free will. In this way, social justice becomes dependent on a deep and maybe unsolvable metaphysical issue. Roemer (1998) suggested that each society can decide on its own how to define the responsibility/circumstances cut in a political decision; but he also appears to support the view that individuals should not be held responsible for what they do not control, a requirement which is unlikely to be satisfied by political settlements.

The idea that responsibility coincides with control is hard to apply in economic models, where individual decision is described as a mechanical operation of maximization under constraints. If the individual does not control his preferences and his budget set, how can he be held responsible for picking the option in the budget set which is the best according to his preferences? The individual does not appear to be more in "control" than an automaton which has a well-designed servomechanism.

In contrast, responsibility for preferences (and utility functions) is much easier to apply in economic models and is typically adopted in most of the economic literature under consideration here, liberal and utilitarian alike. It is noteworthy that Dworkin rejects the identification of responsibility and control proposed by his critics. His main argument is that individuals can request help for personal handicaps, but cannot view the goals they adopt for their life as handicaps. When they suffer from cravings which they wish they did not have, it makes sense to provide help, because such cravings do not reflect their real goals. But when one deeply identifies with one's goals, Dworkin argues, how could one ask for more resources on the pretext that this goal is a hindrance to satisfaction? This argument,

however, is not ultimately convincing, because one can identify with one's goal while observing that the prevailing conditions are unfavorable to its satisfaction. One would then ask for help not because one's goal is a handicap, but because the environment is not favorable to the satisfaction of this goal.[14]

The defence of responsibility for personal preferences, in Rawls's theory, is a topic of interpretational controversies, but interestingly some authors[15] argue that, unlike Dworkin, Rawls does not claim that responsibility for preferences has immediate ethical value. Instead, responsibility for one's preferences would follow from other fairness principles, in particular the idea that in a system of fair cooperation between autonomous moral agents, it is appropriate that they share resources equally, on the understanding that each of them will assume responsibility for his own goals and for his view of the good life. Responsibility in this perspective is not a pre-institutional notion which can justify some inequalities; it is a consequence of fair arrangements in which people assume certain roles and functions.

Rawls's description of the fairness principles which justify assigning responsibility to people for their preferences is rather cryptic, which is why many commentators view his approach as a mere precursor of Dworkin's. There is a simpler approach, which relies on the idea that freedom, understood as the ordinary activity of choice, is an essential dimension of human life. Sen (1992), for instance, argues that one cannot simply evaluate individual situations in terms of outcomes, because this neglects the freedom aspect. Now, one cannot offer people a substantial array of options if their choice does not have consequences. In other words, responsibility for one's choice must somehow follow directly from freedom of choice.

Making responsibility a notion derivative of freedom instead of a basic principle has many advantages. First, it eliminates the dependence of social justice on the metaphysics of free will, because one can recognize the importance of the ordinary activity of choice for a flourishing life even if one is a hard determinist (i.e. believes that there is no free will). Secondly, it justifies putting limits on the possible consequences of responsible choice. Indeed, freedom is not enhanced by introducing the possibility of living in misery or under the dominion of others. Freedom has value when one is offered a good menu of options, more than when the menu contains bad and destructive options. Thirdly, it replaces the backward-looking, punitive, and moralizing justification of disadvantages that is pervasive in theories of equal opportunities, with a forward-looking, enhancing, and nonmoralistic approach, which can in particular advocate providing fresh starts for those who regret past decisions, no matter how responsible they are for those decisions.

[14] See Cohen (1989) and Matravers (2002). This rebuttal of Dworkin's defense does not automatically support the control view of responsibility. One can be in control of one's goals, identify with a chosen goal, and still reasonably complain that this goal is hard to achieve in the prevailing environment. The fact that one could adopt other goals and more easily achieve them does not seem sufficient to reject such complaints. Satisfaction of *any* goal is not the ultimate goal.

[15] See esp. Scheffler (2003).

It was suggested above that freedom of choice implies responsibility for choices, but this is not quite true. It does not make sense to hold people directly responsible for their choices when they have different menus of options, as is typically the case even under strongly egalitarian policies. For instance, one cannot hold people responsible for their supply of labor when they have different wage rates, because their choice is influenced by the difference in their budget sets. One can only hold them responsible for their preferences, and it does indeed make sense, from the freedom perspective, to hold people responsible for their preferences, at least the preferences they identify with, because these preferences are the underlying principle which would govern their choices if they were in good conditions of choice. This responsibility assignment involves a respect for people's goals in life. The framework presented above is well adapted to this assignment of responsibility and has mostly been used in this way in economic applications, as already mentioned.

As explained above, Sen defends his "capabilities" approach primarily in terms of freedom rather than responsibility, but the difference from theories of equal opportunity which rely on a pre-institutional notion of responsibility is ultimately unclear. His main argument is that what is really important is not the actual level of achievement ("functionings") but the access to achievements, and this notion of access may be vulnerable to metaphysical worries. When he opposes fasting to starving, for instance, one may be worried that the fasting individual is actually under some influence that puts satisfactory nutrition out of reach for him.

What was proposed above was an alternative use of the notion of freedom, which refers to the ordinary activity of choice, the scope of which is defined by institutions. In this context one no longer asks whether the fasting individual has enough opportunities, but whether endangering one's health without any concern on behalf of social institutions is a valuable option to put on the menu.

The variant of Sen's theory which is defined in terms of "refined functionings"—namely, functionings associated with the capability sets from which they are chosen—is more interesting, because it makes it possible to record individual achievements and the way in which individuals value these achievements together with other possibilities. Sen unfortunately argues that refined functionings and capabilities are equivalent, because the chosen functionings are a part of the capability set. This ignores the obvious informational difference between saying that "Jones has access to food" (a capability information) and saying that "Jones has access to food but fasts" (a refined functioning information). The capability approach will consider that Jones is well off, even though he may be close to dying, whereas the refined functioning approach permits a more comprehensive evaluation. The freedom approach that was sketched above is still different, and will ask whether Jones has what he wants from a *good* menu, which may not be the case if the option "fasting without limit and without medical monitoring" should not be on a good menu.

In conclusion, the economic analysis of responsibility-sensitive redistribution, which typically holds individuals responsible for their preferences, can be justified either from a doctrine of equality of resources *à la* Rawls and Dworkin, or, preferably, from a freedom-oriented approach which considers that it is important to grant substantial freedom of choice to individuals and to respect their preferences. The concern for a "good menu" can be introduced into the analysis by imposing subsistence requirements and bans on certain kinds of subordination. Under such safeguards, the application of the criteria presented above seems a reasonable way to articulate fairness, freedom, and responsibility.

References

Anderson, E. S. (1999). What is the Point of Equality? *Ethics*, 109, 287–337.

Arneson, R. J. (1989). Equality and Equal Opportunity for Welfare. *Philosophical Studies*, 56, 77–93.

—— (2000). Luck Egalitarianism and Prioritarianism. *Ethics*, 110, 339–49.

Barry, B. (1991). *Liberty and Justice: Essays in Political Theory*, ii. Oxford: Oxford University Press.

Bossert, W. (1995). Redistribution Mechanisms Based on Individual Characteristics. *Mathematical Social Sciences*, 29, 1–17.

—— and Fleurbaey, M. (1996). Redistribution and Compensation. *Social Choice and Welfare*, 13, 343–55.

—— —— and Van de Gaer, D. (1999). Responsibility, Talent, and Compensation: A Second-Best Analysis. *Review of Economic Design*, 4, 35–56.

Cappelen, A. W., and Tungodden, B. (2002). Responsibility and Reward. *FinanzArchiv*, 59, 120–40.

—— —— (2003). Reward and Responsibility: How Should we be Affected When Others Change their Effort?. *Politics, Philosophy & Economics*, 2, 191–211.

Cohen, G. A. (1989). On the Currency of Egalitarian Justice. *Ethics*, 99, 906–44.

Dworkin, R. (1981). What is Equality? Part 2: Equality of Resources. *Philosophy & Public Affairs*, 10, 283–345.

—— (2000). *Sovereign Virtue: The Theory and Practice of Equality.* Cambridge, MA.: Harvard University Press.

Fleurbaey, M. (1994). On Fair Compensation. *Theory and Decision*, 36, 277–307.

—— (1995*a*). Equal Opportunity or Equal Social Outcome? *Economics and Philosophy*, 11, 25–56.

—— (1995*b*). Three Solutions for the Compensation Problem. *Journal of Economic Theory*, 65, 505–21.

—— (2008). *Fairness, Responsibility and Welfare.* Oxford: Oxford University Press.

—— and Maniquet, F. (1996). Fair Allocation with Unequal Production Skills: The No-Envy Approach to Compensation. *Mathematical Social Sciences*, 32, 71–93.

—— —— (1999). Fair Allocation with Unequal Production Skills: The Solidarity Approach to Compensation. *Social Choice and Welfare*, 16, 569–83.

——— ——— (2005). Fair Orderings with Unequal Production Skills. *Social Choice and Welfare*, 24, 93–128.
——— ——— (2006). Fair Income Tax. *Review of Economic Studies*, 73, 55–83.
——— ——— (2007). Help the Low-Skilled or Reward the Hard-Working: A Study of Fairness in Optimal Income Taxation. *Journal of Public Economic Theory*, 9, 467–500.
——— ——— (2009). Compensation and Responsibility. In K. J. Arrow, A. K. Sen, and K. Suzumura (eds.), *Handbook of Social Choice and Welfare*, ii. Amsterdam: North-Holland.
Gaspart, F. (1998). Objective Measures of Well-Being and the Cooperative Production Problem. *Social Choice and Welfare*, 15, 95–112.
Hild, M., and Voorhoeve, A. (2004). Equality of Opportunity and Opportunity Dominance. *Economics and Philosophy*, 20, 117–46.
Hurley, S. L. (2003). *Justice, Luck, and Knowledge*. Cambridge, MA.: Harvard University Press.
Iturbe-Ormaetxe, I. (1997). Redistribution and Individual Characteristics. *Review of Economic Design*, 3, 45–55.
——— and Nieto, J. (1996). On Fair Allocations and Monetary Compensations. *Economic Theory*, 7, 125–38.
Kolm, S. C. (1972). *Justice et équité*. Paris: Ed. du CNRS. Repr. and trans. as *Justice and Equity*, Cambridge, MA.: MIT Press, 1999.
——— (2004). *Macrojustice: The Political Economy of Fairness.* New York: Cambridge University Press.
Maniquet, F. (1998). An Equal-Right Solution to the Compensation–Responsibility Dilemma. *Mathematical Social Sciences*, 35, 185–202.
——— (2004). On the Equivalence between Welfarism and Equality of Opportunity. *Social Choice and Welfare*, 23, 127–48.
Matravers, M. (2002). Responsibility, Luck, and the "Equality of What?" Debate. *Political Studies*, 50, 558–72.
Moulin, H. (1994). La présence d'envie: comment s'en accommoder? *Recherches Economiques de Louvain*, 60, 63–72.
O'Neill, D., Sweetman, O., and Van de gaer, D. (2000). Equality of Opportunity and Kernel Density Estimation: An Application to Intergenerational Mobility. In T. B. Fomby and R. C. Hill (eds.), *Applying Kernel and Nonparametric Estimation to Economic Topics*, Advances in Econometrics, 14. Stamford, CT.: JAI Press.
Ooghe, E., and Lauwers, L. (2005). Non-Dictatorial Extensive Social Choice. *Economic Theory*, 25, 721–43.
——— Schokkaert, E., and Van de gaer, D. (2007). Equality of Opportunity versus Equality of Opportunity Sets. *Social Choice and Welfare*, 28, 209–30.
Pazner, E., and Schmeidler, D. (1974). A Difficulty in the Concept of Fairness. *Review of Economic Studies*, 41, 441–3.
Peragine, V. (1999). The Distribution and Redistribution of Opportunity. *Journal of Economic Surveys*, 13, 37–69.
——— (2002). Opportunity Egalitarianism and Income Inequality. *Mathematical Social Sciences*, 44, 45–64.
——— (2004). Measuring and Implementing Equality of Opportunity for Income. *Social Choice and Welfare*, 22, 187–210.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA.: Harvard University Press.
—— (1982). Social Unity and Primary Goods. In A. K. Sen and B. Williams (eds.), *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.
Roemer, J. E. (1985). Equality of Talent. *Economics and Philosophy*, 1, 151–87.
—— (1993). A Pragmatic Theory of Responsibility for the Egalitarian Planner. *Philosophy and Public Affairs*, 22, 146–66.
—— (1998). *Equality of Opportunity*. Cambridge, MA.: Harvard University Press.
—— (2002*a*). Egalitarianism Against the Veil of Ignorance. *Journal of Philosophy*, 99, 167–84.
—— (2002*b*). Equality of Opportunity: A Progress Report. *Social Choice and Welfare*, 19, 455–71.
—— (2004). Equal Opportunity and Intergenerational Mobility: Going beyond Intergenerational Income Transition Matrices. In M. Corak (ed.), *Generational Income Mobility in North America and Europe*. Cambridge: Cambridge University Press.
—— *et al.* (2003). To What Extent do Fiscal Regimes Equalize Opportunities for Income Acquisition Among Citizens? *Journal of Public Economics*, 87, 539–65.
Scheffler, S. (2003). What is the Value of Equality? *Philosophy and Public Affairs*, 31, 5–39.
Schluter, C., and Van de gaer, D. (2002). Mobility as Distributional Difference. Mimeo, University of Bristol.
Schokkaert, E., and Van de Voorde, C. (2004). Risk Selection and the Specification of the Conventional Risk Adjustment Formula. *Journal of Health Economics*, 23, 1237–59.
—— Dhaene, G., and Van de Voorde, C. (1998). Risk Adjustment and the Tradeoff between Efficiency and Risk Selection: An Application of the Theory of Fair Compensation. *Health Economics*, 7, 465–80.
—— Van de gaer, D., Vandenbroucke, F., and Luttens, R. (2004). Responsibility-Sensitive Egalitarianism and Optimal Linear Income Taxation. *Mathematical Social Sciences*, 48, 151–82.
Sen, A. K. (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.
—— (1992). *Inequality Reexamined*. Oxford: Clarendon Press.
Sprumont, Y. (1997). Balanced Egalitarian Redistribution of Income. *Mathematical Social Sciences*, 33, 185–202.
Tungodden, B. (2005). Responsibility and Redistribution: The Case of First Best Taxation. *Social Choice and Welfare*, 24, 33–44.
Van de gaer, D. (1993). Equality of Opportunity and Investment in Human Capital. Ph.D. thesis, K. U. Leuven.
—— Martinez, M., and Schokkaert, E. (2001). Three Meanings of Intergenerational Mobility. *Economica*, 68, 519–38.
Vandenbroucke, F. (2001). *Social Justice and Individual Ethics in an Open Society: Equality, Responsibility, and Incentives*. Berlin: Springer-Verlag.
van der Veen, R. (2004). Basic Income versus Wage Subsidies: Competing Instruments in an Optimal Tax Model with a Maximin Objective. *Economics and Philosophy*, 20, 147–84.
Van Parijs, P. (1995). *Real Freedom for All*. Oxford: Oxford University Press.
Varian, H. (1974). Equity, Envy and Efficiency. *Journal of Economic Theory*, 9, 63–91.
Wolff, J. (1998). Fairness, Respect and the Egalitarian Ethos. *Philosophy and Public Affairs*, 27, 97–122.

CHAPTER 17

EQUALITY AND PRIORITY

BERTIL TUNGODDEN

17.1 Introduction

Most people care about inequalities. But why? Scanlon (2000) suggests that this is mainly due to the *instrumental* value of equality.

> I find that my reasons for favoring equality are in fact quite diverse, and that most of them can be traced back to fundamental values other than equality itself. The idea that equality is, in itself a fundamental moral value turns out to play a surprisingly limited role in my reasons for thinking that many of the forms of inequality which we see around us should be eliminated. (p. 21)

A reduction in inequality may, among other things, alleviate suffering, the feeling of inferiority, the dominance of some over the lives of others; and in many cases these effects are of sufficient importance to motivate our concern for the alleviation of inequality (Anderson 1999). But many people think that there are reasons for caring about equality that are independent of its instrumental value, and it is the plausibility of assigning *intrinsic* value to equality that has been seriously questioned in the recent literature on prioritarianism.

In most of the debate on egalitarianism and prioritarianism, the framework has been narrowed down to a comparison of distributions of well-being (Parfit

Much of the material in this paper is also presented in Tungodden (2003). I should like to thank Paul Anand and an anonymous referee for most valuable comments.

1995).[1] Hence, the assumption has been that for any population $N = \{1, \ldots, n\}$, each social alternative is characterized by an n-dimensional well-being vector $x = (x_1, \ldots x_n)$, where x_i is the well-being of person i in society.[2] Moreover, it is standardly assumed that the framework satisfies a minimal condition of anonymity, saying that the identity of an individual should not influence our reasoning (if we consider two alternatives $x = (1, 2, 3)$ and $y = (2, 1, 3)$, then the minimal condition of anonymity says that we should be indifferent between x and y). Within this framework, the question has been how, and to what extent, one should take into account that one alternative is more equal than another when ranking the alternatives in terms of a moral betterness relation.

The chapter is organized as follows. In Section 17.2, I consider what is commonly called the leveling-down objection to egalitarianism and how it relates to the principle of personal good. Section 17.3 contains a discussion of how a concern for equality should affect our social evaluations. In particular, I discuss the link between equality promotion and Rawlsian reasoning and how the value of equality may be combined with utilitarian reasoning. In Sections 17.4 and 17.5, I provide a discussion of prioritarianism and how this perspective relates to egalitarianism. In Section 17.4, I present the standard framework of prioritarianism, whereas in Section 17.5 I review the literature on prioritarianism and uncertainty.

17.2 The Leveling-Down Objection and the Principle of Personal Good

It is commonly believed that egalitarians should accept the following principle:

The Weak Principle of Equality: If one alternative is more equal than another, it is better in one respect.

[1] This is not an uncontroversial restriction of the problem at hand; see among others Rawls (1971, 1993); Sen (1980, 1992); Dworkin (1981); Cohen (1989); and Scanlon (1993). I will also assume that there are no informational biases, such that we have a quantitative notion of well-being. This is in contrast to much of the economics literature in this field, where the focus has been on the implications of informational constraints on our understanding of egalitarianism; see Bossert and Weymark (1999) for a survey. For other surveys on egalitarian reasoning, see among others Kolm (1996); Roemer (1996); Scanlon (1998); Pojman and Westmoreland (1997); Clayton and Williams (2000); and Holtug and Lippert-Rasmussen (2006).

[2] In other words, well-being is ultimately measured in a one-dimensional way. This includes welfarist theories (see Blackorby, Donaldson, and Weymark (1984) for a formal definition), but also allows for many other interpretations of well-being.

However, it has been argued that this principle faces a serious problem, which Parfit (1995) names *the leveling-down objection*.[3] A reduction in inequality can take place by harming the better-off in society without improving the situation of the worse-off. But this cannot be good *in any respect*, contrary to the claim of the weak principle of equality. Hence, according to the objection, inequality cannot be intrinsically bad.

This objection does not attack any particular restriction that egalitarians are committed to impose on the betterness relation.[4] Its target is the way in which egalitarians have to justify any particular betterness ranking in cases where there is a loss for the better-off and no gain for the worse-off. Even though egalitarians may insist that such a loss makes things worse all things considered, they have to accept that it is better in one respect. Or at least, this is what the leveling-down objection claims.

In my view, the leveling-down objection does not really challenge egalitarianism as a viable normative position. Even if one should accept the premises of the leveling-down objection, one is not committed to the view that equality promotion is never valuable. As argued by Kagan (1988), Kamm (1996), and Temkin (2000), a principle may have genuine significance in some settings even if it lacks significance in other settings. Hence, we may defend an egalitarian position saying that equality promotion is relevant only in solving distributive conflicts in society, and that in all other cases we should follow the following version of the Pareto principle (introduced by Broome 1991):

The Principle of Personal Good: For all alternatives x and y, if everyone is at least as well off in x as in y and someone is strictly better off, then x is better than y.

I now turn to a discussion of how one may combine this principle with a concern for equality promotion.

17.3 Moderate Egalitarianism

In order to study the implications of equality promotion, we have to clarify further our understanding of the concept of inequality. It is trivial to say that equality is better than inequality. But we need more than this. We need to compare different unequal distributions. There has been much formal work on this within

[3] See also Temkin (1993, 2000, 2003); Holtug (1998); and Wolff (2000).

[4] Even though the weak principle of equality has some implications for the betterness relation. If x is more equal than y and not worse in any respect, then the weak principle of equality implies that x is better than y. See also Klint Jensen (2003) and Brown (2003).

economics (see among others Atkinson 1970; Sen 1973; Dasgupta, Sen, and Starrett 1973; Kolm 1976 *a*, *b*; Blackorby and Donaldson 1978, 1980; Shorrocks 1980; Bossert and Pfingsten 1990; and for overviews, Lambert 1993; Sen and Foster 1997; and Cowell 2000), but I will take as the point of departure the claim of Vallentyne (2000):

> All plausible conceptions of equality hold that, where perfect equality does not obtain . . . any benefit (no matter how small) to a worst off person that leaves him/her still worst off person has priority (with respect to equality promotion) over any benefit (no matter how large) to a best off person. (p. 1)

This is a very weak restriction on our conception of *equality*, and it is satisfied by all well-known inequality measures. Actually, this is also true for a slightly stronger view, which allows for more than one best-off person.

Strong Conditional Contracting Extremes (on equality): For all alternatives x and y, if (1) all the best-off persons in x are best-off persons in y and their well-being level is strictly lower in x than y; (2) all the worst-off persons in x are worst-off persons in y and their well-being level is strictly higher in x than y; and (3) the well-being of everyone else is the same in x and y; then x is more equal than y.

Let us now consider the case where we care *only* about equality promotion when solving distributive conflicts, captured by the following condition on the betterness relation:

Strict Priority to Equality Promotion: For all alternatives x and y, if (1) there are persons with higher well-being in x than y and persons with higher well-being in y than x, and (2) x is more equal than y, then x is better than y.

We may define *strict moderate egalitarianism* as the position that imposes a minimal condition of anonymity, the principle of personal good, and strict priority to equality promotion on the betterness relation. Given our minimal equality condition, strict priority to equality promotion places the following restriction on the betterness relation:

Strong Conditional Contracting Extremes (on betterness): For all alternatives x and y, if (1) all the best-off persons in x are best-off persons in y and their well-being level is strictly lower in x than y; (2) all the worst-off persons in x are worst-off persons in y and their well-being level is strictly higher in x than y; and (3) the well-being of everyone else is the same in x and y; then x is better than y.

Let me stress that this condition is restricting the betterness relation only with respect to distributive conflicts between the *best-off* and the *worst-off*. For all other cases, it is silent. Hence, it does not rule out the possibility of taking into account the size of gains and losses when there is a conflict between, say, the worst-off and the second worst-off (as long as the second worst-off is not also the best-off). To

illustrate, consider $x = (2, 10, 100)$ and $y = (1, 100, 100)$. Many well-known inequality measures would provide support for the conclusion that there is more inequality in x than in y. If so, then strict priority to equality promotion implies that y is better than x.

However, if we impose transitivity, then the betterness relation must satisfy the following maximin property if it satisfies anonymity, the principle of personal good, and strong conditional contracting extremes on betterness.

Maximin: For all alternatives x and y, if the level of well-being in the worst-off position is strictly higher in x than y, then x is better than y.

Consequently, if we think that maximin sometimes violates equality promotion, then we have an impossibility result.[5] Let me briefly illustrate this impossibility with an example. Suppose that $y = (1, 100, 100)$ is considered more equal than $x = (2, 10, 100)$, and hence that strict priority to equality promotion implies that y is better than x. Compare x with $z = (2, 10, 10)$. From the principle of personal good, it follows that x is better than z. By transitivity, we now have that y is better than z. But this violates strict priority to equality promotion according to the minimal requirement of strong conditional contracting extremes on equality.

Moreover, if we are willing to accept a further restriction on the concept of equality, then we can establish a complete link between strict moderate egalitarianism and the stronger leximin principle.[6] Vallentyne (2000, p. 6) argues that equality is increased if there is a decrease in the well-being of a person above the mean who stays above the mean, an increase in the well-being of a person below the mean who stays below the mean, and no changes elsewhere in the distribution.[7] If we accept this suggestion and impose strict priority on equality promotion, the principle of personal good, anonymity, and transitivity on the betterness relation, then we have a characterization of the leximin principle.[8] In sum, this shows that there is a very close link between equality promotion and Rawlsian reasoning (see also Barry 1989, pp. 229–34).

There is another interesting link between equality promotion and the leximin principle, and that is by imposing a separability condition on the betterness

[5] See Tungodden (2000 *a*, *b*) for a detailed discussion of this result, and Tungodden and Vallentyne (2005) for a discussion of possible ways of escaping this impossibility. See also Bosmans (2006, 2007*a*, 2007*b*).

[6] The leximin principle states that if the worst-off is at the same level in the two alternatives, then we should assign absolute priority to the second worst-off, and so on. For a critical discussion of the link between the leximin principle and the difference principle of Rawls, see Tungodden (1999) and Tungodden and Vallentyne (2006). See also Van Parijs (2001) for a thorough discussion of the difference principle.

[7] This is also suggested by Temkin (1993, p. 25).

[8] See Tungodden (2000*a*) for a further discussion of this result. Note that Hammond (1976, 1979) was the first to show how an objection to inequality between any two groups leads to maximin. I will return to Hammond's result shortly.

relation.[9] Strong separability demands that we also solve distributive conflicts in a way that is independent of the well-being of indifferent people. In order to define this condition formally, let M denote a subgroup of the total population N and $\bar{M}$ the rest of the population.

Strong Separability: For all alternatives x, y, z, w, if (1) for every person $j \in M$, j has the same utility level in x as in z and in y as in w, and (2) for every person $j \in \bar{M}$, j has the same utility level in x as in y and in z as in w, then x is better than y if and only if z is better than w.

As an illustration, consider an example suggested by Broome (forthcoming). We have four alternatives $c = (2, 2, 2, 2, 2, 2, 2, 2)$, $d = (4, 1, 2, 2, 2, 2, 2, 2)$, $e = (2, 2, 1, 1, 1, 1, 1, 1)$, and $f = (4, 1, 1, 1, 1, 1, 1, 1)$. If the betterness relation satisfies strong separability, then we know that c is better than d if and only if e is better than f. However, if we want to solve these conflicts by giving strict priority to equality promotion, then it might seem as if we have to abandon the demand for strong separability. In this example, it is obvious that c is more equal than e, and hence one could think that it is futile, *within an egalitarian framework*, to demand consistency in the way we rank c to d and e to f.

However, I will argue that this is not the case. We may defend a version of moderate egalitarianism along the lines suggested by Nagel (1979, 1991), where we seek a result which is acceptable to each person involved.

> Oddly enough, egalitarianism is based on a more obscure conception of moral equality than either of the less egalitarian theories. . . . Something close to unanimity is being invoked . . . The essence of such a criterion is to try in a moral assessment to include each person's point of view separately, so as to achieve a result which is in a significant sense acceptable to each person involved or affected. (Nagel 1979, pp. 116–23)

Given this framework, we can safely ignore the indifferent people, and moreover we may argue that conflicts should be solved by assigning strict priority to equality promotion *within the group of people involved in the conflict*. In order to state this in a more formal manner, let us for any two alternatives x and y define x^y as the truncated version of x where we have deleted every person being indifferent between x and y. Hence, as an example, if $x = (1, 4, 6, 10, 15)$ and $y = (1, 9, 12, 13, 15)$, then $x^y = (4, 6, 10)$ and $y^x = (9, 12, 13)$.

Strict Priority to Equality Promotion within the Group of People Involved in the Conflict: For all alternatives x and y, if (1) there are persons with higher well-being

[9] This topic is in fact of much practical importance, because a separable betterness relation makes possible a decomposable approach to policy considerations. Sen and Foster (1997) discuss this issue at some length, but remark that "even if one accept the *usefulness* of decomposability, one might still wonder about its *acceptability* as a general condition" (p. 156).

in x than y and persons with higher well-being in y than x, and (2) x^y is more equal than y^x, then x is better than y.

This condition, together with the principle of personal good, imposes strong separability on the betterness relation. Hence, it is possible to combine an a priori demand for strong separability with a version of moderate egalitarianism. Of course, to appeal to equality promotion within a group is certainly not the same as to appeal to equality promotion in society at large; but at the same time it is clearly an egalitarian perspective. It does not appeal to anything other than equality promotion within the group of people involved in the distributive conflict.

Consider now any two-person conflict. It is quite obvious that equality is promoted between the worse-off and the better-off by giving absolute priority to the worse-off, and hence strict priority to equality promotion within the group of people involved in the conflict implies the following condition on the betterness relation, suggested by Hammond (1976, 1979).

The Hammond Equity Condition: For all alternatives x and y, if there exist j and k such that (1) the well-being level of j is strictly lower in x than y, (2) the well-being level of k is strictly higher in x than y, (3) j has strictly higher well-being level than k in x, and (4) the utility of everyone else is the same in x and y, then x is better than y.

To illustrate the condition, consider $x = (1, 3, 7, 8)$, $y = (1, 3, 6, 9)$. Hammond equity implies that x is better than y, and it is easily seen that this promotes equality within the group of people involved in the distributive conflict. $x^y = (7, 8)$ is clearly more equal than $y^x = (6, 9)$ (which also follows from strong conditional contracting extremes on equality). As shown by Hammond, this is all we need to characterize the leximin principle within our framework within a consistent framework satisfying the principle of personal good.

In both philosophy and economics, there has been considerable concern about how to combine egalitarian reasoning with a concern for the utilitarian perspective.[10] Of course, egalitarians do not want to embrace the utilitarian betterness relation, but they may still find the following principle appealing:

Weak Utilitarianism: If one alternative has more total utility than another, it is better in one respect.

If we endorse weak utilitarianism, then we need to clarify how to balance a concern for equality with a concern for total well-being. Before entering into the problem of balancing, though, I believe there is a more fundamental question to ask. If you are an *egalitarian*, then *why* should you care about utilitarian reasoning *at all*? If we read Parfit (1995) on this, it becomes clear that he does not make a distinction between the principle of personal good and utilitarian reasoning.

[10] By introducing utilitarianism, I do not impose a particular interpretation of the concept of well-being. Here, my concern is the idea of assigning value to the total amount of well-being.

Suppose next that the people in some community could all be either (1) equally well off, or (2) equally badly off. The [weak] Principle of Equality does not tell us that (2) would be worse. This principle is about the badness of inequality; and, though it would be clearly worse if everyone were equally worse off, our ground for thinking this cannot be egalitarian.

To explain why (2) would be worse, we might appeal to [weak utilitarianism] . . . When people would be on average better off, or receive a greater net sum of benefits, we can say, for short, that there would be more [*well-being*] . . . If we cared only about equality, we would be *Pure* Egalitarians. If we cared only about [well-being], we would be Pure Utilitarians—or what is normally called Utilitarians. But most of us accept a *pluralist* view: one that appeals to more than one principle or value. (p. 4)

When comparing (1) and (2) in Parfit's example, it would be sufficient to appeal to the principle of personal good. Parfit, on the other hand, defends (1) by appealing to weak utilitarianism. That is unfortunate, because there is a fundamental difference between these two principles. Anyone ought to accept the principle of personal good, whereas weak utilitarianism is more controversial. Actually, many egalitarians seem to reject utilitarian reasoning altogether, and on this basis they might think that they should reject a pluralistic egalitarian theory as well. This is suggested, for example, by McKerlie (1994):

And those egalitarians who believe that there is something fundamentally wrong with the kind of thinking done by the utilitarian principle would not be willing to include it (or any other principle formally like it) in the combined view. (p. 27)

Notice that this view rejects not only the utilitarian betterness relation, but also weak utilitarianism. These egalitarians do not see any value in the total amount of utility in society; it is simply an irrelevant aspect of the situation. However, as I have shown, egalitarians do not have to include utilitarian reasoning in order to have a workable theory. It is sufficient that they accept the principle of personal good.

This is not to say that weak utilitarianism ought to be rejected by egalitarians. As illustrated by Kymlicka (1988), it might be defended as a way of expressing moral equality. And it could be the case that some egalitarians want to combine these two ways of expressing moral equality (see e.g. Nagel 1979, p. 122).[11] Moreover, other egalitarians may want to include utilitarian reasoning even though they reject it as an expression of moral equality, arguing that the appropriate expression of moral equality is not the only value of importance.

Let *weak moderate egalitarianism* be the name of the set of positions that combine a concern for equality with a concern for total well-being. This framework allows for a number of specific approaches, though the nature of these approaches depends on our interpretation of the previous characterization of the leximin principle. If we endorse my favorite interpretation and acknowledge that the leximin principle always promotes equality (in distributive conflicts), then a weak moderate egalitarian

[11] On the other hand, Nagel (1991, p. 78) rejects the idea that utilitarianism represents a reasonable expression of the moral equality of people.

would simply be someone who weighed the utilitarian and the leximin argument (that is, weighed the mean and the well-being of the worst-off). There would be no reason to allow for other weighting schemes, because in this case we think that the leximin principle captures all there is to say about equality promotion. On the other hand, if we think that the leximin principle is an imperfect framework for equality promotion, then we might consider alternative approaches tenable when aiming at combining equality promotion with utilitarian reasoning.

Usually, economists have taken the Pigou–Dalton criterion of transfer as the point of departure for a discussion of moderate egalitarianism.[12]

The Pigou–Dalton Principle of Transfer: For all alternatives x and y, if there exist j and k such that (1) the well-being gain of j is equal to the well-being loss of k when moving from y to x, (2) j has a lower well-being level than k in x, and (3) the utility of everyone else is the same in x and y, then x is better than y.

Even if we accept the Pigou–Dalton principle as a restriction on any egalitarian betterness relation, we should notice that this condition allows for a very broad interpretation of the set of egalitarian betterness relations. There are betterness relations within this framework that do not pay very much attention to equality promotion. The most extreme case would be what we may name quasi-egalitarian utilitarianism, which assigns weight to equality considerations *only* when the total amount of well-being is the same in society. In all other cases, it follows the utilitarian betterness relation. This approach satisfies the Pigou–Dalton principle, but for all practical purposes it is a utilitarian approach. Of course, if we demand a continuous betterness relation, then we exclude this approach and the leximin principle (which is the other extreme of moderate egalitarianism).

17.4 Prioritarianism and Sufficiency

Consider again the case where there is a conflict between the best-off and the worst-off in society. In order to promote *equality*, we have to assign absolute priority to the worst-off in all these cases. And the reason for this is that the other person involved in the conflict is *the best-off*. Hence, it is independent of whether the best-off lives in extreme destitution or has a very good life. But I assume that most people think otherwise. I believe that most people find it much harder to assign absolute priority to the worst-off if both live in destitution. In other words, most of us take

[12] Often, and originally, this condition is stated in the space of income (see Dalton 1920, p. 352), but for our purpose it is appropriate to express it in the space of well-being. See Sen and Foster (1997) for further discussion and definitions, and Tungodden (2003) for a discussion of possible counter-arguments.

into account the *absolute* circumstances of people when evaluating to what extent to assign priority to the worse-off in a distributive conflict.

Roughly speaking, this is the message of prioritarians. And it is an important one. It highlights the fact that there are different ways of *justifying* any distributive principle we impose on the betterness relation. Still, the fact that the absolute circumstances of people affect our evaluations is not news to economists or philosophers, and hence we may wonder why prioritarianism has been considered with so much interest in recent philosophical debate. In order to answer this question, it is useful to have a brief look at how prioritarianism has been introduced among philosophers. The most prominent contribution on prioritarianism is Parfit (1995), who defines the approach as follows:[13]

The Priority View: Benefiting people matters more the worse off these people are.

However, as remarked by Parfit (1995) himself, the definition is imprecise, because it does not clearly distinguish prioritarianism from egalitarianism.

> But this claim by itself, does not define a different view, since it would be made by all Egalitarians. If we believe that we should aim for equality, we shall think it more important to benefit those who are worse off. Such benefits reduce inequality. If that is why we give such benefits priority, we do not hold the Priority View. On this view, as I define it here, we do *not* believe in equality. We give priority to the worse off, not because this will reduce inequality, but for other reasons. (p. 22)

Even if you give priority to the worse-off, you do not necessarily hold the priority view according to the definition of Parfit. What matters is *why* you give priority. In other words, Parfit does not define the distinction between egalitarianism and prioritarianism by different ways of restricting the betterness relation, but by different ways of justifying any principle imposed on the betterness relation.

Prioritarianism can be defended in a negative and a positive way. The positive approach is to defend prioritarianism on its own: that is, to show that it captures an important point of view when reasoning about principles to impose on the betterness relation. The negative approach is to defend it by showing that it represents one way of escaping a number of problems facing standard egalitarian justification. Much of the philosophical literature applies the negative approach. By way of illustration, when Parfit summarizes his discussion on egalitarianism and prioritarianism (1995, p. 34), he introduces the priority view as an option that we can move to when we realize the problems facing the egalitarian approach to distributive justice.[14]

[13] Weirich (1983) is an early philosophical discussion of formal rules capturing the prioritarian intuition.

[14] However, it should also be mentioned that Parfit (1995, p. 22) does not deny the possibility of combining egalitarianism and prioritarianism. In my view, this is an option too easily forgotten in the debate on equality versus priority. See also Peterson and Hansson (2005).

What problems of justification do we avoid when moving from egalitarianism to prioritarianism? First, Parfit (1995, p. 22) suggests that it is an advantage that prioritarianism can be considered a complete moral view, in contrast to any plausible version of egalitarianism that ought to be combined with another principle. This fact is also pointed at by McKerlie (1994, p. 27): "some egalitarians regret the fact that the equality view must be combined with another principle. They want a simpler alternative to utilitarianism, and they object to the intuitive nature of the judgments we must make in weighing the reasons provided by the two principles against one another."[15] But this is odd. Prioritarianism, as it is stated, is also intuitionist (Parfit 1995, p. 20), because it does not tell us how much priority to assign to the worse-off. Hence, the only difference in this respect is that in the egalitarian case we have to rely on intuition when justifying the tradeoff we make between different principles, whereas in the prioritarian case we have to rely on intuition when justifying any particular interpretation of the principle of priority. It is hard to see that this distinction is significant.

Second, Parfit (1995, p. 23) stresses that by endorsing the priority view we avoid the leveling-down objection. Certainly, on the basis of a concern for the absolute circumstances of people, there is nothing to be gained by reducing the level of well-being of the better-off. But I have already argued against the importance of the leveling-down objection, and hence, in my view, this move should not count for much.

Finally, as we have seen in the previous section, many philosophers have been reluctant to include utilitarian reasoning in their justification of any particular betterness relation. In this respect, they have considered the prioritarian approach more appropriate than weak moderate egalitarianism, because it allows for a different way of justifying a concern for gains and losses of the better-off. Eventually, gains and losses are included in the prioritarian framework if we do not assign infinitely more importance to improving the *absolute* circumstances of poorer people than better-off people, and not, as in utilitarianism, because we care about the *total* level of well-being. This is, in my view, an interesting argument, and it has not been properly recognized by economists. Economists have tended to assume that any betterness relation that can be represented as the outcome of a tradeoff between a concern for utility and equality necessarily rests partly on utilitarian justification. Prioritarianism shows that this need not be the case. And even though economists certainly have realized that there is an alternative representation of such betterness relations that avoids any reference to total utility, to wit by a social welfare function defined directly on individual well-being, I think it is fair to say that economists have not acknowledged that this formulation may invite an alternative justification of the way we restrict the betterness relation.

[15] Of course, this is only the case of weak moderate egalitarianism. See also Rawls (1971, pp. 34–40).

More importantly, economists have not seen that this formulation also may invite an alternative justification of our concern for the worse-off. The standard view has been that any betterness relation that can be represented as the outcome of a tradeoff between a concern for utility and a concern for equality necessarily reflects a concern for equality (see e.g. Sen and Foster 1997, p. 123). Hence, even though economists have been aware of the fact that we may care about both the absolute and the relative circumstances of the worse-off, they have not considered how these aspects may constitute different kinds of justification. Prioritarianism, however, shows that this is the case. Thus, in my view, the main contribution of prioritarianism is not the introduction of a completely new idea (that absolute circumstances should count in distributive reasoning has been suggested by many), but the clarification of how this idea constitutes a distinctive way of justifying a concern for the worse-off. I consider this a positive reason for adopting the prioritarian perspective. Prioritarian justification of priority to the worse-off is essential in its own right, and not only as a way of (possibly) escaping problems facing egalitarian reasoning.

This is most clearly seen if we consider a set of cases where economists have certainly recognized that justification of priority to the worse-off cannot be based on purely egalitarian grounds, even in combination with utilitarian reasoning. The cases I have in mind are those that include an *absolute poverty line.* Most of us recognize the special importance of improving the lives of poor people, and hence should like to include this in our scheme of justification (Raz 1986, p. 240). However, in order to do that, we need to adopt prioritarian reasoning.

It has been argued by some philosophers that an absolute threshold is all that matters in distributive reasoning. In particular, Frankfurt (1987, p. 22) suggests *the doctrine of sufficiency*:[16] "If everyone had enough, it would be of no moral consequence whether some had more than others" (p. 21). Hence, according to Frankfurt, we should assign priority to those below this sufficiency threshold in a conflict with people who have enough; but there is no reason to assign priority to the worse-off among people who have enough. Even though this is not usually considered a prioritarian doctrine, I believe it highlights an essential issue within prioritarianism: namely, to what extent an absolute threshold should affect our justification of priority to the worse-off.

The sufficiency approach faces at least two challenges.[17] First, it needs to explain what it means to say that someone has enough.[18] Second, we need to know why

[16] For a critical discussion of Frankfurt's argument, see Goodin (1987).

[17] See also Crisp (2000), who outlines a version of sufficiency based on the notion of compassion.

[18] Rosenberg (1995, p. 66) argues that "[o]perationalizing sufficiency is probably far easier than establishing equal shares". Surely, it is hard to operationalize the ideal of equality, but in order to compare this task with the doctrine of sufficiency, we have to determine what it means that someone has enough. Hence, a priori it is hard to say whether the need for a practical standard counts in favor of a doctrine of sufficiency or not.

we should assign priority only to those below the sufficiency threshold. As I see it, there are two ways of understanding the idea of having enough. One is to argue that there is this feeling of contentment (or absence of distress) which can be satisfied with a certain amount of money, and which we can argue should be included as a need in an expanded version of the idea of an absolute poverty line.[19] The other interpretation, relying on Frankfurt's claim that reasonable people ought to feel content at a certain level of well-being, is moral, and is that there is no reason (from a person's point of view) to object to unequal distributions of well-being as long as this person has enough. In other words, the sufficiency level defines the level of well-being above which there is no reason to complain.

If we accept this latter definition, it follows directly that we should pay no attention to people above the sufficiency level in a distributive conflict. Arneson (2000), Nagel (1991, p. 81), and Temkin (2003) clearly reject such a view of distributive justice, and in my view, a more plausible reading of an absolute threshold is that it represents a level of well-being where there is a fundamental change in the moral significance of people's claims in a distributive conflict. This does not rule out a concern for people above the absolute threshold, and it does not rule out the possibility of assigning priority to the worse-off within this group.

Of course, it is not easy to draw any such line, and in that respect it is important to notice the work of economists on fuzzy poverty lines.[20] But I think that most people share the intuition that there is a fundamental difference between the complaints of a person living in destitution and the complaints of a person living a good life. We may say that this illustrates a case where the better-off person has enough (in order to fulfill all important needs), and hence where we assign strong (maybe absolute) priority to the poor person (without rejecting the relevance of the claim of the better-off person).

I believe that the notion of an absolute threshold is of fundamental importance, and that it represents the most important reason for including the prioritarian point of view within any reasonable moral conception of the distributive problem. Of course, it is hard to determine how much more importance to assign to the needs of a poor person in a conflict with a person above the threshold. But I think that this particular aspect of prioritarianism is fairly well recognized by economists (even though some economists will insist on a purely relative notion of poverty), and that the more fundamental lesson learned by the recent contribution of prioritarian philosophers is that our concern for the absolute circumstances of people can be

[19] The inclusion of the feeling of contentment in the definition of an absolute threshold may cause a relative threshold in the space of income (as pointed out more generally in Sen 1983). See also Rosenberg (1995, p. 67), who defends the doctrine of sufficiency on the basis of an idea about what are the "real interests" of a person.

[20] Again, see Sen and Foster (1997, pp. 188–91) for an overview. See also Lemmi and Betti (2006) for a number of interesting contributions on the fuzzy set approach to multidimensional poverty measurement.

expanded to a more general theory of justification (as suggested by Nagel (1991, pp. 69–70) among others).

So far I have talked mainly about prioritarianism and the sufficiency approach as ways of justifying a concern for the worse-off. Let me now comment on how prioritarian justification restricts the betterness relation. It should be clear that any prioritarian betterness relation needs to satisfy the principle of personal good and the Pigou–Dalton principle. The essence of prioritarianism is to improve the absolute circumstances of people (which implies endorsement of the principle of good) and moreover to assign more priority to the worse-off on the basis of absolute circumstances (which implies endorsement of the Pigou–Dalton principle). It may be worthwhile to stress that the Pigou–Dalton principle is an *unquestionable* restriction on a prioritarian betterness relation, because the level of well-being of indifferent people is of no importance when assigning priority on the basis of the absolute circumstances of people. Hence, as an illustration, a rank-order weighting scheme (like the Gini-based ranking rule) cannot be part of a purely prioritarian framework, because such a ranking scheme implies that the level of well-being of indifferent people may play a role in the overall evaluation. More generally, any prioritarian betterness relation must be strongly *separable* in the well-being of individuals.

In sum, an egalitarian position and a prioritarian position potentially differ in two respects; first, in the way they restrict the betterness relation, and second, in the way they justify the restrictions imposed on the betterness relation.[21] But are there betterness relations that cannot be justified on egalitarian and prioritarian grounds? There are two ways of approaching this question. One is to look for *implausible cases*; the other is to look for *impossible cases*. Economists have been eager to look for the impossible cases (Broome, forthcoming; Fleurbaey, forthcoming), whereas philosophers have been more concerned with the implausible cases (McKerlie 1994; Parfit 1995).

Let us first look at the impossible cases. Any betterness relation violating strong separability needs to be justified on the basis of egalitarianism. By way of illustration, compare $x = (1, 4, 4)$, $y = (1, 3, 6)$, $z = (10, 4, 4)$, and $w = (10, 3, 6)$. In this case, suppose that the betterness relation in question states that x is better than y and w is better than z. On the basis of prioritarian justification, we cannot support this conclusion, because in order to do that, we need to assign importance to the well-being level of indifferent people in our evaluation. However, it is not the case that any betterness relation satisfying strong separability ought to be justified on the basis of prioritarian reasoning. As I have argued in Section 17.3, it is certainly possible to defend a strongly separable betterness relation on egalitarian grounds.

Is there any prioritarian betterness relation that cannot be defended on the basis of egalitarian reasoning? Fleurbaey (forthcoming) does not think so.

[21] See also Arneson (1999, 2000).

In short, a prioritarian will always find an egalitarian who advocates the same social ranking. When comparing distributions with the same total amount of benefits, the prioritarian will agree with any egalitarian who measures inequality with the same index that is implicit in the prioritarian's social ranking. (pp. 8–9)

In evaluating this claim, the real issue is whether any inequality index will do the work within an egalitarian framework. Certainly, if the prioritarian betterness relation assigns absolute priority to people below an absolute threshold, but not absolute priority to the worse-off more generally, then it is impossible to defend the index implicit in the ranking as an inequality index that can be established on egalitarian grounds. Leaving aside the idea of an absolute threshold (which is not discussed by Fleurbaey), however, I believe that there are no other cases where we can say that it is impossible to justify a prioritarian betterness relation on the basis of egalitarian reasoning. There might be more cases where this is implausible, but in order to defend such a view, one would have to impose further restrictions on our understanding of inequality.

What about the implausible cases? In the philosophical literature, there has been some discussion about the strength of the leximin argument if derived from prioritarian reasoning and not from some version of egalitarianism. The intuition of philosophers like Parfit (1995) and McKerlie (1994) is that the leximin principle is quite implausible as some version of the priority view.

If we are not concerned with relative levels, why should the smallest benefit to the worst-off person count for infinitely more than much greater benefits to other representative people? (Parfit 1995, p. 39)

If the difference principle is a version of the priority view, it is more vulnerable to the intuitive objection. The objection seems to show that, although we might give greater priority to helping the very worst-off, we do not give it absolute priority. We think that a small gain for them can be morally outweighed by a much larger gain for others who are also badly-off. (McKerlie 1994, p. 33)

It is clear that within the egalitarian framework, we can derive the leximin principle from a set of first principles and thereby avoid intuitionism (Rawls 1971, p. 34), whereas, as I see it, a prioritarian defense of the leximin principle has to be based on intuitive reasoning.[22] This is an important difference, and it might be the case that our intuitions undermine the prioritarian justification of the leximin principle.

In any case, I believe that this discussion of implausible cases points to the most fundamental concern in distributive reasoning: to wit, *how much* priority to assign to the worse-off. On this issue we find strong practical political disagreement, and not on the question about whether we should adopt a separable or nonseparable

[22] Of course, we could imagine deriving leximin from general principles introduced within prioritarianism, but I find it hard to see how this should be done. See also Arneson (1999).

perspective. This is not to say that it is unimportant to clarify the different possible modes of justification.[23] But I think that this exercise is of particular importance if it can guide us on the essential question about the extent of priority to assign to the worse-off.

17.5 Prioritarianism and Uncertainty

So far we have not considered how the prioritarian proposal should be understood in the context of uncertainty. There are two main proposals considered in the literature: *ex ante* prioritarianism and *ex post* prioritarianism (Rabinowicz 2001, 2002; McCarthy 2006, 2007). The analysis of these proposals builds on the work that shows that the axioms of expected utility, applied to a betterness relation over lotteries, yield a representation of the utilitarian betterness relation in line with expected utility theory (Harsanyi 1955, 1975; Broome 1991, 2004).

Ex post prioritarianism takes as a starting point that the priority to the worse-off should be assigned on the basis of the goodness that individuals attain in particular outcomes, whereas *ex ante* prioritarianism assigns priority to the worse-off on the basis of the expected goodness that each person gets from the complete lottery. The main message in the literature is that both perspectives face problems. *Ex ante* prioritarianism implies that the betterness relation is not strongly separable, whereas *ex post* prioritarianism implies a violation of the principle of personal good.

To illustrate the problem of *ex post* prioritarianism, consider a society with two individuals and two lotteries. One lottery x gives a certain outcome of 10 to both individuals, whereas the other lottery y contains two equi-probable outcomes. Let us denote this lottery $y = ([16,5],[5,16])$, where person 1 and person 2 get 16 and 5, respectively, in the first outcome, and vice versa in the second outcome. How should we rank these two lotteries? If we view the numbers as reflecting how good the different outcomes are for each person, and we assume, in line with expected utility theory, that the goodness of a lottery for a person equals expected goodness, then it follows that y is better for both individuals. Hence, according to the principle of personal good defined on lotteries, y should be considered better than x. *Ex post* prioritarianism, however, assigns priority to the worse-off in each outcome, and let us assume that we consider a version of *ex post* prioritarianism that considers the loss of person 2 in the first possible outcome in y as outweighing the gain of person 1 in this outcome in a comparison with x (and vice versa for the second possible outcome in y); that is, the loss from 10 to 5 outweighs the gain from 10 to 16.

[23] For a critical view on this literature, see Hausman (forthcoming).

In this case, *ex post* prioritarianism will conclude that both the possible outcomes in y are worse than the certain outcome in x, and consequently y is considered as worse than x. This violates the principle of personal good, which states that y should be considered better than x since everyone is better off in y than in x. Similar examples can be constructed for all other versions of *ex post* prioritarianism, and consequently we have a conflict between *ex post* prioritarianism and the principle of personal good. On the basis of this kind of analysis, Rabinowicz (2001, 2002) concludes that prioritarians should accept the principle of personal good when defined on outcomes, but not when defined on lotteries (as in the example above).

McCarthy (2006, 2007) rejects *ex post* prioritarianism, and argues that the more plausible position is *ex ante* prioritarianism. To illustrate this view, consider another lottery z also containing two equi-probable outcomes $z = ([10, 0], [0, 32])$. The expected goodness of person 1 and person 2 are 5 and 16 in z, and these are the numbers that the *ex ante* prioritarian relies on when evaluating the alternatives. Hence, it follows immediately that *ex ante* prioritarianism satisfies the principle of personal good defined on lotteries. Let us now compare x and z. The expected goodness of person 1 is greater in x than z, whereas the expected goodness of person 2 is greater in z than x. So the principle of personal good is silent in this case, since we have a distributive conflict in terms of expected goodness of each individual. In terms of total expected goodness, z is better than x, but an *ex ante* prioritarian may still argue that x is better than z since he wants to give priority to the worse-off. The loss in expected goodness of person 1 may therefore outweigh the gain in expected goodness of person 2 for an *ex ante* prioritarian.

The problem facing *ex ante* prioritarianism is that it does not satisfy the strong independence condition of expected utility theory. To illustrate, consider the famous example of Diamond (1967), where we compare the two lotteries $x = ([1, 0], [1, 0])$ and $y = ([0, 1], [1, 0])$, with equi-probable outcomes. Given that the second possible outcome is the same in the two lotteries, the strong independence condition requires that we can disregard this part when determining the ranking of x and y. If we consider only the first possible outcome in both alternatives, we see that it follows straightforwardly from anonymity that we have to consider x and y as equally good. An *ex ante* prioritarian, however, would disagree. The relevant comparison for an *ex ante* prioritarian would be that y gives both persons 0.5 in expected goodness, whereas x gives 1 and 0 in expected goodness to person 1 and person 2. Given that an *ex ante* prioritarian assigns priority to the worse-off in terms of expected goodness, he would conclude that y is better than x (which is in line with the conclusion of Diamond, but the reasoning is different). Hence, the *ex ante* prioritarian would violate the strong independence axiom.

The question is then whether there is a reason for an *ex ante* prioritarian to satisfy the strong independence axiom. McCarthy (2006, 2007) thinks so, and thus believes that this kind of example shows that *ex ante* prioritarianism is inconsistent.

This is not obvious to me. In addition to the general normative arguments against independence (see Anand 1993 and Ch. 5 above), an *ex ante* prioritarian cares about the level of expected goodness when assigning priority, and thus should not be interested in disregarding some of the outcomes when evaluating lotteries. Interestingly, as pointed out by McCarthy (2006), a characterization of *ex ante* prioritarianism can be provided along the same lines as the classical characterization of utilitarianism by Harsanyi if we drop the strong independence condition and add the requirement that we should consider y as better than x in the example above (and more generally in all situations of this kind).

17.6 Concluding Remarks

Egalitarian and prioritarian reasoning are discussed beyond the framework covered in this chapter. There is, for example, a substantial literature discussing how a concern for equality can be combined with holding people responsible for their choices, where the basic idea is that inequalities reflecting differences in choices may be fair, but not inequalities due to factors beyond individual control. A main result in this literature is due to Bossert (1995) and Bossert and Fleurbaey (1996), who show that it is impossible to combine egalitarianism with a very common conception of individual responsibility. Based on this impossibility result, there has been extensive discussion about how to reformulate the egalitarian view or the idea of responsibility in order to find a consistent responsibility-sensitive egalitarian approach to distributive justice (see Fleurbaey 2007 for an overview of this literature).

Moreno-Ternero and Roemer (2006) present an interesting version of prioritarian reasoning, which takes the weak equity condition of Sen (1973) as its starting point. They consider a resource allocation problem where people differ in their ability to transform resources into valuable outcomes, and in line with the weak equity condition of Sen they propose a priority condition that states that no individual should dominate another individual in both resources and outcomes. They show that this condition, together with some other rather mild conditions, characterizes a class of resource allocation mechanisms that equalize across individuals on the basis of some index of outcome and resources, which represents a compromise view between equalizing resources or equalizing outcomes. This characterization result may be seen as providing some justification for the common practice of working with indices that combine a concern for resources and outcomes—for example, the United Nations Development Programme's human development indicator.

Egalitarian and prioritarian reasoning play an important role in political life, both explicitly in justifying or rejecting different policies and implicitly in the

various welfare indices that are put forward in the debate. This should motivate further analysis of how to understand the value of equality and the priority to the worse-off, and, importantly, should also motivate more effort in making the lessons from this important literature transparent for the policymakers.

References

Anand, Paul (1993). *Foundations of Rational Choice under Risk*. Oxford: Clarendon Press.

Anderson, Elisabeth (1999). What is the Point of Equality? *Ethics*, 109, 287–337.

Arneson, Richard J. (1999). Egalitarianism and Responsibility. *Journal of Ethics*, 3, 225–47.

—— (2000). Luck Egalitarianism and Prioritarianism. *Ethics*, 100, 339–49.

Atkinson, Anthony B. (1970). On the Measurement of Inequality. *Journal of Economic Theory*, 2, 244–63.

Barry, Brian (1989). *Theories of Justice.* Berkeley: University of California Press.

Blackorby, Charles, and Donaldson, David (1978). Measures of Relative Inequality and their Meaning in Terms of Social Welfare. *Journal of Economic Theory*, 18, 59–80.

—— —— (1980). A Theoretical Treatment of Indices of Absolute Inequality. *International Economic Review*, 21, 107–36.

—— —— and Weymark, John A. (1984). Social Choice with Interpersonal Utility Comparisons: A Diagrammatic Introduction. *International Economic Review*, 25, 327–56.

Bosmans, Kristof (2006). Measuring Economic Inequality and Inequality Aversion Doctoral dissertation, Katholieke Universiteit Leuven.

—— (2007*a*). Comparing Degrees of Inequality Aversion. *Social Choice and Welfare*, 29, 405–28.

—— (2007*b*). Extreme Inequality Aversion without Separability. *Economic Theory*, 32, 589–94.

Bossert, Wolfgang (1995). Redistribution Mechanisms Based on Individual Characteristics. *Mathematical Social Sciences*, 29, 1–17.

—— and Fleurbaey, Marc (1996). Redistribution and Compensation, *Social Choice and Welfare*, 13, 343–55.

—— and Pfingsten, Andreas (1990). Intermediate Inequality: Concepts, Indices and Welfare Implications. *Mathematical Social Sciences*, 19, 117–34.

Bossert, Walter, and Weymark, John A. (1999). Utility in Social Choice. In Salvador Barbera, Peter Hammond, and Christian Seidl (eds.), *Handbook of Utility Theory*, 7–84. Dordrecht: Kluwer.

Broome, John (1991). *Weighing Goods.* Oxford: Blackwell.

—— (2004). *Weighing Lives.* New York: Oxford University Press.

—— (forthcoming). Equality versus Priority: A Useful Distinction. In C. Murray and D. Wikler (eds.), *Fairness and Goodness in Health*. Geneva: WHO.

Brown, Campell (2003). Giving Up Levelling Down. *Economics and Philosophy*, 19/1, 111–34.

Clayton, Matthew, and Williams, Andrew (eds.) (2000). *The Ideal of Equality.* Oxford: Oxford University Press.

Cohen, G. A. (1989). On the Currency of Egalitarian Justice. *Ethics*, 99, 906–44.
Cowell, F. A. (2000). Measurement of Inequality. In Anthony Atkinson and Francois Bourguignon (eds.), *Handbook of Income Distribution.* 000–00. Amsterdam: North-Holland.
Crisp, Roger (2000). Equality, Priority, and Compassion. Typescript, Oxford University.
Dalton, Hugh (1920). The Measurement of the Inequality of Incomes. *Economic Journal*, 30, 348–61.
Dasgupta, Partha, Sen, Amartya, and Starrett, David A. (1973). Notes on the Measurement of Inequality. *Journal of Economic Theory*, 6, 180–7.
Diamond, Peter (1967). Cardinal Welfare, Individual Ethics, and Interpersonal Comparisons of Utility: Comment. *Journal of Political Economy*, 75, 765–6.
Dworkin, Ronald (1981). What is Equality? Pts. I and II. *Philosophy and Public Affairs*, 10, 283–345.
Fleurbaey, Marc (2007). Fairness, Responsibility and Welfare. Mimeo.
—— (forthcoming). Equality versus Priority: How Relevant is the Distinction? In C. Murray and D. Wikler (eds.), *Fairness and Goodness in Health*. Geneva: WHO.
Frankfurt, Harry (1987). Equality as a Moral Ideal. *Ethics*, 98, 21–43.
Goodin, Robert E. (1987). Egalitarianism, Fetishistic and Otherwise. *Ethics*, 98, 44–9.
Hammond, Peter (1976). Equity, Arrow's Condition, and Rawls' Difference Principle. *Econometrica*, 44, 793–804.
—— (1979). Equality in Two-Person Situations—Some Consequences. *Econometrica*, 47, 1127–35.
Harsanyi, John (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility. *Journal of Political Economy*, 63, 309–21.
—— (1975). Nonlinear Social Welfare Functions: Do Welfare Economics have a Special Exemption from Bayesian Rationality? *Theory and Decision*, 6, 311–32.
Hausman, Daniel M. (forthcoming). Equality versus Priority: A Badly Misleading Distinction. In C. Murray and D. Wikler (eds.), *Fairness and Goodness in Health*. Geneva: WHO.
Holtug, Nils (1998). Egalitarianism and the Levelling Down Objection. *Analysis*, 58, 166–74.
—— and Lippert-Rasmussen, Kasper (eds.) (2006). *Egalitarianism: New Essays on the Nature and Value of Equality*. Oxford: Oxford University Press.
Kagan, Shelly (1988). The Additive Fallacy. *Ethics*, 99, 5–31.
Kamm, Frances (1993). *Morality, Mortality*, i. Oxford: Oxford University Press.
—— (1996). *Morality, Mortality*, ii. Oxford: Oxford University Press.
Klint Jensen, Karsten (2003). What is the Difference between (Moderate) Egalitarianism and Prioritarianism? *Economics and Philosophy*, 19/1, 89–109.
Kolm, Serge-Christophe (1976*a*). Unequal Inequalities I. *Journal of Economic Theory*, 12, 416–42.
—— (1976*b*). Unequal Inequalities II. *Journal of Economic Theory*, 13, 82–111.
—— (1996). *Modern Theories of Justice.* Cambridge, MA: MIT Press.
Kymlicka, Will (1988). Rawls on Teleology and Deontology. *Philosophy and Public Affairs*, 17, 173–90.
Lambert, Peter J. (1993). *The Distribution and Redistribution of Income.* Manchester: Manchester University Press.
Lemmi, Archille, and Betti, Gianni (eds.) (2006). *The Fuzzy Set Approach to Multidimensional Poverty Measurement*. Dordrecht: Kluwer.
Mason, Andrew (ed.) (1998). *Ideals of Equality.* Oxford: Blackwell.

McCarthy, David (2006). Utilitarianism and Prioritarianism. *Economics and Philosophy*, 22/3, 1–29.
—— (2007). Utilitarianism and Prioritarianism II. Mimeo.
McKerlie, Dennis (1994). Equality and Priority. *Utilitas*, 6, 25–42.
Moreno-Ternero, Juan D., and Roemer, John E. (2006). Impartiality, Priority, and Solidarity in the Theory of Justice. *Econometrica*, 74/5, 1419–27.
Nagel, Thomas (1979). *Mortal Questions.* Cambridge: Cambridge University Press.
—— (1991). *Equality and Partiality.* Oxford: Oxford University Press.
Parfit, Derek (1995). Equality or Priority. Lindley Lecture, University of Kansas.
Peterson, Martin, and Hansson, Sven Ove (2005). Equality and Priority. *Utilitas*, 17, 299–309.
Pojman, Louis P., and Westmoreland, Robert (eds.) (1997). *Equality: Selected Readings.* Oxford: Oxford University Press.
Rabinowicz, Wlodek (2001). Prioritarianism and Uncertainty: On the Interpersonal Addition Theorem and the Priority View. In D. Egonsson, J. Josefsson, B. Peterson, and T. Rønnow-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values*, 139–65. Aldershot: Ashgate.
—— (2002). Prioritarianism for Prospects. *Utilitas*, 14, 2–21.
Rawls, John (1971). *A Theory of Justice.* Cambridge, MA: Harvard University Press.
—— (1993). *Political Liberalism.* New York: Columbia University Press.
Raz, Joseph (1986). *The Morality of Freedom.* Oxford: Clarendon Press.
Roemer, John (1996). *Theories of Distributive Justice.* Cambridge, MA: Harvard University Press.
Rosenberg, Alexander (1995). Equality, Sufficiency and Opportunity in the Just Society. *Social Philosophy and Policy*, 12/2, 54–71.
Scanlon, Thomas (1993). Value, Desire, and Quality of Life. In Martha C. Nussbaum and Amartya Sen (eds.), *The Quality of Life*, 185–200. Oxford: Clarendon Press.
—— (1998). *What We Owe Each Other.* Cambridge, MA: Harvard University Press.
—— (2000). The Diversity of Objections to Inequality. In Clayton and Williams (2000), 41–59.
Sen, Amartya (1973). *On Economic Inequality.* Oxford: Clarendon Press.
—— (1980). Equality of What. In S. McMurrin (ed.), *Tanner Lectures on Human Values*, 195–220. Cambridge: Cambridge University Press.
—— (1983). Poor, Relatively Speaking. *Oxford Economic Papers*, 35, 153–69.
—— (1992). *Inequality Reexamined.* Cambridge, MA: Harvard University Press.
—— and Foster, James (1997). *On Economic Inequality*, expanded edn. Oxford: Clarendon Press.
Shorrocks, Anthony (1980). The Class of Additively Decomposable Inequality Measures. *Econometrica*, 48, 613–25.
Temkin, Larry (1993). *Inequality.* Oxford: Oxford University Press.
—— (2000). Equality, Priority, and the Levelling Down Objection. In Clayton and Williams (2000), 126–61.
—— (2003). Equality, Priority, or What? *Economics and Philosophy*, 19/1, 61–87.
Tungodden, Bertil (1999). The Distribution Problem and Rawlsian Reasoning. *Social Choice and Welfare*, 16, 599–614.
—— (2000*a*). Egalitarianism: Is Leximin the Only Option? *Economics and Philosophy*, 16, 229–45.

Tungodden, Bertil (2000*b*). Hammond Equity: A Generalization. Discussion Paper, Norwegian School of Economic and Business Administration.

—— (2003). The Value of Equality. *Economics and Philosophy*, 19/1, 1–44.

—— and Vallentyne, Peter (2005). On the Possibility of Paretian Egalitarianism. *Journal of Philosophy*, 102, 126–54.

—— —— (2006). Who Are the Least Advantaged? In Holtug and Lippert-Rasmussen, (2006), 174–95.

Vallentyne, Peter (2000). Equality, Efficiency, and the Priority of the Worse Off. *Economics and Philosophy*, 16, 1–19.

van Parijs, Philippe (2001). Difference Principles. In Samuel Freeman (ed.), *The Cambridge Companion to John Rawls*, 200–40. Cambridge: Cambridge University Press.

Weirich, Paul (1983). Utility Tempered with Equality. *Nous*, 17, 423–39.

Wolff, Jonathan (2000). Levelling Down. In K. Dowding, J. Hughes, and H. Margetts (eds.), *Challenges to Democracy: The PSA Yearbook 2000*, 18–32. London: Macmillan.

CHAPTER 18

RAWLSIAN JUSTICE

FABIENNE PETER

18.1 Introduction

At the outset of *Political Liberalism*, Rawls (1993, p. 4) asks: "[H]ow is it possible for there to exist over time a just and stable society of free and equal citizens, who remain profoundly divided by reasonable religious, philosophical, and moral doctrines?" In other words, how can we think about justice for a society marked by (reasonable) value pluralism—by deep conflicts among individual preferences about how society should be organized?[1] Classical utilitarianism tries to avoid this problem by sacrificing an independent idea of distributive justice. It treats individual utility as the ultimate good and identifies the right social arrangement as the one that maximizes an aggregate of individual utility. Rawls's theory of justice builds on the social contract tradition to offer an alternative to utilitarianism. His "political conception" of justice rests on fundamental values that he identifies as implicit in democratic societies. Rawls argues that they offer a basis for constructing principles of justice which can be accepted by the members of such societies. Rawls's interpretation of the social contract allows him to address questions of justice directly, not via social welfare, as in utilitarianism, and indeed singles

I have received helpful comments from Paul Anand and Serena Olsaretti—many thanks to them.

[1] I shall discuss the exact meaning of "reasonable" below. For the moment, take reasonable pluralism as deep conflicts between individual preferences that are not due to false beliefs, lack of information, lack of reflection, narrow self-interest, etc.

out justice—not maximum welfare or efficiency—as "the first virtue of social institutions".[2]

Rawls's theory of justice has been enormously influential, in philosophy and beyond. It has, from the start, attracted much interest from economists. An important reason for this interest lies, very simply, in the impressive account that Rawls gives in his articles and books. There are, however, also a number of reasons specific to economic theory. First, in the aftermath of Arrow's impossibility result, welfare economists and social choice theorists struggled with the problem of how to accommodate considerations of justice in their theoretical frameworks. Rawls's theory of justice as fairness offered hope for all those economists not content with the predominance of the criterion of efficiency and not ready to give up on justice. Second, in *A Theory of Justice*, Rawls attempted to justify the principles of justice as fairness by reference to individual rational choice. This attempt attracted a lot of criticism from economists (e.g. Harsanyi 1975), and was eventually abandoned by Rawls in favor of an account that stresses the differences between being rational and being reasonable. Even if this episode has created some confusion, Rawls generally tried to make his theory of justice accessible to economists, and many of his ideas have had a lasting effect on economic theorizing. In this chapter I shall focus on Rawls's own presentation of his theory of justice and on how his theory has been received in normative economics.[3]

18.2 Justice as Fairness: The Bare Bones

Let me start with a brief account of Rawls's theory of justice. I shall refer to Rawls's original presentation of justice as fairness in his 1971 book *A Theory of Justice* as well as to views he put forward in later articles (see Rawls 1999) and books (especially in *Political Liberalism* (1993) and in *Justice as Fairness: A Restatement* (2001)). Rawls has revised some of his views over time, and I shall give an account that is in line with the revised interpretation of justice as fairness.[4]

[2] The passage continues: "[L]aws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust. Each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override" (Rawls 1971, p. 3).

[3] Outside of normative economics, there is another development in economic theory which picks up on Rawlsian themes. Evolutionary game theory studies norms and mechanisms of coordination and cooperation and, as part of that, also norms of distributive justice. Ken Binmore, in his recent book *Natural Justice* (2005), argues that some of Rawls's ideas are corroborated by the findings of evolutionary game theory. Unfortunately, I lack space to discuss Binmore's proposal here, but see Peter (2006).

[4] For the sake of simplicity, I shall abstain from discussing how his ideas have developed over time, unless it is directly relevant to the issue that I am discussing.

18.2.1 Some Fundamental Ideas

If a society is characterized by irreducible value pluralism, there is no single moral or religious authority to which people can agree to resort to resolve distributional conflicts. Rawls thus takes it as a starting point that a theory of justice cannot be based on a "comprehensive" moral or religious doctrine.[5] In his attempt to reconcile reasoning about justice and value pluralism, Rawls turns to political values. He grounds the theory of justice as fairness on ideas which he sees as explications of views that are characteristic of the political culture of democratic societies and as having the potential to be widely shared among citizens of democratic societies. The most important ones are the idea of "society as a fair system of cooperation" and the idea of "citizens as free and equal persons".

Let me start with the latter. It expresses a political, not a psychological or metaphysical, conception of the person (Rawls 1993, 29–35; 2001: §7). Its function is both to capture the fact that people have diverse interests and to explain how they can reach an agreement in matters of justice. According to this conception, persons have two fundamental moral powers. These are the "capacity for a conception of the good", on the one hand, and the "capacity for a sense of justice", on the other (Rawls 1993, p. 34). Rawls (1993, p. 302) defines them as follows:

> [T]he capacity for a sense of justice is the capacity to understand, to apply and normally to be moved by an effective desire to act from (and not merely in accordance with) the principles of justice as the fair terms of social cooperation. The capacity for a conception of the good is the capacity to form, to revise, and rationally to pursue such a conception, that is, a conception of what we regard for us as a worthwhile human life.

By virtue of the capacity for a conception of the good, persons know what is to their advantage and are able to act rationally. Rawls works with a broader understanding of rationality than economic theory, as it is not limited to maximizing a consistent set of preferences. He adopts the Kantian conception of practical reason, which includes the capacity to deliberate about ends, to evaluate, prioritize, and—if necessary—revise them, in addition to the capacity to choose the best means to reach a given end (Rawls 1993, p. 50). Moreover, persons are seen not only as potentially rational, but also as potentially reasonable. Being reasonable is defined in terms of the capacity for a sense of justice, and this capacity refers to the second fundamental idea, that of society as a fair system of cooperation—and I shall discuss the capacity to be reasonable in this context. For the moment, just note that there is a difference between the reasonable and the rational. Conceptions of the good are called reasonable if they are in accord with the requirements of the reasonable.

[5] A comprehensive conception "includes conceptions of what is of value in human life, and ideals of personal character, as well as ideals of friendship and of familial and associational relationships, and much else that is to inform our conduct" (Rawls 1993, p. 13).

Justice as fairness views citizens as free and equal with regard to these two capacities. All citizens are assumed to hold these two principal moral powers, and it is in this respect that they are equal. They are free insofar as they can develop and pursue their own reasonable conception of the good.

A second fundamental idea of justice as fairness—that of society as a fair system of cooperation—is closely linked to the idea of citizens as free and equal persons. To understand Rawls's idea of cooperation, compare it with utilitarianism first. In the utilitarian view, the conception of the person is not a political, but a psychological, one—it uses individual utility both to represent what individuals value and to explain their (rational) actions. Taking this as the starting point, the goodness of individual states is assessed exclusively in terms of the utility that individuals derive from it—however utility is defined. Society is seen as a conglomeration of utility-maximizing individuals, and cooperative arrangements should aim at maximal aggregate utility.

In contrast to utilitarianism, Rawls's theory of justice relies on a distinction between what is rational for individuals and what is reasonable. Persons are reasonable insofar as they recognize that, though they have good reasons to hold their own conception of the good, there are good reasons for other citizens to hold different views. Reasonable citizens accept that their society will always contain a plurality of conceptions of the good. They also accept what Rawls calls the "burdens of judgment"—a list of considerations as to why reasonable disagreement over conceptions of the good is likely to persist (Rawls 1993, p. 54). In addition, by virtue of their sense of justice, persons are assumed to be willing to propose fair terms of cooperation, which guarantee fair prospects for all to pursue their respective rational advantage. The persons recognize, thanks to their capacity for a sense of justice, that the rational pursuit of their own advantage needs to be made compatible with the possibility for others to pursue their conception of the good, provided those conceptions are reasonable too. They are willing to refrain from imposing their own conception of the good upon others and will want principles of justice which are compatible with the fact of reasonable pluralism—an irreducible pluralism of reasonable comprehensive conceptions of the good.

Utilitarianism reduces the reasonable—reasons that refer to the regulation of the individual pursuit of a good life through cooperative arrangements—to the rational—reasons that refer to the individual pursuit of a good life. In the theory of justice as fairness, by contrast, the reasonable is an independent idea. Reasonable persons in Rawls's sense "are not moved by the general good as such but desire for its own sake a world in which they, as free and equal, can cooperate with others on terms all can accept" (Rawls 1993, p. 50). His idea of cooperation thus entails an idea of fair cooperation; it is based on reciprocity. Reciprocity refers to generally recognized rules which secure everybody an adequate share of the benefits produced through cooperation. As such, Rawls's idea of fair cooperation has to be distinguished from an idea of mutual advantage, which demands that everyone gains from cooperation. A conception of justice that specifies fair terms

of cooperation respects and insures equal liberties for the citizens to develop and pursue their reasonable conceptions of the good.

According to Rawls, this idea of cooperation not only distinguishes justice as fairness from utilitarianism and mutual advantage theories. There is also an important difference between justice as fairness and the libertarian approach here. Libertarianism tends to view cooperative schemes as voluntary associations—and deflects demands for more equality with reference to the voluntariness of such schemes. Justice as fairness, by contrast, treats membership in society as involuntary—given by birth and such that exit is, in what concerns justice, not an option. Citizens have, *qua* membership in this cooperation, a right to the benefits produced.[6]

The two—related—ideas of society as a fair system of cooperation and of citizens as free and equal persons form the starting point for the theory of justice as fairness. A further fundamental idea specifies the domain of justice. Reasonable pluralism makes it likely that the citizens will rarely—if ever—agree on the moral value of alternative social states.[7] Taking this into account, justice as fairness is conceived of as having a limited domain. The fair terms of cooperation apply to what Rawls calls the "basic structure" of society, and only to that. The basic structure comprises "society's main political, social, and economic institutions, and how they fit together into one unified system of social cooperation from one generation to the next" (Rawls 1993, p. 11).[8] Rawls singles out the basic structure, because inequalities that have their origin there have the most profound impact on the prospects of the individuals in society.[9] Imposition of rules of fairness on the basic structure is an attempt to correct these fundamental inequalities as far as possible and to establish fair conditions of social cooperation. The intuition is that if the basic structure is just, so is the outcome generated by the social and economic processes it specifies and embeds. Thus, being confined to a limited domain distinguishes justice as fairness both from utilitarianism and from those contractualist moral theories which are intended to apply to all questions of social evaluation (e.g. Scanlon 1998). Justice as fairness proposes principles for how to assess society's main institutions, and only them.

These fundamental ideas are in accordance with what Rawls calls a political conception of justice. He defines such a conception in the following way (2001, p. 26): (i) "it is worked out . . . for the basic structure of a democratic society"; (ii) it

[6] For an influential statement of the libertarian doctrine, see Nozick (1974). On this contrast between justice as fairness and libertarianism, see Rawls (1993, pp. 264 f.).

[7] I use the term "social state" in the sense of social choice theory—a full description of all the economic, political, and social circumstances (Arrow 1963).

[8] Rawls discusses the idea of the basic structure as the first subject of justice extensively in Rawls (1971, §2; 1993, VII; 2001, IV).

[9] "The basic structure is the primary subject of justice because its effects are so profound and present from the start. The intuitive notion here is that this structure contains various social positions and that men born into different positions have different expectations of life determined, in part, by the political system as well as by economic and social circumstances. In this way the institutions of society favor certain starting places over others. These are especially deep inequalities" (Rawls 1971, p. 7).

"does not presuppose accepting any particular comprehensive doctrine"; and (iii) it "is formulated so far as possible solely in terms of fundamental ideas familiar from, or implicit in, the public political culture of a democratic society". Because a political conception of justice does not rest on a comprehensive moral or religious doctrine but builds on fundamental political values instead, it circumvents the problem of value pluralism. That such a theory of justice is restricted in scope—that it does not apply to all moral questions, but only to the problem of the justice of the basic structure of society—is for Rawls a small price to pay.

18.2.2 The Original Position and the Idea of a Public Conception of Justice

In *A Theory of Justice* Rawls famously—but in places misleadingly—condensed these fundamental ideas into the thought-experiment of the original position. The thought-experiment interprets the question that justice as fairness tries to answer as: what principles of justice would free and equal persons choose to regulate the main terms of their cooperation? It thus takes into account the fact of reasonable pluralism and demands that what constitutes the fairness of the basic structure be determined by what persons can agree to. But not any agreement will do. It must respect the idea of equal liberty and the restrictions that the reasonable imposes on the rational. The following hypothetical situation—the original position—is designed to represent such fair conditions. Rawls asks persons to abstract from their actual preferences about their individual advantage and their present position in society. They should do so by imagining that they deliberate about the principles of justice that should apply to the basic structure of society behind a veil of ignorance. The veil of ignorance separates persons in the original position from knowledge of their particular conception of the good, their specific position in society, or their talents and abilities. Only the most general knowledge about society, such as the basic political, economic, sociological, and psychological principles, is allowed to seep through.[10] The veil of ignorance insures that the justification of the principles of justice will not be affected by arguments that are related to defending a particular position in society. It also eliminates bargaining (Rawls 1993, p. 23).

The argument from the original position is concerned, first of all, with the justification of principles of justice.[11] The original position represents the conditions and constraints under which persons should deliberate about adequate principles of justice. As such, it serves the role of a selection device. It is designed to facilitate the selection of principles of justice from a list, not to derive principles of justice directly from it (1971, §21; 2001, p. 83).

[10] See Rawls (1971, §24).

[11] There is a second aspect to the argument from the original position, which I will not be able to discuss here. This concerns the question of stability; see Rawls (1993, lecture IV).

The argument from the original position has sometimes been misinterpreted—one might argue partly because of Rawls's own misleading original characterization of it. What has created problems has been the question of how the persons in the original position deliberate about and choose the principles of justice. In *A Theory of Justice*, Rawls cast justice as fairness as part of rational choice theory (1971, pp. 16 and 47). This view would render the—Kantian—distinction between the reasonable and the rational unintelligible. In his later writing, Rawls treats this claim about the link between rational choice theory and justice as fairness as a mistake (e.g. Rawls 2001, p. 82 n. 2) and insists on the importance of the idea of the reasonable in justifying principles of justice. According to the revised view, we ought to imagine the persons in the original position as follows. They are rational—they have the capacity to formulate, revise, and efficiently pursue a conception of the good. They are also mutually disinterested, in the sense that they are not motivated by feelings of envy or a desire to have power over others. In addition to being rational in this sense, they are reasonable; that is, they are willing both to propose fair terms of cooperation and to act from such principles.

The justification for a theory of justice, according to Rawls, ought to satisfy a publicity constraint. This constraint entails the following: that "everyone accepts, and knows that everyone else accepts", the same principles of justice; secondly, that "society's basic structure . . . is publicly known, or with good reason believed, to satisfy those principles"; and thirdly, that "citizens have a normally effective sense of justice"; that is, they can understand and act from the principles of justice (Rawls 2001, pp. 8 f.). He calls the ideal of a society that is "effectively regulated by a public conception of justice" a "well-ordered society" (2001, p. 8). In a well-ordered society, "the public conception of justice provides a mutually recognized point of view from which citizens can adjudicate their claims of political right on their political institutions or against one another" (Rawls 2001, p. 9).

18.2.3 The Principles of Justice

Which public principles of justice would citizens who think of themselves as free and equal, and who think of their society as a fair system of cooperation, choose to regulate the basic structure of society? Rawls (2001, pp. 42 f.) argues that they could agree on the following two principles of justice:

(a) Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all; and

(b) Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle).

The first principle guarantees the citizens' equality with respect to a list of basic liberties and rights. These basic liberties and rights are the following: political liberties (i.e. the right to vote and to be eligible for public office) and freedom of speech and assembly, liberty of conscience and freedom of thought, freedom of the person and the right to hold personal property, and the freedom from arbitrary arrest and seizure (Rawls 1971, p. 61). It is essential that the first principle is interpreted with regard to such a list of liberties and not, as *A Theory of Justice* has suggested, as a principle of "basic liberty". Justice as fairness assigns special value not to freedom as such, only to a specific list of rights and liberties.

The political liberties are part of that list. Rawls emphasizes that the first principle must include a guarantee that everybody has a fair chance to participate in the political process. This requirement is discussed in *A Theory of Justice* under the heading of the principle of participation (cf. Rawls 1971, §§36 and 37). In *Political Liberalism*, Rawls included it in the first principle of justice in the form of a guarantee of the fair value of the political liberties. The fair value of the political liberties requires that "citizens similarly gifted and motivated have roughly an equal chance of influencing the government's policy and of attaining positions of authority irrespective of their economic and social class" (Rawls 1993, p. 358). The first principle thus insures that, in a well-ordered society, members of all social groups are able to participate in the political process on an equal basis.

The second principle is divided into two parts. The first part is called the principle of fair equality of opportunity, and the second the difference principle. Fair equality of opportunity contrasts with merely formal equality. Formal equality of opportunity is satisfied if there are no discriminating legal barriers that bar some groups in society from access to social institutions. Fair equality of opportunity is more demanding. It requires that no factual social barriers exist which make access dependent upon social and economic position. According to Rawls (1971, p. 73):

> [T]hose who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system, that is, irrespective of the income class into which they are born. In all sectors of society there should be roughly equal prospects of culture and achievement for everyone similarly motivated and endowed. The expectations of those with the same abilities and aspirations should not be affected by their social class.

Rawls argues that, from the point of view of the original position, it is reasonable to want to impose the principles of equality of basic liberties and rights and of fair equality of opportunity on the basic structure. Prevented from knowledge of one's social position, abilities, talents, and preferences, citizens who view themselves as free and equal have no reason to depart in these fundamental matters from the position of equality that characterizes the original position. Looking at it the other way around, the setup of the original position provides a rationale for selecting

egalitarian principles of justice that aim at dissociating the prospects that citizens face from their social endowments.

In spite of this egalitarian commitment, Rawls also emphasizes that some economic and social inequalities are justifiable. The question is which, and the difference principle, the second part of the second principle of justice, answers this question. This well-known principle demands that the basic structure of society should be organized so as to admit social and economic inequalities to the extent that they are beneficial to the least advantaged group. It allows deviations from the situation of equal liberty that characterizes the original position insofar as these deviations allow for a more productive scheme of cooperation that leaves the least advantaged group better off than it would be were the basic structure set up so as to preserve absolute equality. Given a range of possibilities for how to design the institutions of the basic structure, the difference principle says that it should be the institutional arrangement which maximizes the prospects of the least advantaged group. What the difference principle rules out is a basic structure which grants greater benefits for more advantaged groups while worsening the prospects of the least advantaged group.

The principles are ordered lexicographically. The first principle of justice has priority over the second, and the principle of fair equality of opportunity has priority over the difference principle.[12] This implies that the equality of basic liberties and rights, including the fair value of the political liberties, is not to be overridden by other considerations. Furthermore, the difference principle has to be interpreted as applying to an environment in which the lexicographically prior principles are already in place. The social and economic inequalities that the difference principle might justify are those that do not undermine the equality of basic rights and liberties and the principle of fair equality of opportunities. Finally, the lexicographic structure also implies that, in social evaluation, the principles of justice taken together have priority over other considerations—for example, over considerations of maximum welfare or efficiency.

18.2.4 Primary Goods

The two principles of justice require some kind of interpersonal comparisons to check whether the fair value of the political liberties and fair equality of opportunity hold, and in order to apply the difference principle. Justice as fairness hence demands an informational framework which allows for interpersonal comparisons. But given value pluralism, what basis could there be for interpersonal comparisons?

In economic theory, there is a tendency to deny the possibility of interpersonal comparisons because of the diversity of people's values. Arrow (1984, p. 160), for example, writes:

[12] See Rawls (1971, pp. 42 f. and 302 f.).

In a way that I cannot articulate well and am none too sure about defending, the autonomy of the individuals, an element of mutual incommensurability among people, seems denied by the possibility of interpersonal comparisons. No doubt it is some such feeling as this that has made me so reluctant to shift from pure ordinalism, despite my desire to seek a basis for a theory of justice.

Arrow recognizes that his skepticism about interpersonal comparisons makes it difficult to overcome the implications of his famous impossibility theorem and to formulate a theory of justice. What Rawls has to show, in reply, is that justice as fairness offers a theoretical approach to interpersonal comparisons that respects this "mutual incommensurability" between different conceptions of the good.

Rawls proposes the framework of primary goods, which does not have its basis in individual conceptions of the good. Primary goods are an account of general institutional features of the basic structure of society which affect the prospects of individuals, whatever their ideas of the good life are. Rawls counts the following as primary goods (Rawls 1993, p. 181):

— basic liberties and rights, given by a list
— freedom of movement and free choice of occupation against a background of diverse opportunities
— powers and prerogatives of offices and positions of responsibility in the political and economic institutions of the basic structure
— income and wealth
— the social bases of self-respect.

Unlike the preference framework used by economists, the primary goods framework is objective. The value of these goods is derived from fundamental political values of democratic societies, and not from individual ideas of the good. That is to say, it is an account of what the relevant goods are, on the basis of which it can be assessed whether the basic structure establishes fair cooperation and respects people as free and equals. Primary goods are not intended to be a surrogate for individual well-being, but a measure of people's access to basic institutions—of the institutional conditions for the realization of the two fundamental moral powers that persons have. Because primary goods are thus not determined by a comprehensive conception of the good, but rest on shared political values, the framework overcomes the problem of "mutual incommensurability" that Arrow mentions.

An individual's endowment with primary goods is measured by an index.[13] Although the construction of such an index is not easy, I want to leave aside for the moment the practical difficulties and assume that it is at least approximately possible.[14] Furthermore, the construction of an index is facilitated by the way in

[13] See Rawls (1971, §15; 1982).

[14] I shall come back to this problem below. On problems that may arise in the construction of an index, see Gibbard (1979); Hohm (1983); and Blair (1988). But see also Sen (1991), who argues

which the principles of justice shape the distribution of the different primary goods. By the first principle and the principle of fair equality of opportunity, the first two goods on the list must be distributed equally. Fair equality of opportunity also demands that access to powers and prerogatives of offices and to positions of responsibility in important political and economic institutions, the third primary good, be distributed equally. Holdings of such positions, in contrast, may be distributed unequally. Their distribution and the distribution of income and wealth are regulated by the difference principle.

The case of the last primary good on the list, entailing the social bases of self-respect, is more complicated.[15] Note, first, that the emphasis lies on the social bases of self-respect, and not on self-respect directly. It thus refers to the basic structure of society as well, and not, say, to an individual state of mind. Take a feudal society as an example of a society in which some groups are not granted the social bases for self-respect. This primary good is of a different kind from the others on the list. While the first four primary goods describe general means to develop and pursue a conception of the good, the social bases of self-respect insure that the citizens have the possibility to experience it as worthwhile to do so. In this regard they are important for the development of a sense of justice, the second moral power attributed to citizens. It is for these reasons that Rawls regards the social bases of self-respect as the most important primary good on the list. The distribution of this primary good is affected by the regulation of the basic structure of society. Since the social bases of self-respect are in part dependent on the social and economic status of the individuals, they will not be equalized under justice as fairness. Here the difference principle applies, requiring that social and economic inequalities do not undermine a minimal social basis of self-respect.

This raises the more general question of how the relationship between primary goods and the equality and freedom of the citizens is to be understood in Rawls's theory of justice. We saw that fairness requires that the basic structure must grant more than formal freedom to pursue individual ends. Justice as fairness includes some demand that the individuals can actually make use of their freedom. But to what extent? Rawls's distinction between liberty and the worth of liberty helps to clarify this issue.[16] The worth of liberty is measured by the index of primary goods that a citizen holds. Whereas the basic liberties are equalized by virtue of the first principle, the worth of these liberties is not always equal, since primary goods are not equalized in justice as fairness. For the political liberties, however, and only for them, Rawls includes a guarantee of the fair value. The idea is, in analogy to what fair equality of opportunity demands, that "citizens similarly gifted and motivated have roughly an equal chance of influencing the government's policy

that the technical details of constructing an index are less important than determining what kind of information should be included.

[15] On the social bases of self-respect, see especially Rawls (1971, §§67 and 82).

[16] See Rawls (1971, pp. 204 f.; 1993, pp. 324–31).

and of attaining positions of authority irrespective of their economic and social class".[17] The worth of the other basic liberties is allowed to be affected by social and economic inequalities. By not guaranteeing an equal value of all liberties, justice as fairness is not, however, indifferent to the worth of liberty, since social and economic inequalities are limited by the difference principle. The point of reference is an equal distribution of primary goods in the original position. Inequalities are then introduced in order to achieve a more elaborate organization of society. But the inequalities in the holdings of primary goods—and thus in the worth of liberties—are restricted by the requirement that they must be to the benefit of the least advantaged group. From the point of view of the original position, a social state with these inequalities is preferred to a state with equal distribution. In Rawls's words, what justice as fairness requires is that "the basic structure is to be arranged to maximize the worth to the least advantaged of the complete scheme of equal liberty shared by all. This defines the end of social justice" (Rawls 1971, p. 205). Inequalities in the worth of liberty for the citizens are justified because the combined impact of the principles of justice on the regulation of the basic structure of society is to insure a sufficient worth of liberty for the least advantaged group.

18.3 Rawlsian Justice in Normative Economics

Economists were quick to realize the significance of Rawls's *A Theory of Justice*, and there are many reviews of that book written by economists (e.g. Arrow 1973; Musgrave 1973; Alexander 1974; Harsanyi 1975). At first, the tendency was to discuss—and criticize—Rawls's ideas from the welfarist angle that dominated economic theory. Arrow, for example, writes that "Rawls . . . starting from the same premises [as Harsanyi and Vickrey], derives the statement that society should maximize min u_i [instead of the sum of utilities]" (1984, p. 102). This short statement is illustrative of how Rawls's theory of justice as fairness was interpreted early on by economists, for multiple reasons. First, it focuses on the difference principle and neglects its relation to the other elements of Rawls's principles of justice. Secondly, by linking Rawls directly to ideas independently put forward by Vickrey and Harsanyi, it reduces the original position to a hypothetical choice situation behind a veil of ignorance and ignores the way in which Rawls's thought-experiment of

[17] Rawls (1993, p. 358). The original statement of the two principles of justice in Rawls (1971) did not mention the particular role of the political liberties. But Rawls discusses this issue in *A Theory of Justice*, §§36 and 37.

the original position is an attempt to represent fundamental political values. And finally, it takes no notice of primary goods and interprets the difference principle in terms of the utility framework. Let me discuss this interpretation of Rawlsian justice and its problems in some detail.

18.3.1 Maximin and Welfarist Justice

The version of the difference principle that most economists are familiar with is best called the maximin, or the Rawlsian, social welfare function—to differentiate it from Rawls's own principle. Maximin is the social welfare function that identifies the social optimum as the social state of affairs in which the worst-off person enjoys the highest utility. It ranks alternative social states on the basis of how well-off the worst-off person is.[18]

Maximin contrasts, for example, with the utilitarian social welfare function, according to which the social optimum is the social state that maximizes the sum of individual utilities. Both, however, demand interpersonal comparisons of welfare; the maximin social welfare function relies on interpersonal comparability at the ordinal level. Given the negative stance taken by many economists on the interpersonal comparability of utility, the question has to be answered as to how such comparisons could be made. One answer for comparisons at the ordinal level invokes extended sympathy, where individuals are assumed to have preferences not just about their own position in different social states, but about the position of other individuals as well. Individual preferences are then over ordered pairs of the form (x, i), which stands for the position of individual i in social state x. Interpersonal comparisons take the form: individual i is better off in state x than individual j in state y.[19] Compared to the utilitarian social welfare function, which requires interpersonal comparisons at the cardinal level, maximin is perceived to be more parsimonious informationally and therefore less problematic by many economists (cf. Arrow 1984).

Both maximin and the utilitarian social welfare function are variants of the same—welfarist—view of justice. In the words of Charles Blackorby, Walter Bossert, and David Donaldson (2002, p. 545), this view "asserts that a just society is a good society: good for the individual people that comprise it. To implement such an approach to justice, the social good is identified and used to rank social alternatives. Of the alternatives that are feasible, given the constraints of human nature and

[18] More technically speaking, maximin says that a social state x is judged at least as good as another, y, if there is an individual k in y such that no position in state x is perceived to be worse than $k's$. This formulation is from Sen (1970*a*, pp. 156 ff.). Another popular formulation is in terms of a veto of the worst-off (e.g. Strasnick 1976). Tungodden (1999) has recently pointed out that the veto interpretation is misleading.

[19] On extended sympathy, see Strasnick (1976) and Arrow (1984). In fact, extended sympathy is a form of *intra*personal comparisons.

history, the best is identified with justice." The "social good", in this welfarist view, is an aggregate of individual utility. In other words, the only thing that matters about social states is how much utility individuals enjoy. Different theories of justice, different social welfare functions, specify different ways for aggregating individual utility. One difference between Rawls's and the welfarist theory of justice is thus that the latter is, while the former is not, specified in terms of individual utility. To my knowledge, it was Sen who first pointed out the parallel between Rawls's difference principle and the problem of social choice in his book *Collective Choice and Social Welfare* (Sen 1970*a*). But already there he noted that Rawls pursues a different aim: "[Rawls's] main interest is not so much in the ordering of social states, which is our concern, but with finding just institutions as opposed to unjust ones, which is a somewhat different problem" (Sen 1970*a*, p. 140). And in a later article he pointed out the differences between the informational frameworks used.[20]

I shall come back to this difference between primary goods and utility below. I first want to comment on further differences between maximin and Rawls's difference principle. Recall that the latter, being embedded in the conception of justice as fairness as a whole, is limited by two types of constraints. The first is the lexicographic ordering of the two principles of justice. The difference principle is meant to apply only to those social states which satisfy the first principle of justice as fairness and the principle of fair equality of opportunity. As a result, Rawls did not, as some have suggested (e.g. Harsanyi 1975), require giving absolute priority to the worst-off. The interests of others are protected by the principle of equal liberty and the principle of fair equality of opportunity. In addition, the difference principle applies only to the setting up of society's basic structure. Rawls did not defend this principle as a general principle for settling isolated problems of distributive justice. He did not, for example, suggest that a just educational system is one in which all the resources are spent on the improvement of the situation of the worst-off. The maximin principle also neglects how the difference principle is meant to apply to the dynamic aspects of the basic structure. Justice as fairness is grounded on a procedural interpretation of justice: it compares not alternative distributions as such, but alternative ways of how the basic structure generates distributions.[21] Rawls (1971, p. 88) writes:

[20] "Rawls' (1971) 'difference principle' in his theory of justice, in which a person's disadvantage is judged in terms of his access to "primary social goods", and not in terms of utility as such (as in the apocryphal version popular among economists), will clash violently with welfarism" (Sen 1979, p. 548).

[21] On procedural justice, see Rawls (1971, pp. 84 ff.). Rawls discriminates between different types of procedural justice. One distinction is between "perfect" and "imperfect" procedural justice. Perfect procedural justice applies if the desired outcome is known and if a procedure exists that can bring about this outcome. Imperfect procedural justice is used when the desired outcome is known, but no procedure exists that can bring it about with absolute certainty. The procedure has then to be set up so that it promotes this goal with maximum likelihood. *Pure* procedural justice, in contrast, applies if the desired outcome is not known. This is the type of procedural justice that justice as fairness uses.

A distribution cannot be judged in isolation from the system of which it is the outcome or from what individuals have done in good faith in the light of established expectations. If it is asked in the abstract whether one distribution of a given stock of things to definite individuals with known desires and preferences is better than another, then there is simply no answer to this question.

Procedural justice focuses on the constraints which set the framework for acceptable outcomes, and not on the outcomes themselves. It focuses on the background against which these transactions take place. The claim is that if the basic structure of society is just (as assessed against the principles of justice primarily chosen), the outcomes it generates are just too. The two principles of justice together serve as a selection device for alternative institutional structures. The difference principle is thus meant not to evaluate alternative distributions within a particular basic structure, but to help make a selection from alternative sets of institutions to regulate the basic structure. If the maximin social welfare function is isolated from the context of justice as fairness and applied to rank alternative distributions, this is likely to lead to very different recommendations from Rawls's own, and Rawls should not be blamed if many of these recommendations are highly implausible.

18.3.2 Decision-Making Behind a Veil of Ignorance

Arrow and many others draw an analogy between Rawls's and John Harsanyi's (1955, 1975) use of the idea of the veil of ignorance. Harsanyi, even before Rawls published his first article on justice as fairness, invoked a hypothetical situation of uncertainty to justify average utilitarianism—a social welfare function that maximizes average utility. Harsanyi's argument distinguishes between personal preferences and moral preferences, and defines the latter in line with the utilitarian tradition. Unlike personal preferences, which are self-regarding, the set of moral preferences is "based on a serious attempt to give the same weight to the interests of every member of society, in accordance with the principle of average utility" (Harsanyi 1975, p. 598). The link between his notion of moral preferences and utilitarianism is an equiprobability assumption: people evaluate alternative institutional arrangements as if they each had equal chances to end up in any given social position. The veil of ignorance is a metaphor for this equiprobability assumption. Harsanyi then shows that if people are rational, they will choose that institutional arrangement which on average gives them the highest expected utility. In his view, maximin would be a highly irrational choice, as it neglects how much good can be produced in society and looks only at the interests of the worst-off.

At the heart of this controversy is the question of the role of probabilities in people's reasoning and of attitudes towards risk in the argument from the original position. Rawls objects to "the very use of probabilities in the original position"

(Harsanyi 1975, p. 598), as well as to the use of von Neumann–Morgenstern (vNM) utility functions on the grounds that people's attitudes towards risk—which are part of how these utility functions are construed—are not automatically relevant when it comes to justifying an institutional arrangement. Harsanyi admits that disregarding subjective probabilities makes maximin-type rules more plausible, but thinks that doing so would be irrational. He also defends using vNM utility functions in the argument from the original position on the grounds that they are, first of all, an account of people's preferences. Many economists, Arrow included, have tended to accept Harsanyi's arguments.[22]

Some of Rawls's early replies to his critics may have provided economists with additional arguments for why their criticism is valid. In the article "Some Reasons for the Maximin Criterion", published in the *American Economic Review*, Rawls (1974) uses the language of utility and tries to argue for the difference principle via a defense of the assumption that persons in the original position are risk-averse. The easy reply for economists is to ask just how much risk aversion is plausibly assumed and to argue that the assumption of extreme risk aversion that is needed to justify maximin is not plausible. Harsanyi (1975) has a critique of maximin along those lines.

But Rawls's view can be supported on different grounds, and in later writings, Rawls clarified his views (e.g. Rawls 2001, §31). As I have argued above, Rawls's idea of the original position is intended to condense the fundamental ideas on which his theory of justice is based. While Rawls accepts Harsanyi's equiprobability assumption, the idea of the original position has additional normative constraints built in. What is particularly important is that the people behind the Rawlsian veil of ignorance are not just rational, but also reasonable. The persons in the original position enjoy equal liberty, and, by virtue of their two moral powers, can reason not only about how to use their liberty to best advance a conception of the good, but also about how to respect the equal liberty of everyone. The Rawlsian thought-experiment thus places people in a situation not just of equiprobability but also of reciprocity.

The kind of reasoning that Rawls imagines persons behind the veil of ignorance to be engaged in supports his argument from 'the strains of commitment'. This argument defends the difference principle against the principle of maximizing average utility (1999, p. 229; 2001, p. 103). As Rawls puts it, those behind the veil of ignorance "must ask themselves whether those they represent can reasonably be expected to honor the principles agreed to in the manner required by the idea of an agreement" (Rawls 2001, p. 103). Because of the strains of commitment, Rawls argues, it will be difficult for those who end up in the lower positions in society to sustain utilitarianism. Utilitarianism "asks them to view the greater advantages of others who have more as a sufficient reason for having still lower prospects of

[22] For a recent discussion of these issues, see Kolm (1998).

life than otherwise they could be allowed" (1999, p. 230). The difference principle, "by contrast, . . . assures the less favored that inequalities work to their advantage" (Rawls 1999, p. 230). The difference principle thus takes into account the strains of commitment placed on the worse-off. A defender of utilitarianism will probably object that this argument can be turned around and used to attack the difference principle from the perspective of the better-off. Rawls argues that this is not the case. Although the better-off in a society that satisfies the two principles of justice as fairness would have received more if the utilitarian principle was implemented, they give up less than the worse-off do in a utilitarian society compared with the just society because the respect for equal liberty is respected in one case but not in the other. The strains of commitment operate asymmetrically.

18.3.3 Interpersonal Comparisons of What?

Maximin, as I have defined it, is explicitly based on individual utility. As such, it contrasts with Rawls's difference principle, which refers to primary goods. This raises a question about what the appropriate informational framework is for evaluating justice. Is it utility, primary goods, or some alternative informational framework? Ever since Sen's Tanner Lecture with the title "Equality of What?" (Sen 1980), this question has been intensely debated in economics, philosophy, and political theory.[23]

The debate in economics is influenced by the problem of interpersonal comparisons. The transition from the old to the new welfare economics and to social choice theory was brought about, among other things, by the rejection of the assumption that utility functions are interpersonally comparable (Robbins 1938). After Arrow published his impossibility result, this assumption was revisited. It can be shown that if the axioms that social welfare functions have to satisfy are modified in order to allow for interpersonal comparisons, a variety of social welfare functions become possible. Note that cardinal measurability of utility functions alone is not sufficient to invalidate the impossibility result—comparability is necessary (Sen 1970*a*, thm. 8*2).[24]

Against this background, two broad positions can be identified in normative economics. One strand defends welfarism and is inclined to accept the implications of the impossibility of interpersonal comparisons of utility.[25] We saw above that this is, for example, Arrow's position. The other strand rejects welfarism and pursues the possibilities that alternative informational framework offer. Sen's work exemplifies this position.

[23] Recent contributions are from Pogge (2002) and Vallentyne (2005); see also the "equality exchange" webpage, <http://mora.rente.nhh.no/projects/EqualityExchange/>.

[24] On this issue, see also Hammond (1976).

[25] In a recent paper, Fleurbaey (2003) has argued that this does not follow.

Welfarism, particularly if combined with the Pareto principle, is usually regarded as a very weak value judgment, and thus as an attractive framework for social evaluation. There are, however, two major lines of criticism of welfarism. One objects to the identification of individual well-being with the satisfaction of individual preferences. The other is directed against the assumption of preferences being given. Pursuing these criticisms will reveal that the value judgments implicit in this framework are not so weak after all.

The identification of individual well-being with the satisfaction of individual preferences is especially problematic if welfare judgments are based on actual preferences. Actual preferences may be revealed, according to the theory of revealed preferences, through the choices individuals make. But what an individual chooses is not necessarily linked to his or her well-being. For preferences may be based on false information or beliefs. Equally, an individual may decide according to values other than those maximizing his or her well-being, such as responsibility, moral commitments, standards of politeness, etc. and make choices that reduce his or her well-being.[26] Another problem is that welfarism may exclude important information from becoming relevant in social evaluation. Sen (1979, pp. 547 f.) discusses the following example. Take two individuals in two situations (two social states). Assume that there is the same pattern of individual utilities in the two states. From the point of view of welfarism, they should be regarded as equally good. But if we know that in one state one individual is tortured by the other, we would certainly not want to treat the two states in the same way.

A further problem arises if preferences are not independent of social states. This is the case, for example, with adaptive preferences, when individuals unintentionally tend to adjust their preferences to their possibilities.[27] Adaptive preferences pose a serious obstacle for social evaluation based on individual preferences. For one, if preferences are formed adaptively, the satisfaction of these preferences will yield a distorted picture of individual well-being. More generally, if these preferences vary with the social states, how should the social choice between the alternative states be made? If there is a different preference profile for every alternative in the choice set, a decision regarding the relevant profile in terms of which the social choice is to be made is needed first.[28]

The so-called expensive tastes argument also poses a challenge for welfarism.[29] Whereas the problem of adaptive preferences points to the inappropriateness of using given preferences as the basis for moral assessment due to their dependence on the individual situation to be evaluated, the expensive tastes argument focuses on the opposite issue. That is, it is concerned with the individual capacity to reflect upon preferences and to exert a certain control over them. To the extent to which individuals have this capacity, they can also be held responsible for them.

[26] On this see Sen (1977). [27] The term "adaptive preferences" is from Elster (1983).
[28] Voorhoeve (2006) revisits this problem.
[29] See e.g. Rawls (1999, p. 369). See also Arneson (1990*b*); Cohen (1989); and Daniels (1990).

And, so the argument goes, if some individual has voluntarily cultivated expensive tastes, these preferences cannot be taken as the given basis for claims on society's resources. It is necessary to discriminate among preferences with respect to their origins if they are still to be the informational framework for social evaluation. Thus, while the problem of expensive tastes is just an example, the more fundamental issue is that preferences over which the individuals have control and for which they can be held responsible may not form an adequate informational framework in matters of distributive justice. "Luck egalitarians" (Anderson 1999) in particular have argued that considerations of responsibility matter in assessments of justice.[30]

One answer to these problems with welfarism is to argue for some kind of correction of actual preferences.[31] But this solution is affected by the further difficulty of needs to answer the question of who should "launder" actual preferences—the individuals themselves or someone else on their behalf?

Many, including Rawls, have argued that a further reason to leave the utility framework behind is the difficulty it poses for making interpersonal comparisons, and they have suggested alternative informational frameworks. Besides primary goods and other resourcist frameworks, Sen's capability approach is unquestionably the most important proposal. Sen's proposal was inspired by his rejection of both utility and Rawls's primary goods. His main objection to welfarism is that it does not contain enough information about social states. Sen's criticism of Rawls's primary goods, in short, is that focusing on the means that individuals have to pursue their ends is too rigid an informational framework. In particular, it does not respond to individual differences in the ability to make effective use of primary goods in the pursuit of their respective ends. Sen's capability approach is designed to overcome this shortcoming (see also Chapter 23 below). He suggests evaluating alternative social states on the basis of the possibilities that individuals have to achieve valuable "functionings". Functionings are a description of the various things an individual can do or be in a particular state, such as being well-nourished, being able to read, etc. Capabilities are then defined over the space of functionings. They reflect "the alternative combinations of functionings the person can achieve, and from which he or she can choose one collection" (Sen 1993, p. 31).

The capability approach is objective, in that it focuses on a set of valuable functions which will shape social evaluation independently of a particular individual's goals and interests. The basic idea is that the individuals need to have access to these functionings in order to pursue their different values and interests. Furthermore, the capability approach is deliberately open with respect to the set of functionings that will be identified as the relevant one for different purposes of social evaluation. The capability approach, as interpreted by Sen, is restricted to providing a general

[30] See Dworkin (1981); Arneson (1989); Cohen (1989); Roemer (1996); Vallentyne (2005).
[31] See e.g. Griffin (1986).

framework for social evaluation, without specifying a substantive list of functionings which would invariably determine the goodness of individual states.[32]

By focusing on opportunities people have, the capability approach leaves the purely outcome-oriented terrain of the utility framework. Other examples for consequentialist opportunities-based frameworks are Richard Arneson's opportunities for welfare and Gerald Cohen's midfare.[33] What these approaches share with welfarism, however, is the consequentialist structure. Common to consequentialist frameworks is the premise that social evaluation has to be made on the basis of an informational framework which captures what is intrinsically valuable to the individuals. Consequentialist theories differ in their definition of the good life. In utilitarianism, the good is identified with individual utility. It is, therefore, a form of "welfarist consequentialism" (Sen and Williams 1982): that is, of consequentialist theories which require the assessment of consequences in terms of individual utility. Although utilitarianism is only a particular form of consequentialism, it has been the most influential one, not least in economic theory. An alternative to utilitarianism is perfectionism. Perfectionist theories start from some ideal of the good life which is general in the sense that it applies to all individuals alike. It forms the basis on which the moral goodness and rightness of social states is to be assessed (e.g. Griffin 1986, p. 56).

Rawls's theory of justice rejects both welfarism and consequentialism. Approaches in welfare economics and social choice theory which remain wedded to welfarism and/or consequentialism thus miss an important aspect of his theory. One aspect of Rawls's non-consequentialism is the emphasis on procedural justice. The idea of procedural justice is not one that is easily incorporated into the framework of social choice theory. Because there are many reasons for why procedural justice is an important value (Anand 2001), social choice theorists have started to explore this idea. The emphasis tends to be on the "process aspect" of freedom (Sen 2002)—be it with regard to individual choice (e.g. Pattanaik and Xu 1990, 2000) or social choice (Suzumura and Xu 2004).

Another aspect of Rawls's non-consequentialism is the primary goods framework. As explained above, primary goods are not to be understood as a surrogate for individual well-being; nor do they need to be seen by people as intrinsically valuable. But what is it that is evaluated, if not individual well-being or the goodness of lives? To answer this, the primary goods framework must be seen as embedded in the political conception of justice in the same way as the informational framework of individual utility is embedded in utilitarian moral philosophy. It is dependent upon both the conception of the person and the conception of society that characterize the two theories. Primary goods are an answer to the question of what goods the citizens in the original position would want to have in

[32] This in contrast to Nussbaum (1988, 1993).

[33] Arneson (1989, 1990*a*, 1990*b*); Cohen (1989, 1993). See also Roemer (1996).

order to ensure fair prospects for developing, pursuing, and, if necessary, revising reasonable conceptions of the good, whatever they may be.[34] With the primary goods framework, Rawls hopes to circumvent the problems of both welfarism and perfectionism.

Even if these advantages of the primary goods framework are granted, there are further questions that need to be addressed. Many economists are of the view that an important problem affects the construction of an index of primary goods. The problem has its origin in the following dilemma.[35] If the index of primary goods ignores people's own evaluations of these goods, recommendations based on such an index will tend to violate the Pareto principle—people will often be willing to trade and exchange the bundle that has been assigned to them for one that they prefer. If the index takes people's own evaluation into account, however, does that not mean that the approach collapses into welfarism? The question is whether this dilemma is as damning for non-welfarist approaches to justice as it may appear. Many have argued—rightly, in my view—that this is not the case. First, as Sen has argued, taking into account people's own evaluations does not commit one to welfarism. Such evaluations may consist of an appreciation of the requirements of justice (Sen 1991). Indeed, one might add, that is the whole point of the Rawlsian contractarian approach. This thus addresses the second horn of the dilemma, the alleged collapse into welfarism. Brun and Tungodden (2004) have recently argued that while Sen is right, this solution does not eliminate the first horn of the dilemma, the possibility that evaluations based on such an index violate the Pareto principle. That is correct, of course. But is it a problem? Recall that, in the Rawlsian view, primary goods apply to the evaluation of alternative arrangements of the basic structure, not to the evaluation of resulting distributions directly. If a resulting just distribution is such that there are, say, two individuals who would still be willing to trade, this need not imply that they would consider the basic structure unjust. The distinction between what is rational and what is reasonable is helpful for grasping the difference. The individuals may perceive further trading to be to their rational advantage, while judging the prevailing arrangement of the basic structure to be justified on grounds of what they consider reasonable. A further consideration is that the Pareto criterion, as it has been extensively discussed in the literature, may select social states which are highly unappealing otherwise. Such social states may, for example, undermine the protection of individual rights (e.g. Sen 1970*b*, 1979). This raises the question of how to prioritize between the different evaluative criteria. In the Rawlsian perspective, efficiency is not taken to be the "first virtue" of social institutions; justice is. Violations of the Pareto criterion do not automatically constitute a transgression.

[34] See Rawls (1993, p. 180).

[35] See Plott (1978); Gibbard (1979); Blair (1988); Arneson (1990*b*); Sen (1991); Roemer (1996); Brun and Tungodden (2004); Fleurbaey (2005).

References

Alexander, S. S. (1974). Social Evaluation through Notional Choice. *Quarterly Journal of Economics*, 88, 597–624.

Anand, Paul (2001). Procedural Fairness in Economic and Social Choice: Evidence from a Survey of Voters. *Journal of Economic Psychology*, 22, 247–70.

Anderson, Elizabeth (1999). What is the Point of Equality? *Ethics*, 109/2, 287–337.

Arneson, Richard J. (1989). Equality and Equal Opportunity for Welfare. *Philosophical Studies*, 56, 77–93.

—— (1990*a*). Liberalism, Distributive Subjectivism, and Equal Opportunity for Welfare. *Philosophy and Public Affairs*, 19, 158–94.

—— (1990*b*). Primary Goods Reconsidered. *Noûs*, 24, 429–54.

Arrow, Kenneth (1963). *Social Choice and Individual Values*. New Haven: Yale University Press.

—— (1973). Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice. *Journal of Philosophy*, 70, 245–67. Repr. in Arrow (1984), 96–114.

—— (1984). *Collected Papers*, i: *Social Choice and Justice*. Oxford: Blackwell.

Blackorby, Charles, Bossert, Walter, and Donaldson, David (2002). Utilitarianism and the Theory of Justice. In Kenneth Arrow, Amartya Sen, and Kotaro Suzumura (eds.), *Handbook of Social Choice and Welfare*, i. 543–96. Amsterdam: Elsevier Science.

Binmore, Ken (2005). *Natural Justice*. Oxford: Oxford University Press.

Blair, Douglas H. (1988). The Primary-Goods Indexation Problem in Rawls's Theory of Justice. *Theory and Decision*, 24, 239–52.

Brun, Berndt Christian, and Tungodden, Bertil (2004). Non-Welfaristic Theories of Justice: Is the "Intersection Approach" a Solution to the Indexing Impasse? *Social Choice and Welfare*, 22, 49–60.

Cohen, Gerald A. (1989). On the Currency of Egalitarian Justice. *Ethics*, 99, 906–44.

—— (1993). Equality of What? On Welfare, Goods, and Capabilities. In Martha Nussbaum, and Amartya Sen (eds.), *Quality of Life*, 9–29. Oxford: Clarendon Press.

Daniels, Norman (1990). Equality of What: Welfare, Resources, or Capabilities? *Philosophy and Phenomenological Research*, 50, 273–96.

Dworkin, Ronald (1981). What is Equality? Part 2: Equality of Resources. *Philosophy and Public Affairs*, 10, 283–345.

Elster, Jon (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Fleurbaey, Marc (2003). On the Informational Basis of Social Choice. *Social Choice and Welfare*, 21, 347–84.

—— (2005). Social Choice and the Indexing Dilemma. <http://cerses.shs.univ-paris5.fr/publifleurbaey/scatid.pdf>.

Gibbard, Allan (1979). Disparate Goods and Rawls' Difference Principle: A Social Choice Theoretic Treatment. *Theory and Decision*, 11, 267–88.

Griffin, James (1986). *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press.

Hammond, Peter J. (1976). Equity, Arrow's Conditions, and Rawls' Difference Principle. *Econometrica*, 44, 793–804.

Harsanyi, John C. (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, 62, 309–21.

——(1975). Can the Maximin Principle Serve as a Basis of Morality? A Critique of John Rawls's Theory. *American Political Science Review*, 69, 594–606.

Hohm, Larry (1983). Formulating Rawls's Principles of Justice. *Theory and Decision*, 15, 337–47.

Kolm, Serge (1998). Chance and Justice: Social Policy and the Harsanyi–Vickrey–Rawls Problem. *European Economic Review*, 42, 1393–416.

Musgrave, R. A. (1973). Maximin, Uncertainty, and the Leisure Tradeoff. *Quarterly Journal of Economics*, 87, 625–32.

Nozick, Robert (1974). *Anarchy, State, and Utopia*. Oxford: Blackwell.

Nussbaum, Martha (1988). Nature, Function, and Capability: Aristotle on Political Distribution. *Oxford Studies in Ancient Philosophy*, suppl. vol., 145–84.

——(1993). Non-Relative Virtues: An Aristotelian Approach. In Martha Nussbaum and Amartya Sen (eds.), *Quality of Life*, 242–69. Oxford: Clarendon Press.

Pattanaik, Prasanta, and Xu, Yongsheng (1990). On Ranking Opportunity Sets in Terms of Freedom of Choice. *Recherches Economiques de Louvain*, 56, 383–90.

————(2000). On Ranking Opportunity Sets in Economic Environments. *Journal of Economic Theory*, 93, 48–71.

Peter, Fabienne (2006). Justice: Political, not Natural. *Analyse und Kritik*, 28/1, 382–97.

Plott, Charles R. (1978). Rawls' Theory of Justice: An Impossibility Result. In H. W. Gottinger and W. Leinfeller (eds.), *Decision Theory and Social Ethics*, 201–14. Dordrecht: Reidel.

Pogge, Thomas (2002). Can the Capability Approach be Justified? *Philosophical Forum*, 30/2, 167–228.

Rawls, John (1971). *A Theory of Justice as Fairness*. Cambridge, MA: Harvard University Press.

——(1974). Some Reasons for the Maximin Criterion. *American Economic Review*, 64: 141–6. Repr. in Rawls (1999), 225–31.

——(1993). *Political Liberalism*. New York: Columbia University Press.

——(1999). *Collected Papers*, ed. Samuel Freeman. Cambridge, MA: Harvard University Press.

——(2001). *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.

Robbins, Lionel (1938). Interpersonal Comparisons of Utility: A Comment. *Economic Journal*, 48, 635–41.

Roemer, John (1996). *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Scanlon, Thomas (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Sen, Amartya (1970*a*). *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

——(1970*b*). The Impossibility of a Paretian Liberal. *Journal of Political Economy*, 78, 152–7.

——(1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs*, 6, 317–44.

——(1979). Personal Utilities and Public Judgements; or: What's Wrong with Welfare Economics? *The Economic Journal*, 89, 537–58.

——(1980). Equality of What?" In S. M. Mc Murrin (ed.), *The Tanner Lectures on Human Values*, i. 195–220. Salt Lake City: University of Utah Press.

——(1991). On Indexing Primary Goods and Capabilities. Mimeo, Harvard University.

Sen, Amartya (1993). Capabilities and Well-being. In Martha Nussbaum and Amartya Sen (eds.), *Quality of Life*, 30–53. Oxford: Clarendon Press.

—— (2002). *Rationality and Freedom*. Cambridge, MA: Harvard University Press.

—— and Williams, Bernard (eds.) (1982). *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Strasnick, Steven (1976). The Problem of Social Choice: Arrow to Rawls. *Philosophy and Public Affairs*, 5, 241–73.

Suzumura, Kotaro, and Xu, Yongsheng (2004). Welfarist-Consequentialism, Similarity of Attitudes, and Arrow's General Impossibility Theorem. *Social Choice and Welfare*, 22/1, 237–51.

Tungodden, Bertil (1999). The Distribution Problem and Rawlsian Reasoning. *Social Choice and Welfare*, 16, 599–614.

Vallentyne, Peter (2005). Capabilities versus Opportunities for Well-being. *Journal of Political Philosophy*, 13, 359–71.

Voorhoeve, Alex (2006). Preference Change and Interpersonal Comparisons of Welfare. *Royal Institute of Philosophy Supplement*, 81/59, 265–79.

CHAPTER 19

JUDGMENT AGGREGATION

CHRISTIAN LIST
CLEMENS PUPPE

19.1 Introduction

Judgment aggregation is the subject of a growing body of work in economics, political science, philosophy, and related disciplines. Although the literature on judgment aggregation has been influenced by earlier work in social choice theory, the recent interest in the problem was sparked by the so-called "doctrinal paradox" in law and economics (Kornhauser and Sager 1986). Suppose a three-member court has to reach a verdict in a breach-of-contract case. According to legal doctrine, the defendant is liable (the *conclusion*, here denoted c) if and only if he or she did a particular action *and* had a contractual obligation not to do it (the two *premises*, here denoted a and b). The doctrinal paradox consists in the fact that majority voting on the premises may support a different verdict from majority voting on the conclusion. As illustrated in Table 19.1, suppose the first judge holds both premises to be true; the second holds the first premise, but not the second, to be true; and the third holds the second premise, but not the first, to

We are grateful to Franz Dietrich, Ron Holzman, Philippe Mongin, Klaus Nehring, and the participants of the Workshop on Judgement Aggregation, September 2007, in Freudenstadt (Germany), for helpful comments and discussion.

Table 19.1. Doctrinal paradox/discursive dilemma

	Action Done, *a*	Obligation Held, *b*	*c* if and only if (*a* and *b*)	Liable *c*
Judge 1	True	True	True	True
Judge 2	True	False	True	False
Judge 3	False	True	True	False
Majority	True	True	True	False

be true. Then a majority of judges holds each premise to be true, which seems to support a "liable" verdict, and yet a majority of judges holds the conclusion to be false. Although the first discussions of this problem focused on the distinction between "premise-based" and "conclusion-based" methods of decision-making, the doctrinal paradox illustrates a more general point, which Pettit (2001) has called the "discursive dilemma": majority voting on multiple, interconnected propositions may lead to an inconsistent set of collective judgments. In the court example, majorities accept *a*, *b*, [*c* if and only if (*a* and *b*)], and the negation of *c*, an inconsistent set of propositions in the standard sense of logic (see also Brennan 2001).

Naturally, the observation that majority voting may fail to produce consistent collective judgments raises several questions. In particular, how general is the problem? Is it restricted to majority voting, or does it extend to other decision methods? And does it occur only in special situations, such as the breach-of-contract case, or does it arise more generally?

In response to these questions, List and Pettit (2002, 2004) proposed a model of judgment aggregation combining a logical representation of propositions with an axiomatic approach inspired by Arrovian social choice theory. Using this model, they proved a simple impossibility theorem showing that if judgments are to be made on at least two atomic propositions and at least one suitable composite proposition (and their negations), there exists no judgment aggregation rule satisfying four conditions: universal domain (all combinations of rational individual judgments are admissible as inputs), collective rationality (only rational collective judgments are admissible as outputs), anonymity (the aggregation is invariant under permutations of the individuals), and systematicity (the collective judgment on each proposition is the same function of individual judgments on that proposition).

This result, however, gives only a partial answer to the questions raised above. Its conditions on both the aggregation rules and the decision problems under consideration can be significantly generalized or otherwise refined (e.g. Pauly and Van Hees 2006; Dietrich 2006*a*, 2007*b*; Nehring and Puppe 2005*a*). Moreover, instead of producing mere impossibility results, the literature has now provided

several general characterizations of both aggregation rules and decision problems with salient properties (e.g. Dokow and Holzman 2005; Nehring and Puppe 2005*b*, 2007*b*; Dietrich and List 2007*a*). Some of these draw on other branches of aggregation theory that are closely related cousins of the logic-based framework, including the aggregation of binary evaluations (Wilson 1975; Rubinstein and Fishburn 1986) and the theory of strategy-proof social choice on generalized single-peaked domains (Nehring and Puppe 2002). The interest in the problem of judgment aggregation is enhanced by the observation that the classical preference aggregation problem of social choice theory is a special case, by representing preference relations as sets of binary ranking judgments (List and Pettit 2004; Nehring 2003; Dietrich and List 2007*a*), as explained in Section 19.3.2.1. An earlier precursor is Guilbaud's (1966) logical reformulation of Arrow's theorem. More generally, by representing decision problems not in standard propositional logic but in other more expressive logics, many realistic decision problems can be expressed as judgment aggregation problems (Dietrich 2007*b*). Judgment aggregation is also related to the theory of belief merging in computer science (Konieczny and Pino-Perez 2002).

Our aim in this survey article is to provide an accessible overview of some key results and questions in the theory of judgment aggregation. We omit proofs and technical details, focusing instead on concepts and underlying ideas. But our perspective in this survey is a social-choice-theoretic one; we do not attempt to review the related legal and philosophical literatures. After introducing and discussing the formal framework in Section 19.2, we devote the bulk of our discussion to propositionwise aggregation rules, i.e. ones satisfying an independence condition (Section 19.3). For the purpose of this survey, this focus is justified by the fact that most of the technical results in the literature pertain to aggregation rules satisfying independence. We do not unreservedly endorse the independence condition, however, and discuss its relaxation in Section 19.4. In Section 19.5 we address other themes and developments in the literature.

19.2 Modeling Judgment Aggregation

19.2.1 The Logic-Based Framework

We use Dietrich's (2007*b*) model of judgment aggregation in general logics, which extends List and Pettit's (2002) original model in standard propositional logic. We consider a set of individuals $N = \{1, 2, \ldots, n\}$ (with $n \geq 3$). They are faced with a decision problem that requires making collective judgments on logically interconnected propositions. Propositions are represented in formal logic. The

language of the logic, $\mathcal{L}$, can be any set of sentences (called *propositions*) closed under negation (i.e. $p \in \mathcal{L}$ implies $\neg p \in \mathcal{L}$). The best-known example is standard propositional logic; here $\mathcal{L}$ is the smallest set containing (i) given atomic propositions $a, b, c, \ldots$ and (ii) for any $p, q \in \mathcal{L}$, the composite propositions $\neg p$, $(p \wedge q)$, $(p \vee q)$, $(p \rightarrow q)$, $(p \leftrightarrow q)$ with logical connectives $\neg$ ('not'), $\wedge$ ('and'), $\vee$ ('or'), $\rightarrow$ ('if-then') and $\leftrightarrow$ ('if and only if'). Other logics have languages involving other logical connectives, which often feature in realistic judgment aggregation problems (see Section 19.5.2). The logic is endowed with a notion of *consistency*, which satisfies some regularity conditions.[1] In standard propositional logic, for instance, a set of propositions $S \subseteq \mathcal{L}$ is *consistent* if there exists a truth-value assignment (with standard properties) making all propositions in S true, and *inconsistent* otherwise. For example, the set $\{a, a \vee b\}$ is consistent, while $\{\neg a, a \wedge b\}$ is inconsistent.

A decision problem is given by an *agenda* $X \subseteq \mathcal{L}$, interpreted as the set of propositions on which judgments are to be made. We assume that X is finite and closed under negation (i.e. if $p \in X$, then $\neg p \in X$) and identify each doubly-negated proposition $\neg\neg p$ with the nonnegated proposition p. We also exclude tautologies and contradictions from the agenda.[2]

An (individual or collective) *judgment set* is a subset $J \subseteq X$, interpreted as the set of accepted propositions in the agenda. For our purposes, to accept p means to believe p, but different interpretations of "acceptance" can be given (for instance, in terms of desire). The notion of belief is very general here, applicable both to positive propositions (e.g. "current CO_2 emissions lead to global warming") and to normative ones (e.g. "we should reduce CO_2 emissions"). Judgment aggregation problems cannot be resolved simply by statistical information-pooling techniques, since individuals may agree to disagree on the propositions, particularly if these are normative. We call J *consistent* if it is a consistent set in $\mathcal{L}$, and *complete* if $p \in J$ or $\neg p \in J$ for any proposition $p \in X$. A *profile* is an n-tuple $(J_1, J_2, \ldots, J_n)$ of individual judgment sets.

A *(judgment) aggregation rule* F is a mapping which assigns to each profile $(J_1, J_2, \ldots, J_n)$ of individual judgment sets (in some domain) a collective judgment set $J = F(J_1, J_2, \ldots, J_n)$. An aggregation rule F has *universal domain* if its domain is the set of all profiles of consistent and complete judgment sets; it is *collectively rational* if it generates a consistent and complete collective judgment set $F(J_1, J_2, \ldots, J_n)$ for every profile $(J_1, J_2, \ldots, J_n)$ in its domain. Until Section 19.5, we focus on aggregation rules with these two properties.

[1] The three conditions are: (C1) all sets $\{p, \neg p\} \subseteq \mathcal{L}$ are inconsistent; (C2) all subsets of consistent sets $S \subseteq \mathcal{L}$ are consistent; (C3) $\emptyset$ is consistent and each consistent set $S \subseteq \mathcal{L}$ has a consistent superset $T \subseteq \mathcal{L}$ containing a member of each pair $p, \neg p \in \mathcal{L}$.

[2] A proposition $p \in \mathcal{L}$ is a *tautology* if $\{\neg p\}$ is inconsistent, and a *contradiction* if $\{p\}$ is inconsistent.

19.2.2 An Example

Let us consider an illustrative decision problem. A three-member cabinet, $N = \{1, 2, 3\}$, has to make judgments on the following propositions:

a: Current CO_2 emissions lead to global warming.
$a \rightarrow b$: If current CO_2 emissions lead to global warming, then we should reduce CO_2 emissions.
b: We should reduce CO_2 emissions.

The agenda is the set $X = \{a, \neg a, a \rightarrow b, \neg(a \rightarrow b), b, \neg b\}$. The cabinet members' judgments as shown in Table 19.2 are given by the following individual judgment sets: $J_1 = \{a, a \rightarrow b, b\}$, $J_2 = \{a, \neg(a \rightarrow b), \neg b\}$, $J_3 = \{\neg a, a \rightarrow b, \neg b\}$.

If we use (propositionwise) majority voting as the aggregation rule, we obtain the same problem as identified above: an inconsistent collective set of judgments $J = \{a, a \rightarrow b, \neg b\}$. Thus, under universal domain, majority voting on the present agenda X is not collectively rational. By contrast, a dictatorship of minister 1—say, the prime minister—obviously guarantees a consistent collective judgment set. As we observe below, there are nondictatorial and collectively rational aggregation rules with universal domain for this agenda.

19.2.3 The Abstract Aggregation Framework

It is useful to relate the present logic-based framework of judgment aggregation to the framework employed in abstract aggregation theory, following Wilson (1975).[3] In abstract aggregation, individual vectors of yes/no evaluations over multiple binary issues are aggregated into a collective such vector, subject to feasibility constraints.[4] An (*abstract*) *aggregation rule* is a mapping $f : Z^n \rightarrow Z$, where $Z \subseteq \{0, 1\}^k$ represents the set of feasible yes/no evaluation vectors over k (≥ 1) binary

Table 19.2. CO_2 emissions

	Global warming, a	$a \rightarrow b$	Reduce emissions, b
Minister 1	True	True	True
Minister 2	True	False	False
Minister 3	False	True	False
Majority	True	True	False

[3] The property space framework of Nehring and Puppe (2002) is informationally equivalent to the abstract aggregation framework. Any property space can be uniquely embedded in $\{0, 1\}^k$ for some k up to isomorphism; conversely, any subset of $\{0, 1\}^k$ (with at least two elements) uniquely defines a property space.

[4] This is generalized to nonbinary issues in Rubinstein and Fishburn (1986).

issues. A judgment aggregation problem with agenda X can be represented in this framework by defining Z as the set of admissible truth-value assignments over the (unnegated) propositions in X, identifying binary issues with proposition-negation pairs (thus $k = \frac{|X|}{2}$).[5] To illustrate, consider the agenda of the global warming example above, $X = \{a,\ \neg a,\ a \rightarrow b,\ \neg(a \rightarrow b),\ b,\ \neg b\}$. The set of admissible truth-value assignments over the unnegated propositions a, $a \rightarrow b$, and b is $Z = \{(1, 1, 1), (1, 0, 0), (0, 1, 1), (0, 1, 0)\}$.[6]

There is a loss of information by moving from the logic-based framework to the abstract one. It consists in the fact that, while every agenda X uniquely defines a subset $Z \subseteq \{0, 1\}^k$, the same subset Z may arise from different agendas. For example, consider the agendas $X_1 = \{a, \neg a, a \vee b, \neg(a \vee b)\}$ and $X_2 = \{a \wedge b, \neg(a \wedge b), a \rightarrow b, \neg(a \rightarrow b)\}$. As is easily verified, the set of admissible truth-value assignments corresponding to both X_1 and X_2 is $Z = \{(1, 1), (0, 1), (0, 0)\}$. But obviously X_1 and X_2 are different in terms of both interpretation and syntax. In particular, X_2 contains only composite propositions, whereas X_1 also contains an atomic one, a fact that one may wish to use in handling the two aggregation problems (see Section 19.5 for examples). In what follows we use the logic-based framework but make cross-references to abstract aggregation at various points.

19.2.4 Conditions on Aggregation Rules

We now turn to conditions one may wish to impose on an aggregation rule F. We begin with the uncontroversial requirement that if all individuals unanimously submit the same judgment set, this judgment set should be the collective one.

Unanimity: For any unanimous profile $(J, \ldots, J)$ in the domain, $F(J, \ldots, J) = J$.[7]

The next condition requires that the collective judgment on each proposition p should depend only on individual judgments on p, not on individual judgments on other propositions.

Independence: For any $p \in X$ and any profiles $(J_1, \ldots, J_n), (J'_1, \ldots, J'_n)$ in the domain, if [for all $i \in N$, $p \in J_i \Leftrightarrow p \in J'_i$], then [$p \in F(J_1, \ldots, J_n) \Leftrightarrow p \in F(J'_1, \ldots, J'_n)$].

[5] Under this representation, the conditions of universal domain and collective rationality, explicitly imposed on the judgment aggregation rule F, are implicitly built into the definition of the abstract aggregation rule f.

[6] Here the conditional $\rightarrow$ is interpreted as the "material" conditional of standard propositional logic. In Sect. 19.3.2.3, we contrast this with a "subjunctive" interpretation.

[7] If F is also required to satisfy monotonicity as defined below, unanimity follows from the even weaker condition of *sovereignty*, whereby F has all complete and consistent judgment sets in its range.

While not uncontroversial, independence has some prima facie appeal in that it guarantees a propositionwise approach to aggregation. A stronger condition results from combining independence with a neutrality condition, requiring in addition equal treatment of all propositions.

Systematicity: For any $p, q \in X$, and any profiles $(J_1, \ldots, J_n), (J_1', \ldots, J_n')$ in the domain, if [for all $i \in N$, $p \in J_i \Leftrightarrow q \in J_i'$], then [$p \in F(J_1, \ldots, J_n) \Leftrightarrow q \in F(J_1', \ldots, J_n')$].

The following monotonicity condition states that if one individual switches from rejecting to accepting a collectively accepted proposition (keeping fixed all other individuals' judgments), this proposition should remain collectively accepted. In the presence of independence, monotonicity seems a natural requirement. To state it formally, we call two profiles *i-variants* if they coincide for all individuals except possibly i.

Monotonicity: For any $p \in X$, $i \in N$ and i-variants $(J_1, \ldots, J_i, \ldots, J_n)$, $(J_1, \ldots, J_i', \ldots, J_n)$ in the domain, if [$p \notin J_i$, $p \in J_i'$, and $p \in F(J_1, \ldots, J_i, \ldots, J_n)$], then $p \in F(J_1, \ldots, J_i', \ldots, J_n)$.

The final two conditions are basic democratic requirements. The first requires that no single individual should always determine the collective judgment set; the second requires that all individuals should have equal weight in the aggregation.

Nondictatorship: There exists no $i \in N$ such that, for any profile $(J_1, \ldots, J_n)$ in the domain, $F(J_1, \ldots, J_n) = J_i$.

Anonymity: For any profiles $(J_1, \ldots, J_n), (J_1', \ldots, J_n')$ in the domain that are permutations of each other, $F(J_1, \ldots, J_n) = F(J_1', \ldots, J_n')$.

19.3 Propositionwise Aggregation

We are now ready to express the fundamental questions raised by the discursive dilemma more formally. First, is the failure to achieve collective rationality restricted to majority voting, or does it extend to other aggregation rules? And second, how large is the class of agendas for which the problem arises? Notice the difference in focus between these two questions. The first concerns the class of *aggregation rules* that guarantee collective rationality for a given agenda, whereas the second concerns the class of *agendas* for which collectively rational aggregation rules with specific additional properties (such as the ones just introduced) exist.

The original theorem by List and Pettit (2002) answers these questions for a special class of aggregation rules (those satisfying universal domain, collective

rationality, systematicity, and anonymity) and a special class of agendas (those containing at least two atomic propositions and at least one suitable composite proposition in standard propositional logic).

Theorem 1 (List and Pettit 2002). If $X \supseteq \{a, b, a \wedge b\}$ (where $\wedge$ could be replaced by $\vee$ or $\rightarrow$), there exists no aggregation rule satisfying universal domain, collective rationality, systematicity, and anonymity.

A stronger version of this result, due to Pauly and Van Hees (2006), weakens anonymity to nondictatorship. These results, however, should only be seen as first "baseline" results. They provide only sufficient, but not necessary, conditions on the agenda for an impossibility to arise, and for many agendas, the systematicity condition on the aggregation rule can be weakened to independence alone in the presence of some other conditions. Let us therefore review several more general characterization results.

19.3.1 Characterization Results

In all results reviewed in this subsection, we consider collectively rational aggregation rules with universal domain satisfying unanimity and independence. We ask whether such rules can also be nondictatorial and/or anonymous. We distinguish between results with and without the requirement of monotonicity, and within each category between results with and without the requirement of neutrality, i.e. independence strengthened to systematicity. For our exposition, we call an aggregation rule *regular* if it is collectively rational, has universal domain, and satisfies unanimity.

19.3.1.1 *Results with Monotonicity*

The main advantage of assuming monotonicity is the resulting simple characterization of the class of propositionwise aggregation rules for any given agenda X, as follows (Nehring and Puppe 2002, 2007*b*). Let $\mathcal{W}$ denote a nonempty family of subsets of N, closed under taking supersets and interpreted as a family of "winning coalitions" of individuals. A *structure of winning coalitions* assigns to each proposition $p \in X$ a family $\mathcal{W}_p$ with these properties such that $W \in \mathcal{W}_p \Leftrightarrow (N \setminus W) \notin \mathcal{W}_{\neg p}$. An aggregation rule F is called *voting by issues* if, for all $p \in X$ and all profiles $(J_1, \ldots, J_n)$ in the universal domain,

$$p \in F(J_1, \ldots, J_n) \Leftrightarrow \{i : p \in J_i\} \in \mathcal{W}_p.$$

Thus a proposition is collectively accepted if and only if the set of individuals accepting it is a winning coalition for that proposition. As shown in Nehring and Puppe (2007*b*), an aggregation rule with universal domain satisfies unanimity,

independence, and monotonicity, and always accepts exactly one member of each pair $p, \neg p \in X$ if and only if it is voting by issues.

It is important to note, however, that voting by issues does not generally guarantee collective rationality. This can be seen from the fact that majority voting is an instance of voting by issues, where the family of winning coalitions $\mathcal{W}_p$ for each proposition p consists of all subsets of N with more than $n/2$ members. The necessary and sufficient condition for collective rationality of voting by issues is the following. Call a set of propositions S *minimally inconsistent* if S is inconsistent and every proper subset of S is consistent. Examples of minimally inconsistent sets are $\{a, \neg a\}$ and $\{a, a \rightarrow b, \neg b\}$; by contrast, the set $\{\neg a, \neg b, a \wedge b\}$ is inconsistent, but not minimally so.

Theorem 2 (Nehring and Puppe 2002, 2007*b*). Voting by issues on X with winning coalitions $(\mathcal{W}_p)_{p \in X}$ is collectively rational if and only if, for all minimally inconsistent subsets $\{p_1, \ldots, p_l\} \subseteq X$ and all selections $W_j \in \mathcal{W}_{p_j}$, $\bigcap_{j=1}^{l} W_j \neq \emptyset$.

The characterizing condition in Theorem 2 is called the *intersection property*.[8] It provides a powerful tool for determining both the class of regular, independent, and monotonic aggregation rules for a given agenda and the class of agendas admitting such rules with additional properties. Theorems 3, 4, and 5 in this section are instances of the second category of results; we illustrate the first category of results in Section 19.3.2.2 below.

We distinguish classes of agendas in terms of their logical complexity. The first result uses the following condition:

Median Property: All minimally inconsistent subsets of the agenda X contain exactly two propositions.

The median property says that the agenda is "simple" in the sense that direct interconnections between propositions are confined to pairs.[9]

Theorem 3 (Nehring and Puppe 2007*b*). There exist regular, monotonic, systematic, and nondictatorial aggregation rules on the agenda X if and only if X has the median property.

Using Theorem 3 it can easily be shown that if the number of individuals is odd, an agenda with the median property even admits regular, monotonic, and systematic rules that are *anonymous*. In other words, with an odd number of individuals the median property is also necessary and sufficient for majority voting to be collectively rational (see Nehring and Puppe 2007*b*; Dietrich and List 2007*b*).

[8] Dietrich and List (2007*b*) show that the intersection property can be generalized to one for collective consistency and one for collective deductive closure (each are weakenings of collective rationality).

[9] The terminology stems from the fact that agendas satisfying the median property correspond to so-called median spaces when embedded in the property space framework.

The next result uses a stronger condition on the agenda. Say that p *conditionally entails* q if $p \neq \neg q$ and there exists a minimally inconsistent subset $Y \subseteq X$ such that $p, \neg q \in Y$. Intuitively, this means that q can be deduced from p, using other propositions in the agenda. We write $p \rhd q$ if there exists a sequence $p_1, p_2, \ldots, p_m$ with $p = p_1$ and $q = p_m$ such that p_1 conditionally entails p_2, p_2 conditionally entails $p_3, \ldots,$ and p_{m-1} conditionally entails p_m.

Total Blockedness: The agenda X is *totally blocked* if for any pair of propositions $p, q \in X$, $p \rhd q$.

Total blockedness says that any proposition in the agenda can be reached from any other proposition in it via a sequence of conditional entailments. One can show that if an agenda satisfies the median property it cannot be totally blocked.

Theorem 4 (Nehring and Puppe 2005*b*). There exist regular, monotonic, independent, and nondictatorial aggregation rules on the agenda X if and only if X is not totally blocked.

Viewed as a possibility result, Theorem 4 is not completely satisfactory, since it admits quite degenerate possibilities such as local dictatorships, i.e. dictatorships on particular propositions. Agendas admitting regular, monotonic, independent, and locally nondictatorial aggregation rules can be characterized as well, using the intersection property (Nehring and Puppe 2005*b*). They coincide with the class of agendas admitting regular, monotonic, independent, and anonymous rules with an odd number of individuals, but the characterizing condition ("quasi-blockedness") is somewhat complicated. A much simpler characterization is obtained by requiring anonymity for an arbitrary number of individuals.

Blockedness: The agenda X is *blocked* if for some proposition $p \in X$, $p \rhd \neg p$, and $\neg p \rhd p$.

Theorem 5 (Nehring and Puppe 2005*b*). There exist, for any number of individuals n, regular, monotonic, independent, and anonymous aggregation rules on the agenda X if and only if X is not blocked.

For many agendas, all anonymous rules are unanimity rules (with certain default judgments in the absence of unanimity), and, without anonymity, all admissible rules are "oligarchic" (with certain default judgments when the oligarchs disagree), as defined in Section 19.3.2.2 below. In particular, as shown in Nehring (2006), under regularity, monotonicity, and independence, a nontrivial agenda[10] admits only oligarchic rules in this sense if and only if, for all $p, q \in X$, $[p \rhd q$ and $q \rhd p]$ or $[p \rhd \neg q$ and $\neg q \rhd p]$ ("semi-blockedness"). In Section 19.5.1 we review some

[10] An agenda is called *nontrivial* if it contains at least two propositions p and q such that p is logically equivalent neither to q nor to $\neg q$.

characterizations of another kind of oligarchic rules (where disagreements between the oligarchs lead to incomplete judgments).

19.3.1.2 *Results without Monotonicity*

While monotonicity is arguably an appealing condition in the presence of independence, it is not used in many classic results in standard social choice theory, notably Arrow's theorem, and in several early results on judgment aggregation, including Theorem 1. We may therefore ask whether it is needed to arrive at general characterization results of the above kind. Without requiring monotonicity, a characterization of aggregation rules in terms of structures of winning coalitions, along the lines of the intersection property, is not known. However, characterization results can be obtained by introducing an additional agenda complexity condition.

Even-Number-Negation Property: The agenda X has a minimally inconsistent subset Y such that $(Y \setminus Z) \cup \{\neg p : p \in Z\}$ is consistent for some subset $Z \subseteq Y$ of even size.

This condition says that the agenda has a minimally inconsistent subset that can be made consistent by negating an even number of propositions in it (Dietrich 2007*b*; Dietrich and List 2007*a*). An equivalent algebraic condition is *non-affineness* (Dokow and Holzman 2005), which in turn is equivalent to the requirement that the agenda is not structurally equivalent to a set of propositions whose only logical connectives are $\neg$ and $\leftrightarrow$. The agendas in the discursive dilemma and global warming examples above, for instance, satisfy the even-number-negation property; we give further examples in Section 19.3.2. The following theorem generalizes the earlier results on systematicity by List and Pettit (2002) and Pauly and Van Hees (2006) (i.e. Theorem 1 above and its strengthening, respectively).

Theorem 6 (Dietrich and List 2007*a*). There exist regular, systematic, and non-dictatorial aggregation rules on the agenda X if and only if X satisfies the median property or violates the even-number-negation property.

By contraposition, Theorem 6 says that the class of regular and systematic aggregation rules are precisely the dictatorships if and only if the agenda satisfies the even-number-negation property and does not have the median property. Variants of this result continue to hold even when, as in the first theorems on systematicity just cited, no unanimity requirement is imposed on the aggregation rule. If an agenda satisfies the even-number-negation property and does not have the median property, the class of admissible aggregation rules then grows to contain all dictatorial and inverse dictatorial rules, provided the latter are consistent (Dietrich and List 2007*a*). If the agenda in addition has an inconsistent subset $Y \subseteq X$ such that $\{\neg p : p \in Y\}$ is consistent, then systematicity alone (under universal domain and collective rationality) suffices to characterize dictatorships (Dietrich 2007*b*).

Pauly and Van Hees (2006) derived the first impossibility theorem on judgment aggregation with systematicity weakened to independence, followed by Dietrich (2006*a*). These results made strong assumptions on the agenda, but—like the systematicity results just reviewed—imposed no unanimity requirement on the aggregation rule. Instead, aggregation rules were required merely to be nonconstant.

With unanimity, Dokow and Holzman (2005) showed that if an agenda is totally blocked, then regular, independent, and nondictatorial rules exist if and only if the agenda violates the even-number-negation property. The aggregation rules needed for this result are derived from the *parity rule*, under which a proposition is collectively accepted if and only if it is accepted by an odd number of individuals (assuming, for simplicity, that n is odd); clearly, this defines a regular, systematic, anonymous (and hence, nondictatorial) but nonmonotonic aggregation rule.[11] Combined with Theorem 4, one thus obtains the following result. Its "only if" direction is also contained in Dietrich and List (2007*a*).

Theorem 7 (Dokow and Holzman 2005). There exist regular, independent, and nondictatorial aggregation rules on the agenda X if and only if X is not totally blocked or violates the even-number-negation property.

A simple example of a totally blocked agenda that violates the even-number-negation property (and for which parity rules are collectively rational) is the agenda $X = \{a, \neg a, b, \neg b, a \leftrightarrow b, \neg(a \leftrightarrow b)\}$, which shows that the even-number-negation property is essential in Theorems 6 and 7.

Table 19.3 summarizes the main results just presented. Each cell gives a necessary and sufficient condition for an agenda to admit regular and independent aggregation rules with the additional properties stated in the respective row and column; the even-number-negation property is abbreviated by "e-n-n". The results corresponding to anonymous but possibly nonmonotonic, respectively nonoligarchic and possibly nonmonotonic, rules have recently been obtained by Dietrich and List in an unpublished note. A characterization of the agendas admitting regular, independent, and locally nondictatorial aggregation rules without monotonicity is not yet known.[12]

19.3.2 Specific Agendas

To illustrate the applicability of the results just reviewed, we turn to agendas with specific additional structure.

[11] The general class of "parity rules" emerges by fixing any subset $M \subseteq N$ with an odd number of individuals and applying the rule stated in the text to subprofiles restricted to M.

[12] Note that the characterizing condition of the agendas admitting regular, independent, and anonymous rules (for all n) is the same regardless of whether or not monotonicity is required.

Table 19.3. Characterization of agendas for possibility results

	Monotonic	Possibly nonmonotonic
Neutral and nondictatorial	median	median or (not e-n-n)
Anonymous (for every *n*)	not blocked	not blocked
Locally nondictatorial	not quasi-blocked	?
Nonoligarchic	(not semi-blocked) or trivial	(not semi-blocked) or (not e-n-n)
Nondictatorial	not totally blocked	(not totally blocked) or (not e-n-n)

19.3.2.1 *Preference Agendas*

An important class of agendas are the *preference agendas*, in terms of which preference aggregation problems can be represented in the judgment aggregation model.[13] Here the additional structure stems from the rationality conditions on preferences. To represent preference relations, we use a simple language of predicate logic, with a two-place predicate symbol P representing strict preference and a set of constant symbols $K = \{x, y, z, \ldots\}$ representing social alternatives. A preference agenda is of the form $X = \{xPy, \neg xPy : x, y \in K \text{ with } x \neq y\}$. To capture rationality conditions on preferences, one has to define consistency appropriately: a set of propositions $S \subseteq X$ is deemed *consistent* if $S \cup H$ is consistent in the standard logical sense, where H is the appropriate set of rationality conditions on preference relations. In the case of strict preference orderings, these are asymmetry, transitivity, and connectedness.[14] It is easily seen that the judgment aggregation problem on X represents a classical preference aggregation problem, with each consistent and complete judgment set representing a fully rational preference relation.

There has been a sequence of contributions on how the results on judgment aggregation apply to preference aggregation. In a companion paper to their original paper, List and Pettit (2004) adapted their proof of Theorem 1 above to the preference agenda, showing that there is no anonymous, systematic, and collectively rational aggregation rule with universal domain here. Nehring (2003) subsequently proved that the preference agenda is totally blocked and hence, applying Theorem 4 above, showed that all regular, independent, and monotonic aggregation rules are dictatorial. Dietrich and List (2007*a*) and Dokow and Holzman (2005) showed that the preference agenda in addition satisfies the even-number-negation property

[13] There are various ways of representing preference aggregation problems in judgment aggregation or abstract aggregation. The present construction using predicate logic is based on Dietrich and List (2007*a*), extending List and Pettit (2004). For other related approaches, see Wilson (1975); Nehring (2003); Dokow and Holzman (2005); and Nehring and Puppe (2005*a*). An early construction was given by Guilbaud (1966).

[14] For example, transitivity is represented by the proposition $(\forall v_1)(\forall v_2)(\forall v_3)((v_1 P v_2 \wedge v_2 P v_3) \rightarrow v_1 P v_3)$.

(equivalently, non-affineness) and, by applying the "only if" part of Theorem 7, showed that all regular and independent aggregation rules are dictatorial. The latter result is Arrow's theorem for strict preferences.[15]

19.3.2.2 *Truth-Functional Agendas*

An important feature of the agenda in the original doctrinal paradox is that there is a conclusion (e.g. liability of the defendant) whose truth-value is uniquely determined by the truth-values of several premises (e.g. action and obligation). An agenda is called *truth-functional* if it can be partitioned into a sub-agenda of premises and a sub-agenda of conclusions such that each conclusion is truth-functionally determined by the premises. Nehring and Puppe (2005*a*) and Dokow and Holzman (2005) characterized classes of regular aggregation rules satisfying certain conditions on truth-functional agendas.

The bottom line is that all regular, independent, and monotonic rules on such agendas are oligarchic.[16] An *oligarchic rule* with default $J^0 \subseteq X$ specifies a nonempty set $M \subseteq N$ (the "oligarchs") such that, for all $p \in X$ and all profiles $(J_1, \ldots, J_n)$ in the universal domain,

$$p \in F(J_1, \ldots, J_n) \Leftrightarrow \begin{cases} & p \in J_i \text{ for all } i \in M \\ \text{or} & p \in J^0 \text{ and } [p \in J_i \text{ for some } i \in M] \end{cases}$$

Clearly, dictatorships are special cases (with M singleton). Nehring and Puppe (2005*a*) identified the truth-functional agendas admitting nondictatorial oligarchic rules ensuring collective rationality. For instance, in the global warming example above with agenda $X = \{a, \neg a, a \to b, \neg(a \to b), b, \neg b\}$, any oligarchic rule with default $J^0 = \{\neg a, a \to b, b\}$ is collectively rational.

19.3.2.3 *Agendas with Subjunctive Implications*

Dietrich (2006*b*) argued that, in many contexts, the material interpretation of the implication operator is not natural. To illustrate, consider again the global warming example above. Under a material interpretation of implication, the set of propositions $\{\neg a, \neg(a \to b), \neg b\}$ is inconsistent, since negating the antecedent a makes the material implication $a \to b$ true by definition. In everyday language, however, negating a (i.e. negating the proposition that current emissions lead to global warming) and negating b (i.e. negating the proposition that one should reduce emissions) seems perfectly consistent with the negation of any implication

[15] Dokow and Holzman (2006) and Dietrich (2007*a*) provided derivations of Arrow's theorem for weak preferences in judgment aggregation, each using different constructions.

[16] Dokow and Holzman (2005) did not assume monotonicity and weakened the unanimity requirement to surjectivity of the aggregation rule; therefore these authors obtained slightly different characterizations, depending on the complexity of the (truth-functional) agenda.

between a and b. Accordingly, a "subjunctive" interpretation of the implication operator (Lewis 1973) renders the set $\{\neg a, \neg(a \to b), \neg b\}$ consistent. Dietrich (2006*b*) showed that on this interpretation, the agenda in the global warming example admits collectively rational supermajority ("quota") rules (which are anonymous, monotonic, and independent). Generalizing the anonymous version of the intersection property, Dietrich (2006*b*) characterized the admissible quota rules on a large class of agendas with subjunctive implications.

19.3.2.4 *Non-Truth-Functional Agendas with a Premise/Conclusion Structure*

Agendas with subjunctive implications are usually not truth-functional. For instance, in the global warming example, affirming a and negating $a \to b$ is consistent with either affirming or negating b under a subjunctive interpretation of the implication. Nehring and Puppe (2007*a*) studied agendas containing a "conclusion" (a "decision") that depends in a general, not necessarily truth-functional way on some "premises" (the "decision criteria") from the viewpoint of *justifying* the collective decision. They provided several characterization results, including a characterization of the logical interrelations between the premises and the conclusion that enable independent and monotonic aggregation rules with majority voting on the conclusion. While such rules cannot exist in the truth-functional case, they do exist under reasonable circumstances in the non-truth-functional one. For instance, in the global warming example, the rule according to which b is decided by majority voting while a and $a \to b$ are affirmed if and only if each reaches a quota of at least 3/4 is consistent on the subjunctive interpretation.

19.3.2.5 *The Group Identification Problem*

In the group identification problem introduced by Kasher and Rubinstein (1997), each individual makes a judgment on which individuals belong to a particular social group subject to the constraint that the social group is neither empty nor universal. List (2008) formalized this problem in the judgment aggregation model and showed that the corresponding agenda is totally blocked and satisfies the even-number-negation property; therefore, by the "only if" part of Theorem 7 above, all regular and independent aggregation rules for the group identification problem are dictatorial. Dietrich and List (2006*b*) investigated the group identification problem in the case where the membership status of some individuals can be left undecided and showed that all regular and independent aggregation rules are oligarchies with empty default (see Section 19.5.1 below). A. Miller (2007) developed a model in which individuals make judgments about their membership in several social groups simultaneously.

19.3.3 Why Independence?

The independence condition in judgment aggregation is often challenged on the grounds that it fails to do justice to the fact that propositions are logically interconnected, which is the essence of the judgment aggregation problem (e.g. Chapman 2002; Mongin 2005). In this subsection, we put forward a possible "instrumental" justification of independence on the basis of strategy-proofness.[17] In fact, this justification also supports monotonicity.

The simplest way to implement the idea of strategy-proofness is in terms of the following nonmanipulability condition. Say that one judgment set J *agrees* with another J' on some proposition p if $[p \in J \Leftrightarrow p \in J']$. An aggregation rule F is *nonmanipulable* if there is no individual $i \in N$, no proposition $p \in X$, and no profile $(J_1, \ldots, J_n)$ in the domain such that, for some i-variant $(J_1, \ldots, J_i', \ldots, J_n)$ in the domain, $F(J_1, \ldots, J_n)$ does not agree with J_i on p and $F(J_1, \ldots, J_i', \ldots, J_n)$ agrees with J_i on p. Dietrich and List (2007*d*) showed that, under universal domain, an aggregation rule is nonmanipulable if and only if it is independent and monotonic, which allows the application of Theorems 2–5 above.

In fact, nonmanipulability corresponds to a standard social-choice-theoretic notion of strategy-proofness as follows. Assume that each individual i has a (reflexive and transitive) preference relation $\succcurlyeq_i$ over consistent and complete judgment sets such that, for some (unique) "ideal" judgment set J_i, we have $[J \cap J_i \supsetneq J' \cap J_i] \Rightarrow J \succ_i J'$ for any pair of judgment sets J, J'. Call such preferences *generalized single-peaked* (Nehring and Puppe 2002). A social choice function $\mathcal{F}$ mapping profiles of such preference relations to collective judgment sets is *strategy-proof* if, for all individuals i and all i-variants $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$, $(\succcurlyeq_1, \ldots, \succcurlyeq_i', \ldots, \succcurlyeq_n)$ in the domain,

$$\mathcal{F}(\succcurlyeq_1, \ldots, \succcurlyeq_n) \succcurlyeq_i \mathcal{F}(\succcurlyeq_1, \ldots, \succcurlyeq_i', \ldots, \succcurlyeq_n).$$

It can be shown that any such strategy-proof social choice function $\mathcal{F}$ depends only on the ideal judgment sets and thus induces a judgment aggregation rule F defined by $F(J_1, \ldots, J_n) := \mathcal{F}(\succcurlyeq_1, \ldots, \succcurlyeq_n)$, where, for all i, $\succcurlyeq_i$ is some generalized single-peaked preference relation with ideal judgment set J_i . The induced judgment aggregation rule is independent and monotonic; conversely, any independent and monotonic judgment aggregation rule F satisfying universal domain and collective rationality induces a strategy-proof social choice function $\mathcal{F}$ on the domain of generalized single-peaked preferences by appropriately reversing this construction (Nehring and Puppe 2002, 2007*b*). A definition of strategy-proofness of F, as opposed to strategy-proofness of $\mathcal{F}$, is given in Dietrich and List (2007*d*), extending List (2004).

[17] A closely related argument could be based on the absence of manipulations by the agenda-setter; see List (2004) and Dietrich (2006*a*).

19.4 Relaxing Independence

19.4.1 Premise-Based and Related Approaches

Perhaps the most discussed alternative to majority voting and propositionwise aggregation more generally is the class of premise-based procedures, applicable to truth-functional agendas in which the sub-agenda of premises can be chosen so as to consist of mutually independent proposition-negation pairs (see, among others, Kornhauser and Sager 1986; Pettit 2001; List and Pettit 2002; List 2005; Dietrich 2006*a*; Bovens and Rabinowicz 2006). A *premise-based procedure* is given by applying a suitable propositionwise aggregation rule (such as majority voting) to the premises and deducing the collective judgments on all other propositions (i.e. the conclusions) by logical implication. As an illustration, consider the doctrinal paradox example with individual judgments as shown in Table 19.1. If the premises are taken to be a, b, $c \leftrightarrow (a \wedge b)$ (and negations) and the conclusion is taken to be c (and its negation), the premise-based procedure (based on majority voting) yields the collective judgment set $\{a,\ b,\ c \leftrightarrow (a \wedge b),\ c\}$, i.e. a "liable" verdict.

The appeal of a premise-based procedure is that it is collectively rational and that the independence requirement is confined to logically independent propositions. Dietrich (2006*a*) characterized the premise-based procedure in terms of such a weakened independence condition. A problem, however, is that there does not always exist a unique way to specify premises and conclusions, and that different such specifications may lead to different collective judgment sets. For example, on the above agenda containing a, b, $c \leftrightarrow (a \wedge b)$, c (and negations), any three unnegated propositions (and their negations) can form a sub-agenda of mutually independent premises, setting interpretational issues aside. Using majority voting on the premises, each of these leads to a different collective judgment set in Table 19.1.

The same example also illustrates another problem of the premise-based procedure: majority voting on the premises may overrule a unanimous judgment on the conclusion, as can be seen by taking a, b, c (and negations) as the premises in Table 19.1. More generally, Nehring (2005) characterized truth-functional relations between multiple premises and one conclusion in terms of the admitted aggregation rules satisfying independence and monotonicity on the premises and respecting unanimous judgments on the conclusion; for sufficiently complex truth-functional relations, only dictatorial rules have these properties. Relatedly, Mongin (2005) proved that, for sufficiently rich agendas, the only regular aggregation rules satisfying independence restricted to atomic propositions (which one might view as premises) and a propositionwise unanimity condition are dictatorships. A conceptual difference between the two contributions lies in the

interpretation of the unanimity requirement on the outcome decision. Nehring (2005) interpreted it as a condition of Paretian welfare rationality, suggesting a potentially deep tension between "judgment rationality" (reason-basedness) and consequentialist outcome rationality. Mongin (2005) did not adopt the Paretian interpretation, applying the unanimity condition instead to every proposition. His analysis sought to show the robustness of an impossibility under weakening independence.

Bovens and Rabinowicz (2006) and subsequently List (2005) investigated the truth-tracking reliability of the premise-based procedure in cases where the propositions in question have independent truth conditions. Adapting the Condorcet jury theorem to the case of multiple interconnected propositions, they showed that, under a broad range of assumptions, the premise-based procedure leads to more reliable decisions than majority voting on the conclusion. Within this framework, List (2005) also calculated the probability of disagreements between the two procedures under various assumptions and, by implication, the probability of the occurrence of a majority inconsistency.

19.4.2 The Sequential Priority Approach

A premise-based procedure is a special case of a *sequential priority procedure* (List 2004; Dietrich and List 2007*b*), which can be defined for any agenda. Let an order of priority over the propositions in the agenda be given. Earlier propositions in that order may be interpreted as "prior to" later ones. For any profile, the collective judgment set is determined as follows. Consider the propositions in the agenda in the given order. For any proposition p, if the collective judgment on p is logically constrained by the collective judgments on propositions considered earlier, then it is deduced from those prior judgments by logical implication. If it is not constrained in this way, then it is made by majority voting or another suitable propositionwise aggregation rule.

By construction, any sequential priority procedure guarantees consistent collective judgment sets. Moreover, for truth-functional agendas, a sequential priority procedure can mimic a premise-based procedure if the premises precede the conclusions in the specified order of priority. But clearly sequential priority procedures can also be defined for non-truth-functional agendas. A key feature of sequential priority procedures is their *path dependence*: the collective judgment set may vary with changes in the order of priority over the propositions. Necessary and sufficient conditions for such path dependence were given by List (2004). Dietrich and List (2007*b*) further showed that the absence of path dependence is equivalent to strategy-proofness in a sequential priority procedure.

19.4.3 The Distance-Based Approach

In analogy to the corresponding approach in social choice theory (see e.g. Kemeny 1959), an alternative to propositionwise aggregation is a distance-based approach. Suppose that for a given agenda there is a metric which specifies the distance $d(J, J')$ between any two judgment sets. A *distance-based* aggregation rule determines the collective judgment set so as to minimize the sum of the individual distances. Formally, the collective judgment set for the profile $(J_1, \ldots, J_n)$ is a solution to

$$\min_J \sum_{i=1}^{n} d(J, J_i),$$

where the minimum is taken over all consistent and complete judgment sets.[18] A natural special case arises by taking d to be the *Hamming distance*, where $d(J, J')$ is the number of propositions in the agenda on which J and J' do not agree. This was proposed and analyzed by Pigozzi (2006) under the name "fusion operator" (see also Eckert and Klamler 2007), drawing on the theory of belief merging in computer science (Konieczny and Pino-Perez 2002). When applied to the preference agenda in Section 19.3.2.1 above, this aggregation rule is known as the "Kemeny rule" (Kemeny 1959; see Merlin and Saari 2000 for a modern treatment).

19.4.4 The Relevance Approach

Generalizing each of these specific approaches to relaxing independence, Dietrich (2007*a*) introduced a *relevance relation* between the propositions in the agenda, reflecting the idea that some propositions are relevant to others. For example, premises or prior propositions may be relevant to conclusions or posterior ones. Aggregation rules are now required to satisfy *independence of irrelevant information*: the collective judgment on any proposition p should depend only on the individuals' judgments on propositions relevant to p. The strength of this constraint depends on how many or few propositions are deemed relevant to each proposition: the fewer such relevant propositions, the stronger the constraint. In the limiting case where each proposition is relevant only to itself, the constraint is maximally strong and reduces to the standard independence condition.

The premise-based, sequential priority and distance-based approaches can all be seen as drawing on particular relevance relations: namely, premisehood, linear, and total (i.e. maximally permissive) relevance relations, respectively. Dietrich (2007*a*) proved several results on aggregation rules induced by general relevance

[18] A more general approach could allow also for other functions than the *sum* of individual distances.

relations, such as arbitrary premisehood or priority relations, which often induce a directed acyclic network over the propositions in the agenda. Whether there exist nondegenerate aggregation rules satisfying independence of irrelevant information depends on the interplay between logical connections and relevance connections.[19]

19.5 Other Themes and Contributions

At the time of writing this survey, judgment aggregation is still a very active research field in its developing stage. While the results for independent (i.e. propositionwise) aggregation in the case of two-valued logic seem to be near-definitive, many important aspects of judgment aggregation are not yet fully explored. In this concluding section, we briefly sketch several other themes and contributions that point towards directions for future research.

19.5.1 Rationality Relaxations

The possibility of judgment aggregation under weaker rationality constraints is the subject of several contributions. List and Pettit (2002) observed that, for a sufficiently large supermajority threshold, (*symmetrical*) *supermajority rules*—where any proposition is accepted if and only if it is accepted by a specified supermajority of individuals—guarantee consistency of collective judgments,[20] and unanimity rule in addition guarantees *deductive closure* (i.e. implications of accepted propositions are also accepted). More generally, Dietrich and List (2007*d*) provided necessary and sufficient conditions under which quota rules satisfy each of consistency, deductive closure, and completeness.

Gärdenfors (2006) proved an impossibility theorem showing that, under a particular agenda-richness assumption, any independent aggregation rule satisfying universal domain and unanimity, and generating consistent and deductively closed (but not necessarily complete) collective judgments is *weakly oligarchic*, in the sense that there exists a smallest subset $M \subseteq N$ such that, for all profiles $(J_1, \ldots, J_n)$, $F(J_1, \ldots, J_n) \supseteq \bigcap_{i \in M} J_i$.

[19] Arrow's theorem for weak preferences turns out to be a corollary of one of these results (see n. 15 above).

[20] As a sufficient condition on the supermajority threshold q, they gave $q > \frac{k-1}{k}$, where $k = \frac{|X|}{2}$, a result from List (2001); it also follows from the intersection property, generalized to the case of collective consistency. Dietrich and List (2007*b*) showed that this can be improved to a necessary and sufficient condition by defining k to be the size of the largest minimally inconsistent subset of X.

Generalizations of this result were given by Dietrich and List (2006*b*) and Dokow and Holzman (2006). The common finding is that, if collective rationality is weakened to the conjunction of consistency and deductive closure (and also if the completeness requirement is dropped at the individual level), the agenda conditions leading to dictatorships in the full-rationality case lead to *oligarchies* (*with empty default*), whereby there exists a subset $M \subseteq N$ such that, for all profiles $(J_1, \ldots, J_n)$, $F(J_1, \ldots, J_n) = \bigcap_{i \in M} J_i$. More precisely, Theorems 3, 4, 6, and 7 continue to hold if in their respective statements "nondictatorial" is strengthened to "nonoligarchic" and "full rationality" is weakened to "consistent and deductively closed" (optionally, full rationality at the individual level can also be replaced by consistency and deductive closure).[21] Dietrich and List (2006*b*) provided applications to the aggregation of partial orderings (including a variant of Gibbard's oligarchy theorem for strict preferences) and to the group identification problem (see above); Dokow and Holzman (2006) derived Gibbard's original oligarchy theorem and Arrow's theorem for weak preferences as corollaries.

More recently, Dietrich and List (2007*c*) provided a characterization of agendas leading to dictatorships when the rationality requirement at both individual and collective levels is weakened to consistency alone, dropping both completeness and deductive closure.

19.5.2 Multi-Valued Logic and General Logics

Pauly and Van Hees (2006) and Van Hees (2007) extended the model of judgment aggregation by allowing more than two degrees of acceptance, at both individual and collective levels. Thus they considered the aggregation of multi-valued truth functions. Building on, and generalizing, their impossibility results on systematicity and independence for two-valued logic, they showed that strong impossibility results arise even in this multi-valued context.

As mentioned above, Dietrich (2007*b*) developed a model of judgment aggregation in general logics, which allows the agenda to contain more expressive propositions than those of standard propositional logic. He argued that most realistic judgment aggregation problems and most standard examples of the discursive dilemma involve propositions that contain not only classical operators ("not", "and", "or",...) but also nonclassical ones, such as subjunctive conditionals (see above), modal operators ("it is necessary/possible that"), or deontic operators ("it is obligatory/permissible that"). The general logics model uses an arbitrary language $\mathcal{L}$ with a notion of consistency satisfying the minimal conditions stated in note 1. This includes many familiar logics: classical and nonclassical ones, propositional and predicate ones, and logics whose logical connections are defined

[21] The Dietrich and List (2006*b*) results in addition drop the consistency requirement at the collective level.

relative to a given set of constraints $C \subseteq \mathcal{L}$ such as the rationality constraints in the preference aggregation problem. Most theorems of the literature hold in general logics.

Pauly (2007) explored the role of language in judgment aggregation from a different perspective. He investigated the richness of the language required to express the conditions (such as unanimity, independence, systematicity, etc.) needed for characterizing various aggregation rules. This approach allowed him to derive some nonaxiomatizability results, showing that certain aggregation rules cannot be axiomatically characterized unless a sufficiently rich language is used to express the axioms.

19.5.3 Domain Restrictions

If the condition of universal domain is dropped and the domain of admissible profiles of individual judgment sets is suitably restricted, it becomes possible to satisfy all the other conditions on aggregation rules introduced above. Several domains are known on which majority voting is consistent. One such domain is the set of all profiles of consistent and complete individual judgment sets satisfying a condition called unidimensional alignment (List 2003). A profile is *unidimensionally aligned* if the individuals can be aligned from left to right such that, for each proposition in the agenda, the individuals accepting the proposition are either all to the left, or all to the right, of those rejecting it. Dietrich and List (2006*a*) provided several more general domain restriction conditions guaranteeing consistent majority judgments, including a local variant of unidimensional alignment, under which the relevant left–right alignment of the individuals can be different for each minimally inconsistent subset of the agenda, and some conditions that do not require complete individual judgment sets.

19.5.4 Judgment Aggregation with Disagreements on Connections between Premises and Conclusion

M. Miller (2007) offered a generalization of the truth-functional case of judgment aggregation by considering agendas consisting of several premises and a conclusion, where individuals may disagree about the logical connection between the former and the latter. The rationale behind this extension is that different individuals may reason in different ways and thus use different decision principles for the same decision. M. Miller (2007) proved an impossibility result showing that, again, certain types of oligarchic rules are the only collectively rational aggregation rules satisfying some reasonable conditions.

19.5.5 Liberal Paradox

In some judgment aggregation problems, some individuals or subgroups may have expert knowledge on certain propositions or be particularly affected by them. One may then wish to assign to these individuals or subgroups the right to determine the collective judgment on those propositions. Dietrich and List (2004) investigated how such rights constrain the available aggregation rules. Among other results, they showed that, for a large class of agendas, the assignment of rights to two or more individuals or subgroups is inconsistent with the unanimity condition. This result generalizes Sen's famous "liberal paradox" (1970), as it also applies to the preference agenda, where its conditions reduce to Sen's original conditions. Dietrich and List (2004) further identified domain restriction conditions under which the conflict between rights and the unanimity condition can be avoided.

In a related vein, Nehring (2005) shows that if an aggregation rule treats a proposition and its negation symmetrically, any differential treatment of voters as experts across propositions leads to potential violations of unanimity.

19.5.6 Bayesian Approaches

A natural step is to abandon the discrete, and mostly binary, nature of the evaluation of propositions. Continuous evaluations of propositions arise, for example, from a probabilistic interpretation of propositions or from their interpretation as economic variables. Claussen and Røisland (2005) analyzed the discursive dilemma in economic environments in which judgements involve quantitative assessments of variables. They showed that the original discursive dilemma (with majority voting on "premise variables") is robust with respect to the generalization to a continuous setting.

Nehring (2007) proposed a "Bayesian" model of group choice and showed that there does not generally exist any anonymous aggregation rule that is independent on the premises and always respects individuals' unanimous preferences over the outcome.

Bradley, Dietrich, and List (2006) applied insights from the theory of judgment aggregation to the aggregation of Bayesian networks, which consist of a causal relevance relation over some variables and a probability distribution over them. While some standard impossibility and possibility results also apply to the aggregation of causal relevance relations, a possibility result holds for the aggregation of the associated probability distributions.

Although these contributions underline the robustness of some of the impossibility results derived in the binary case, the Bayesian approach seems to offer new possibilities not yet explored.

References

Bovens, L., and Rabinowicz, W. (2006). Democratic Answers to Complex Questions—An Epistemic Perspective. *Synthese*, 150, 131–53.

Bradley, R., Dietrich, F., and List, C. (2006). Aggregating Causal Judgments. Working paper, London School of Economics.

Brennan, G. (2001). Collective Coherence? *International Review of Law and Economics*, 21/2, 197–211.

Chapman, B. (2002). Rational Aggregation. *Politics, Philosophy and Economics*, 1/3, 337–54.

Claussen, C. A., and Røisland, O. (2005). Collective Economic Decisions and the Discursive Paradox. Norges Bank Working Paper.

Dietrich, F. (2006*a*). Judgment Aggregation: (Im)Possibility Theorems. *Journal of Economic Theory*, 126/1, 286–98.

—— (2006*b*). The Possibility of Judgment Aggregation on Agendas with Subjunctive Implications. *Journal of Economic Theory*, forthcoming.

—— (2007*a*). Aggregation Theory and the Relevance of Some Issues to Others. Working Paper, London School of Economics.

—— (2007*b*). A Generalised Model of Judgment Aggregation. *Social Choice and Welfare*, 28/4, 529–65.

—— and List, C. (2004). A Liberal Paradox for Judgment Aggregation. *Social Choice and Welfare*, forthcoming.

—— —— (2006*a*). Judgment Aggregation on Restricted Domains. Working Paper, University of Maastricht.

—— —— (2006*b*). Judgment Aggregation without Full Rationality. *Social Choice and Welfare*, forthcoming.

—— —— (2007*a*). Arrow's Theorem in Judgment Aggregation. *Social Choice and Welfare*, 29/1, 19–33.

—— —— (2007*b*). Judgment Aggregation by Quota Rules: Majority Voting Generalized. *Journal of Theoretical Politics*, 19/4, 391–424.

—— —— (2007*c*). Judgment Aggregation with Consistency Alone. Working Paper, London School of Economics.

—— —— (2007*d*). Strategy-Proof Judgment Aggregation. *Economics and Philosophy*, 23/3, 269–300.

Dokow, E., and Holzman, R. (2005). Aggregation of Binary Evaluations. *Journal of Economic Theory*, forthcoming.

—— —— (2006). Aggregation of Binary Evaluations with Abstentions. Working Paper, Technion Israel Institute of Technology.

Eckert, D., and Klamler, C. (2007). How Puzzling is Judgment Aggregation? Antipodality in Distance-Based Aggregation Rules. Working Paper, University of Graz.

Gärdenfors, P. (2006). An Arrow-Like Theorem for Voting with Logical Consequences. *Economics and Philosophy*, 22, 181–90.

Guilbaud, G. Th. (1966). Theories of the General Interest, and the Logical Problem of Aggregation. In P. F. Lazarsfeld and N. W. Henry (eds.), *Readings in Mathematical Social Science*, 262–307. Cambridge, MA: MIT Press.

Kasher, A., and Rubinstein, A. (1997). On the Question "Who is a J", a Social Choice Approach. *Logique et Analyse*, 160, 385–95.

Kemeny, J. (1959). Mathematics without Numbers. *Daedalus*, 88/4, 577–91.
Konieczny, S., and Pino-Perez, R. (2002). Merging Information under Constraints: A Logical Framework. *Journal of Logic and Computation*, 12, 773–808.
Kornhauser, L. A., and Sager, L. G. (1986). Unpacking the Court. *Yale Law Journal*, 96/1, 82–117.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
List, C. (2001). Mission Impossible? The Problem of Democratic Aggregation in the Face of Arrow's Theorem. D. Phil. thesis, University of Oxford.
—— (2003). A Possibility Theorem on Aggregation over Multiple Interconnected Propositions. *Mathematical Social Sciences*, 45/1, 1–13 (Corrigendum in *Mathematical Social Sciences*, 52, 109–10).
—— (2004). A Model of Path Dependence in Decisions over Multiple Propositions. *American Political Science Review*, 98/3, 495–513.
—— (2005). The Probability of Inconsistencies in Complex Collective Decisions. *Social Choice and Welfare*, 24/1, 3–32.
—— (2008). Which Worlds are Possible? A Judgment Aggregation Problem. *Journal of Philosophical Logic*, 37, 57–65.
—— and Pettit, P. (2002). Aggregating Sets of Judgments: An Impossibility Result. *Economics and Philosophy*, 18, 89–110.
—— —— (2004). Aggregating Sets of Judgments: Two Impossibility Results Compared. *Synthese*, 140/1–2, 207–35.
Merlin, V., and Saari, D. (2000). A Geometric Examination of Kemeny's Rule. *Social Choice and Welfare*, 17/3, 403–38.
Miller, A. (2007). Group Identification. Working Paper, CalTech.
Miller, M. (2007). Judgment Aggregation and Subjective Decision-Making. *Economics and Philosophy*, forthcoming.
Mongin, P. (2005). Factoring Out the Impossibility of Logical Aggregation. *Journal of Economic Theory*, forthcoming.
Nehring, K. (2003). Arrow's Theorem as a Corollary. *Economics Letters*, 80/3, 379–82.
—— (2005). The Impossibility of a Paretian Rational. Working Paper, University of California at Davis.
—— (2006). Oligarchies in Judgment Aggregation. Working Paper, University of California at Davis.
—— (2007). The Impossibility of a Paretian Rational: A Bayesian Perspective. *Economics Letters*, 96, 45–50.
—— and Puppe, C. (2002). Strategy-Proof Social Choice on Single-Peaked Domains: Possibility, Impossibility and the Space Between. Working Paper, University of California at Davis.
—— —— (2005*a*). Consistent Judgement Aggregation: The Truth-Functional Case. *Social Choice and Welfare*, forthcoming.
—— —— (2005*b*). The Structure of Strategy-Proof Social Choice. Part II: Non-Dictatorship, Anonymity and Neutrality. Working Paper, University of Karlsruhe.
—— —— (2007*a*). Justifiable Group Choice. Working Paper, University of Karlsruhe.
—— —— (2007*b*). The Structure of Strategy-Proof Social Choice. Part I: General Characterization and Possibility Results on Median Spaces. *Journal of Economic Theory*, 135, 269–305.

Pauly, M. (2007). Axiomatizing Collective Judgement Sets in a Minimal Logical Language. *Synthese*, 158, 233–50.

—— and Van Hees, M. (2006). Logical Constraints on Judgment Aggregation. *Journal of Philosophical Logic*, 35, 569–85.

Pettit, P. (2001). Deliberative Democracy and the Discursive Dilemma. *Philosophical Issues*, 11, 268–99.

Pigozzi, G. (2006). Belief Merging and the Discursive Dilemma: An Argument-Based Account to Paradoxes of Judgment Aggregation. *Synthese*, 152/2, 285–98.

Rubinstein, A., and Fishburn, P. (1986). Algebraic Aggregation Theory. *Journal of Economic Theory*, 38, 63–77.

Sen, A. K. (1970). The Impossibility of a Paretian Liberal. *Journal of Political Economy*, 78, 152–7.

Van Hees, M. (2007). The Limits of Epistemic Democracy. *Social Choice and Welfare*, 28/4, 649–66.

Wilson, R. (1975). On the Theory of Aggregation. *Journal of Economic Theory*, 10, 89–99.

CHAPTER 20

POPULATION ETHICS

CHARLES BLACKORBY
WALTER BOSSERT
DAVID DONALDSON

20.1 Introduction

The traditional framework of social choice theory as initiated by Arrow (1951) addresses the issue of aggregating profiles of individual preference relations into a social preference relation. One way of escape from the negative conclusion of his impossibility theorem consists of expanding the informational base of collective choice by assuming that individual preferences are represented by utility functions and allowing for interpersonal comparisons of utility, thereby moving away from the narrow confines of Arrow's assumption of ordinal measurability and interpersonal noncomparability; see, for instance, d'Aspremont and Gevers (1977); Hammond (1979); and Sen (1977) for possibility results and characterizations under various informational assumptions. An extensive survey of the literature on social choice with interpersonal utility comparisons is provided by Bossert and Weymark (2004).

We thank a referee for comments and suggestions. Financial support through grants from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

Most of the literature on social choice theory treats the population as fixed, and the notion of variable population social evaluation has its origins in applied ethics. In particular, Parfit (1976, 1982, 1984) introduced the axiomatic approach to population ethics, and his contribution continues to be one of the most influential in the area; see, for instance, Ryberg and Tännsjö (2004). The approach we follow is welfarist: in order to compare any two alternatives (whose populations may differ), the only information required consists of the sets of those alive in the respective alternatives and their lifetime utilities. The extension of fixed population social evaluation methods to a variable population context is important, because many public policy decisions involve endogenous populations. For instance, when determining public spending on prenatal care, on foreign aid packages with population consequences, or on intergenerational resource allocation, the assumption that the population is fixed is difficult to justify. Therefore, more comprehensive criteria are called for.

Following the usual convention in population ethics, we assume that utilities represent individual lifetime well-being and are normalized so that a lifetime utility of zero represents *neutrality.* Above neutrality, a life, as a whole, is worth living; below neutrality, it is not. From the viewpoint of an individual, a neutral life is a life which is as good as one in which the person has no experiences; see, for instance, Broome (1993; 2004, ch. 8); Heyd (1992, ch. 1); McMahan (1996); and Parfit (1984, app. G) for discussions. People who do not exist do not have interests or preferences, and therefore we take the view that it is not possible to say that an individual can gain (or lose) by being brought into existence with a utility level above (or below) neutrality. Someone who is alive might have an attitude, such as a desire or preference, toward a world in which the person does not exist, but that attitude could hardly be construed as individual betterness or worseness. Similarly, a person who is alive and expresses satisfaction with her or his existence (that is, with having been born) cannot be claiming that existence is better (for him or her) than nonexistence. Note that this does not prevent an individual from gaining or losing by continuing to live—the continuation or termination of life is a matter of length of life, not existence itself.

Sen (1987, p. 11) criticizes welfarism on the grounds that "the battered slave, the broken unemployed, the hopeless destitute, the tamed housewife, may have the courage to desire little". Because we recommend accounts of well-being, such as those of Griffin (1986) and Sumner (1996), that include all aspects of well-being, whether they accord with preferences or not, this difficulty does not arise.

A commonly used principle is *classical utilitarianism*, also referred to as *total utilitarianism.* It ranks any two alternatives by comparing the total utilities of the individuals alive in them. Parfit (1976; 1982; 1984, ch. 19) observed that classical utilitarianism leads to the *repugnant conclusion.* A population principle implies the repugnant conclusion if every alternative in which everyone alive experiences

a utility level above neutrality is ranked as worse than an alternative in which each member of a larger population has a utility level that is above neutrality but may be arbitrarily close to it. This means that population size can always be used to substitute for quality of life as long as lives are (possibly barely) worth living. As Parfit's analysis demonstrates, the repugnant conclusion is implied by any population principle that (i) declares the *ceteris paribus* addition of an individual above neutrality to a given population to be a social improvement; (ii) ranks any alternative with an equal utility distribution as at least as good as any alternative involving the same population, the same total utility, but an unequal distribution of well-being; and (iii) ranks same population, equal utility alternatives by declaring the one with a higher common utility level to be better. We provide a formal statement and proof of this result in Section 20.3.

Although avoidance of the repugnant conclusion is regarded as of paramount importance by most, some have argued that it should be accepted. For a discussion, see Tännsjö (2002).

For any alternative, the *critical level* of utility is that level which, if experienced by an added person without changing the utilities of the existing population, leads to an alternative which is as good as the original. Clearly, the choice of critical levels has important consequences for the properties of a population principle and is closely linked to the possibility of avoiding the repugnant conclusion.

Average utilitarianism uses average, rather than total, utility to rank alternatives. It does not imply the repugnant conclusion, but has other defects, such as declaring the *ceteris paribus* addition of an individual with a lifetime utility well below neutrality desirable as long as the existing population's average utility is even lower. Thus, other population principles are called for, and avoidance of the repugnant conclusion has become an important criterion that acceptable principles should satisfy. We believe that the *critical-level utilitarian* principles with positive critical levels and their generalized counterparts are the most satisfactory; see Blackorby, Bossert, and Donaldson (2005*b*) and Blackorby and Donaldson (1984). Critical-level utilitarianism is a one-parameter family of principles. The parameter is a *fixed* critical level of utility that applies to all alternatives, and the criterion used to rank the alternatives is the sum of the differences between individual utilities and the critical level. *Critical-level generalized utilitarianism* uses transformed utilities and a transformed critical level, thereby allowing for inequality aversion in individual well-being: if the transformation is (strictly) concave, the resulting principle is (strictly) inequality-averse. All critical-level utilitarian principles are also critical-level generalized utilitarian principles.

Due to space limitations, we cannot go beyond a brief introduction to the subject and refer the reader to Blackorby, Bossert, and Donaldson (2005*b*) for an extensive treatment. We focus on critical-level generalized utilitarian principles because we consider those with positive critical levels to be the most suitable for social evaluation. In addition to characterizing these and other critical-level principles, we

discuss an impossibility result as an example of the conflicts that arise in population ethics.

Section 20.2 introduces a welfarist and anonymous approach to population ethics, along with the population principles that are of interest in this survey. Section 20.3 illustrates the dilemmas in population ethics by means of an impossibility result. In Section 20.4, we provide a characterization of critical-level generalized utilitarianism and three of its subclasses. Some issues that are not addressed in the previous sections are discussed briefly in Section 20.5. Section 20.6 concludes.

20.2 Variable Population Anonymous Welfarism

We use $\mathcal{R}$ to denote the set of all real numbers, and $\mathcal{Z}_{++}$ is the set of positive integers. $\mathbf{1}_n$ is the vector consisting of $n \in \mathcal{Z}_{++}$ ones. Suppose there is a set of alternatives X. Each element $x \in X$ is a full description of all relevant aspects of the world, including the identities of everyone alive in x and everything that may affect a person's lifetime well-being. We assume that, for each possible (finite but arbitrarily large) population, there are at least three alternatives with that population. Potential individuals are indexed by positive integers, and for an individual $i \in \mathcal{Z}_{++}$, X_i is the subset of X consisting of all alternatives x such that i is alive in x. We use $N(x) \subseteq \mathcal{Z}_{++}$ to denote the nonempty and finite set of those alive in alternative $x \in X$. An individual utility function for i is a mapping $U_i : X_i \to \mathcal{R}$, interpreted as an indicator of lifetime well-being. Thus, for $x \in X_i$, $U_i(x)$ is i's lifetime utility in alternative x. Note that the domain of U_i is X_i, and therefore i's well-being is defined only for alternatives in which the person exists. A profile of utility functions is an infinite-dimensional vector $U = (U_1, U_2, \ldots, U_i, \ldots)$ containing one utility function for each potential person. The set of all possible utility profiles is $\mathcal{U}$, and we use $U(x)$ to denote the vector of utilities of those alive in $x \in X$; that is, $U(x) = (U_i(x))_{i \in N(x)}$.

A social evaluation functional $F : \mathcal{D} \to \mathcal{O}$ assigns a social ordering of the alternatives to each possible profile in its nonempty domain $\mathcal{D} \subseteq \mathcal{U}$, where $\mathcal{O}$ is the set of all orderings defined on X. An ordering is a reflexive, complete, and transitive binary relation, the social ordering $R_U = F(U)$ is interpreted as a goodness relation, and $x R_U y$ is interpreted to mean that x is socially at least as good as y. The better-than relation corresponding to R_U is P_U, and I_U is the equal goodness relation.

Although it is possible to prove a welfarism theorem without assuming anonymity (see Blackorby, Bossert, and Donaldson 2005*b*), we assume, in this

chapter, that the social evaluation functional is anonymous in the sense that the individual identities are irrelevant—only lifetime utilities and population sizes can influence the social ranking.

If a social evaluation functional is anonymous and welfarist, there is a single anonymous ordering R of all possible utility vectors $\Omega = \bigcup_{n \in \mathcal{Z}_{++}} \mathcal{R}^n$, and alternative $x \in X$ is at least as good as alternative $y \in X$ given the profile U if and only if the utility vector $U(x)$ is at least as good as the utility vector $U(y)$ according to R. P and I are the betterness and equal goodness relations corresponding to R. R is anonymous if and only if, for all $n \in \mathcal{Z}_{++}$ and for all $u, v \in \mathcal{R}^n$ such that u is a permutation of v, $u I v$. Because of this property, we can, without loss of generality, assume that if there are $n \in \mathcal{Z}_{++}$ individuals alive in an alternative, they are the individuals labeled 1 to n. Thus, knowledge of the number of individuals alive in two alternatives and their lifetime utilities is sufficient to rank any two alternatives for any profile.

To state the relevant variable population version of the welfarism theorem, the following axioms are imposed on the social evaluation functional; see Blackorby, Bossert, and Donaldson (2005*b*, ch. 3) for a discussion.

Unlimited Domain: $\mathcal{D} = \mathcal{U}$.

Pareto Indifference: For all $x, y \in X$ such that $N(x) = N(y)$ and for all $U \in \mathcal{D}$, if $U(x) = U(y)$, then $x I_U y$.

Binary Independence of Irrelevant Alternatives: For all $x, y \in X$ and for all $U, V \in \mathcal{D}$, if $U(x) = V(x)$ and $U(y) = V(y)$, then

$$x R_U y \Leftrightarrow x R_V y.$$

Anonymity: For all $x, y \in X$ and for all $U \in \mathcal{D}$, if $U(x)$ is a permutation of $U(y)$, then $x I_U y$.

We obtain

Theorem 1. Suppose a social evaluation functional F satisfies unlimited domain. F satisfies Pareto indifference, binary independence of irrelevant alternatives, and anonymity if and only if there exists an anonymous social evaluation ordering R on Ω such that, for all $x, y \in X$ and for all $U \in \mathcal{D}$,

$$x R_U y \Leftrightarrow U(x) R\, U(y).$$

See Blackorby, Bossert, and Donaldson (2005*b*, ch. 3) for a proof of Theorem 1. Fixed population versions of the welfarism theorem are discussed in Blackorby, Bossert, and Donaldson (2005*a*); Blau (1976); Bossert and Weymark (2004); d'Aspremont and Gevers (1977); Guha (1972); Hammond (1979); and Sen (1977).

We conclude this section by introducing the population principles that are of particular interest in this chapter. For the definition of these principles, it is important to keep in mind that neutrality is normalized to a lifetime utility level of

zero. For other normalizations, the formulation of the principles has to be amended accordingly; see, for instance, Dasgupta (1993), who, somewhat unconventionally, uses a negative utility level to represent neutrality.

The first principle we define is classical utilitarianism, which ranks utility vectors (and thus alternatives) on the basis of the total utilities obtained in them. According to classical utilitarianism,

$$uRv \;\Leftrightarrow\; \sum_{i=1}^{n} u_i \geq \sum_{i=1}^{m} v_i$$

for all $n, m \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $v \in \mathcal{R}^m$.

Average utilitarianism employs average utilities instead of total utilities for social evaluation. Thus, the average utilitarian principle is defined by

$$uRv \;\Leftrightarrow\; \frac{1}{n}\sum_{i=1}^{n} u_i \geq \frac{1}{m}\sum_{i=1}^{m} v_i$$

for all $n, m \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $v \in \mathcal{R}^m$.

Critical-level utilitarianism with a parameter value of $\alpha \in \mathcal{R}$ generalizes classical utilitarianism by replacing utilities with the differences between utilities and the critical-level parameter. This leads to the principle defined by

$$uRv \;\Leftrightarrow\; \sum_{i=1}^{n}[u_i - \alpha] \geq \sum_{i=1}^{m}[v_i - \alpha]$$

for all $n, m \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $v \in \mathcal{R}^m$. Clearly, classical utilitarianism is obtained for $\alpha = 0$.

All of the above principles produce identical fixed population comparisons: namely, those corresponding to utilitarianism.

Generalizations are obtained by replacing utility levels (including critical levels) with transformed utilities. Letting $g : \mathcal{R} \to \mathcal{R}$ be a continuous and increasing function such that $g(0) = 0$, the classical generalized utilitarian principle corresponding to g is defined by

$$uRv \;\Leftrightarrow\; \sum_{i=1}^{n} g(u_i) \geq \sum_{i=1}^{m} g(v_i)$$

for all $n, m \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $v \in \mathcal{R}^m$; average generalized utilitarianism is given by

$$uRv \;\Leftrightarrow\; \frac{1}{n}\sum_{i=1}^{n} g(u_i) \geq \frac{1}{m}\sum_{i=1}^{m} g(v_i)$$

for all $n, m \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $v \in \mathcal{R}^m$; and the critical-level generalized utilitarian principle for α and g is

$$uRv \Leftrightarrow \sum_{i=1}^{n}[g(u_i) - g(\alpha)] \geq \sum_{i=1}^{m}[g(v_i) - g(\alpha)]$$

for all $n, m \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $v \in \mathcal{R}^m$.

Again, fixed population comparisons are the same for all of the generalized principles—they reduce to those according to generalized utilitarianism. If g is (strictly) concave, the resulting principle is (strictly) inequality-averse with respect to lifetime well-being.

20.3 An Impossibility Result

There are numerous impossibility results in the population ethics literature that establish the incompatibility of seemingly plausible axioms. The purpose of this section is to illustrate this observation by means of an impossibility theorem (Blackorby, Bossert, and Donaldson 2006*b*). The axioms that follow are employed.

A weakening of the well-known weak Pareto principle is obtained if social betterness is required whenever one equal utility distribution strictly dominates another equal utility distribution. We call this axiom minimal increasingness; it is satisfied by all population principles introduced in Section 20.2.

Minimal Increasingness: For all $n \in \mathcal{Z}_{++}$ and for all $a, b \in \mathcal{R}$, if $a > b$, then $a\mathbf{1}_n P b\mathbf{1}_n$.

Weak inequality aversion requires that, for any given population and any given total utility, the equal distribution is at least as good as any unequal distribution. The axiom is satisfied by all of the generalized principles (critical-level utilitarian as well as average) associated with a concave transformation g.

Weak Inequality Aversion: For all $n \in \mathcal{Z}_{++}$ and for all $u \in \mathcal{R}^n$, $\left((1/n)\sum_{i=1}^{n} u_i\right)\mathbf{1}_n Ru$.

Sikora (1978) suggests extending the standard Pareto principle to nonexisting individuals, an axiom he calls the Pareto plus principle. It is usually defined as the conjunction of strong Pareto (defined formally in the following section) and the requirement that the addition of an individual above neutrality to a utility-unaffected population is a social improvement. Because the full force of strong Pareto is not required (our impossibility theorem stated below merely assumes minimal increasingness), we retain strong Pareto as a separate axiom and define Pareto plus as follows.

Pareto Plus: For all $n \in \mathcal{Z}_{++}$, for all $u \in \mathcal{R}^n$, and for all $a > 0$, $(u, a)Pu$.

In the axiom statement, the common population in u and (u, a) is unaffected. To defend the axiom, therefore, it must be argued that a life above neutrality is socially better than nonexistence. To interpret the axiom as a Pareto condition, existence must therefore be regarded as *individually* better than nonexistence. Thus, the axiom must be based on the implausible idea that people who do not exist have interests that should be respected. Pareto plus is satisfied by all critical-level generalized utilitarian principles with nonpositive critical levels. Average generalized utilitarianism and critical-level generalized utilitarian principles with positive critical levels do not possess this property.

We follow Parfit (1984) in considering the repugnant conclusion an unacceptable property of a population principle. Thus, our final axiom in this section requires that this conclusion be avoided. A population principle implies the repugnant conclusion if and only if, for any population size $n \in \mathcal{Z}_{++}$, any positive utility level ξ, and any utility level $\epsilon \in (0, \xi)$, there exists a population size $m > n$ such that an m-person alternative in which every individual experiences utility level ϵ is ranked as better than an n-person society in which every individual's utility level is ξ. The axiom that requires the repugnant conclusion to be avoided is defined as follows:

Avoidance of the Repugnant Conclusion: There exist $n \in \mathcal{Z}_{++}$, $\xi \in \mathcal{R}_{++}$, and $\epsilon \in (0, \xi)$ such that, for all $m > n$, $\xi \mathbf{1}_n R \epsilon \mathbf{1}_m$.

As is straightforward to verify, critical-level generalized utilitarianism satisfies Pareto plus if and only if the critical level α is nonpositive and satisfies avoidance of the repugnant conclusion if and only if α is positive. Thus, no critical-level generalized utilitarian principle can satisfy Pareto plus and at the same time avoid the repugnant conclusion. However, this incompatibility extends well beyond these principles. As an illustration, we reproduce an impossibility theorem due to Blackorby, Bossert, and Donaldson (2006*b*). In particular, we show that all minimally increasing and weakly inequality-averse population principles that satisfy Pareto plus lead to the repugnant conclusion. Similar theorems can be found in Arrhenius (2000); Blackorby and Donaldson (1991); Blackorby, Bossert, and Donaldson (2005*b*); Blackorby, Bossert, Donaldson, and Fleurbaey (1998); Carlson (1998); and Ng (1989).

Theorem 2. There exists no anonymous social evaluation ordering R that satisfies minimal increasingness, weak inequality aversion, Pareto plus, and avoidance of the repugnant conclusion.

Proof: Suppose R satisfies minimal increasingness, weak inequality aversion, and Pareto plus. We complete the proof by showing that R must imply the repugnant

conclusion. For any population size $n \in \mathcal{Z}_{++}$, let ξ, ϵ, and δ be utility levels such that $0 < \delta < \epsilon < \xi$. Choose the positive integer r such that

$$r > n\frac{(\xi - \epsilon)}{(\epsilon - \delta)}. \tag{1}$$

Because the numerator and the denominator on the right side of this inequality are both positive, r is positive. By repeated application of Pareto plus and transitivity, $(\xi \mathbf{1}_n, \delta \mathbf{1}_r) P \xi \mathbf{1}_n$. Average utility in $(\xi \mathbf{1}_n, \delta \mathbf{1}_r)$ is $(n\xi + r\delta)/(n + r)$ so, by weak inequality aversion, $[(n\xi + r\delta)/(n + r)]\mathbf{1}_{n+r} R(\xi \mathbf{1}_n, \delta \mathbf{1}_r)$. By (1),

$$\epsilon > \frac{n\xi + r\delta}{n + r}$$

and, by minimal increasingness, $\epsilon \mathbf{1}_{n+r} P [(n\xi + r\delta)/(n + r)]\mathbf{1}_{n+r}$. Using transitivity, it follows that $\epsilon \mathbf{1}_{n+r} P \xi \mathbf{1}_n$, and letting $m = n + r > n$, avoidance of the repugnant conclusion is violated. □

If weak inequality aversion is dropped from the list of axioms in the theorem, the remaining axioms are compatible. For example, *geometrism*, a principle proposed by Sider (1991), satisfies all axioms other than weak inequality aversion. The principle uses a constant $k \in (0, 1)$ and ranks alternatives with a weighted sum of utilities, where the weights are such that the jth-highest nonnegative utility level receives a weight of k^{j-1} and the ℓth-lowest negative utility receives a weight of $k^{\ell-1}$. Critical levels are all zero, and the repugnant conclusion is avoided, but, because weights on higher positive utilities exceed weights on lower ones, the principle prefers inequality of positive utilities over equality (see Arrhenius and Bykvist 1995).

Anonymity is not required for the impossibility result of Theorem 2, but is included in the theorem statement in order to use the same framework throughout. See Blackorby, Bossert, Donaldson, and Fleurbaey (1998) for a proof without anonymity.

20.4 A Characterization of Critical-Level Generalized Utilitarianism

Critical-level generalized utilitarianism can be characterized by means of a set of plausible and intuitively appealing axioms. The first of these applies to fixed population comparisons only. It is the well-known strong Pareto requirement which demands that unanimity be respected.

Strong Pareto: For all $n \in \mathcal{Z}_{++}$ and for all $u, v \in \mathcal{R}^n$, if $u_i \geq v_i$ for all $i \in \{1, \ldots, n\}$ with at least one strict inequality, then $u P v$.

The standard definition of strong Pareto encompasses Pareto indifference, requiring that if everyone in a fixed population has the same level of well-being in two alternatives, the two should be ranked as equally good. In our welfarist framework, this property is automatically satisfied because the relation R is reflexive.

Our second axiom is another fixed population requirement. Continuity requires that small changes in utilities should not lead to large changes in social ranking.

Continuity: For all $n \in \mathcal{Z}_{++}$ and for all $u \in \mathcal{R}^n$, the sets $\{v \in \mathcal{R}^n \mid vRu\}$ and $\{v \in \mathcal{R}^n \mid uRv\}$ are closed in $\mathcal{R}^n$.

Existence of a critical level is an axiom regarding the comparison of alternatives with different population sizes, insuring that at least some nontrivial tradeoffs between population size and well-being are possible. It requires the existence of a critical level for at least one utility vector. Critical levels need not exist for other utility vectors, and if they do, they need not be constant. Thus, the axiom is very weak.

Existence of a Critical Level: There exist $\bar{u} \in \Omega$ and $c \in \mathcal{R}$ such that $(\bar{u}, c) I \bar{u}$.

Strong Pareto, continuity, and existence of a critical level are satisfied by all of the principles introduced in Section 20.2. In contrast, the final axiom used in our characterization is satisfied by all critical-level generalized utilitarian principles, but violated by average utilitarianism and its generalized counterpart. Existence independence is a separability axiom that applies not only to fixed population comparisons but also to those involving different populations. It requires the social ranking to be independent of the existence of the unconcerned—individuals who are not affected by the ranking of two alternatives. See d'Aspremont and Gevers (1977), for example, for a fixed population version of this independence property.

Existence Independence: For all $u, v, w \in \Omega$, $(u, w)R(v, w) \Leftrightarrow uRv$.

These axioms characterize critical-level generalized utilitarianism, as established in the following theorem:

Theorem 3. An anonymous social evaluation ordering R satisfies strong Pareto, continuity, existence of a critical level, and existence independence if and only if R is critical-level generalized utilitarian.

Proof: That critical-level generalized utilitarianism satisfies the required axioms is straightforward to verify. To prove the reverse implication, consider first the case of a fixed population size $n \geq 3$. Applying Debreu's (1959, pp. 56–9) representation theorem, continuity implies the existence of a continuous function $f^n : \mathcal{R}^n \to \mathcal{R}$

such that, for all $u, v \in \mathcal{R}^n$,

$$uRv \;\Leftrightarrow\; f^n(u) \geq f^n(v).$$

By strong Pareto, f^n is increasing in all arguments, and the anonymity of R implies that f^n is symmetric (that is, invariant with respect to permutations of its arguments). Existence independence implies that $\{1, \ldots, n\} \setminus M$ is separable in f^n from its complement M for any choice of M such that $\emptyset \neq M \subset \{1, \ldots, n\}$. Gorman's (1968) theorem on overlapping separable sets of variables (see also Aczél 1966, p. 312, and Blackorby, Primont, and Russell 1978, p. 127) implies that f^n is additively separable. Therefore, there exist continuous and increasing functions $H^n: \mathcal{R} \to \mathcal{R}$ and $g_i^n: \mathcal{R} \to \mathcal{R}$ for all $i \in \{1, \ldots, n\}$ such that

$$f^n(u) = H^n\left(\sum_{i=1}^{n} g_i^n(u_i)\right)$$

for all $u \in \mathcal{R}^n$. Because f^n is symmetric, each g_i^n can be chosen to be independent of i, and we define $g^n = g_i^n$ for all $i \in \{1, \ldots, n\}$. Therefore, because f^n is a representation of the restriction of R to $\mathcal{R}^n$ and H^n is increasing,

$$\begin{aligned} uRv &\Leftrightarrow H^n\Big(\sum_{i=1}^{n} g^n(u_i)\Big) \geq H^n\Big(\sum_{i=1}^{n} g^n(v_i)\Big) \\ &\Leftrightarrow \sum_{i=1}^{n} g^n(u_i) \geq \sum_{i=1}^{n} g^n(v_i) \end{aligned} \tag{2}$$

for all $u, v \in \mathcal{R}^n$. Without loss of generality, g^n can be chosen so that $g^n(0) = 0$.

Next, we prove that there exists a utility level $\alpha \in \mathcal{R}$ which is a critical level for all utility vectors in Ω. Let $u \in \Omega$ be arbitrary. By existence of a critical level, there exist $\bar{u} \in \Omega$ and $c \in \mathcal{R}$ such that $(\bar{u}, c)I\bar{u}$. Applying existence independence twice, we obtain $(u, \bar{u}, c)I(u, \bar{u})$ and $(u, c)Iu$. Thus, c is a critical level not only for $\bar{u}$ but also for any $u \in \Omega$. Letting $\alpha = c$ establishes the claim.

Now we show that, for all $n \geq 3$, the functions g^n and g^{n+1} can be chosen to be the same. Let $u, v \in \mathcal{R}^n$. Because $\alpha \in \mathcal{R}$ is a critical level for all utility vectors in Ω, we have

$$uRv \;\Leftrightarrow\; (u, \alpha)R(v, \alpha). \tag{3}$$

By (2),

$$uRv \;\Leftrightarrow\; \sum_{i=1}^{n} g^n(u_i) \geq \sum_{i=1}^{n} g^n(v_i) \tag{4}$$

and

$$(u, \alpha)R(v, \alpha) \Leftrightarrow \sum_{i=1}^{n} g^{n+1}(u_i) + g^{n+1}(\alpha) \geq \sum_{i=1}^{n} g^{n+1}(v_i) + g^{n+1}(\alpha)$$

$$\Leftrightarrow \sum_{i=1}^{n} g^{n+1}(u_i) \geq \sum_{i=1}^{n} g^{n+1}(v_i). \tag{5}$$

Therefore, using (3), (4), and (5),

$$\sum_{i=1}^{n} g^{n}(u_i) \geq \sum_{i=1}^{n} g^{n}(v_i) \;\Leftrightarrow\; \sum_{i=1}^{n} g^{n+1}(u_i) \geq \sum_{i=1}^{n} g^{n+1}(v_i),$$

which means that the same function can be used for g^n and for g^{n+1}. Because this is true for all $n \geq 3$, it follows that the functions g^n can be chosen independently of n, and we write $g = g^n$ for all $n \geq 3$. Together with (2), it follows that, for all $n \geq 3$ and for all $u, v \in \mathcal{R}^n$,

$$uRv \;\Leftrightarrow\; \sum_{i=1}^{n} g(u_i) \geq \sum_{i=1}^{n} g(v_i). \tag{6}$$

Next, we prove that (6) must be true for $n \in \{1, 2\}$ as well. Let $u, v \in \mathcal{R}^1$. By strong Pareto and the increasingness of g,

$$uRv \;\Leftrightarrow\; u_1 \geq v_1 \Leftrightarrow g(u_1) \geq g(v_1). \tag{7}$$

If $u, v \in \mathcal{R}^2$, existence independence and (6) together imply

$$uRv \;\Leftrightarrow\; (u, \alpha)R(v, \alpha) \Leftrightarrow \sum_{i=1}^{2} g(u_i) + g(\alpha) \geq \sum_{i=1}^{2} g(v_i) + g(\alpha)$$

$$\Leftrightarrow \sum_{i=1}^{2} g(u_i) \geq \sum_{i=1}^{2} g(v_i). \tag{8}$$

(6), (7), and (8) imply that all fixed population comparisons are carried out according to generalized utilitarianism with the same transformation for all population sizes.

To complete the proof, let $n, m \in \mathcal{Z}_{++}$ with $n \neq m$, $u \in \mathcal{R}^n$, and $v \in \mathcal{R}^m$. Suppose $n > m$. By definition of the critical level α,

$$uRv \Leftrightarrow uR(v, \alpha\mathbf{1}_{n-m}) \Leftrightarrow \sum_{i=1}^{n} g(u_i) \geq \sum_{i=1}^{m} g(v_i) + (n - m)g(\alpha)$$

$$\Leftrightarrow \sum_{i=1}^{n} [g(u_i) - g(\alpha)] \geq \sum_{i=1}^{m} [g(v_i) - g(\alpha)].$$

An analogous argument applies to the case $n < m$ and it follows that R is critical-level generalized utilitarian. □

As mentioned earlier, adding Pareto plus (respectively avoidance of the repugnant conclusion) to the axioms of Theorem 3 leads to a characterization of the subclass of critical-level generalized utilitarian principles the members of which have a nonpositive (respectively positive) critical level. These observations are summarized in the following two theorems:

Theorem 4. An anonymous social evaluation ordering R satisfies strong Pareto, continuity, existence of a critical level, existence independence, and Pareto plus if and only if R is critical-level generalized utilitarian with a nonpositive critical-level parameter.

Theorem 5. An anonymous social evaluation ordering R satisfies strong Pareto, continuity, existence of a critical level, existence independence, and avoidance of the repugnant conclusion if and only if R is critical-level generalized utilitarian with a positive critical-level parameter.

Because we consider the repugnant conclusion as unacceptable and the axioms of Theorem 3 as obviously appealing, Theorems 4 and 5 suggest that Pareto plus should be abandoned. Furthermore, we advocate weakly inequality-averse principles satisfying the axioms of Theorem 3, and as a consequence, we recommend the critical-level generalized utilitarian principles with a positive critical level α and a concave utility transformation g characterized in our final theorem. Note that, because the identity map is concave, the critical-level utilitarian principles are weakly inequality-averse.

Theorem 6. An anonymous social-evaluation ordering R satisfies weak inequality aversion, strong Pareto, continuity, existence of a critical level, existence independence, and avoidance of the repugnant conclusion if and only if R is critical-level generalized utilitarian with a positive critical-level parameter and a concave utility transformation.

20.5 Some Issues and Extensions

In this section, we address several additional issues that are not discussed above. Each is examined in Blackorby, Bossert, and Donaldson (2005*b*). Some are present in both fixed population and variable population environments; others appear in variable population environments only.

20.5.1 Utility Measurement and Interpersonal Comparisons

If utilities are ordinally measurable and interpersonally noncomparable, Arrow's (1951) theorem, appropriately modified, leads to an impossibility. Utilities can be assumed to be numerically measurable and interpersonally comparable in order to allow for the largest class of principles. Although this assumption is strong, if utilities are cardinally measurable (unique up to increasing affine transformations) and interpersonally comparable at two utility levels, full numerical comparability results from choosing utility numbers for the two levels.

20.5.2 The Neutrality Normalization

We follow the standard practice in the literature and assign a utility level of zero to neutrality. The idea of neutrality is not necessary for many theorems, including Theorems 2 and 3. Indeed, Dasgupta (1988, 1993) and Hammond (1988) do without neutrality and normalize the critical level in critical-level generalized utilitarianism to zero. Such a normalization is not without its difficulties, however. If critical levels change, individual utilities must also change.

20.5.3 One or Many Profiles

It can be argued that, when comparing complete histories, multiple profiles are inappropriate. Although the single-profile approach is less well developed than the multi-profile approach, we have argued that a richness condition on the set of alternatives together with adapted versions of axioms such as anonymity are sufficient to make the results of the multi-profile case apply in the single-profile environment (Blackorby, Bossert, and Donaldson 2006*a*).

20.5.4 Dynamics

The model presented in this article can be modified to accommodate multiple time periods. If this is done, Pareto indifference rules out discounting of future lifetime utilities. That axiom can be modified to allow discounting of lifetime utilities, however, by making it conditional on birth dates.

Sometimes, population principles are applied to single periods using per period utilities. If this is done and critical levels are not zero, difficulties arise. Suppose, for example, that a person lives one period longer in alternative x than in y with a utility level of zero in the additional period, all else the same. If a per period utility level of zero represents neutrality in the period, every person is equally well off in the two alternatives from the timeless perspective, Pareto indifference requires x

and y to be ranked as equally good, and consistency between per period rankings and the timeless ranking requires the critical level to be zero for the per period ranking. See Blackorby, Bossert, and Donaldson (1995, 1997) for discussions of this intertemporal model.

20.5.5 Uncertainty

The critical-level generalized utilitarian principles can be used to rank actions or combinations of institutional arrangements (including legal and educational ones), customs, and moral rules, taking account of the constraints of history and human nature. If each of these leads with certainty to a particular social alternative, they can be ranked by means of any welfarist principle. But consequences may be uncertain and, in that case, probabilities may be assigned to outcomes, and the resulting uncertain alternatives ranked with extended population principles. One class of such principles, which can be justified axiomatically, consists of the *ex ante* critical-level utilitarian principles; see Blackorby, Bossert, and Donaldson (2005*b*, ch. 7; 2007). These principles employ value functions that are equivalent to the critical-level utilitarian value functions applied to expected utilities.

20.5.6 Incomplete Rankings

There are population principles which declare alternatives to be unranked in some circumstances. One such class of principles is the critical-band generalized utilitarian class, which uses an interval (the band). Two alternatives are ranked if and only if one is declared better than the other by *all* critical-level generalized utilitarian principles with critical levels in the band (see Blackorby, Bossert, and Donaldson 1996; 2005*b*, ch. 7).

20.5.7 Choice Functions

Because many policy decisions have population consequences, it is natural to use population principles to guide them. These decision problems are, in most cases, choice problems: one or more options must be selected from a set of feasible alternatives. The maximizing approach to solving choice problems requires the selection of a best feasible alternative, according to a social ranking. Although this is a reasonable way to proceed, it excludes consideration of choice procedures that are not based on social orderings from the outset.

A natural way to proceed is to focus on choice functions and ask whether the choices can be rationalized by a social ordering. Axioms must therefore be employed that apply to choices rather than rankings of alternatives. This is a complex

problem, but it is possible to find a set of such axioms that characterizes a choice-theoretic version of critical-level generalized utilitarianism (Blackorby, Bossert, and Donaldson 2002; 2005*b*, ch. 10).

20.6 Concluding Remarks

This survey provides but a brief introduction to the many issues that arise in population ethics. There are numerous other principles that have been suggested and analyzed in the literature. For example, number-dampened principles, their restricted counterparts, and restricted versions of critical-level principles (see Blackorby, Bossert, and Donaldson 2005*b*, ch. 5; Hurka 1983, 2000; and Ng 1986) fail to satisfy existence independence, whereas variable population versions of leximin become possible if continuity is dropped as a requirement.

There are open questions in population ethics. Some principles for fixed population social evaluation, such as that corresponding to the the Gini social evaluation function (see Blackorby and Donaldson 1984) are not additively separable. It is not known whether they can be extended to population problems in a reasonable way.

References

Aczél, J. (1966). *Lectures on Functional Equations and their Applications*. New York: Academic Press.

Arrhenius, G. (2000). An Impossibility Theorem for Welfarist Axiologies. *Economics and Philosophy*, 16, 247–66.

——and Bykvist, K. (1995). Future Generations and Interpersonal Compensations. Uppsala Prints and Preprints in Philosophy no. 21, Uppsala University.

Arrow, K. (1951). (2nd edn. 1963). *Social Choice and Individual Values*. New York: Wiley.

Blackorby, C., and Donaldson, D. (1984). Social Criteria for Evaluating Population Change. *Journal of Public Economics*, 25, 13–33.

——— (1991). Normative Population Theory: A Comment. *Social Choice and Welfare*, 8, 261–7.

——Bossert, W., and Donaldson, D. (1995). Intertemporal Population Ethics: Critical-Level Utilitarian Principles. *Econometrica*, 63, 1303–20.

——— ——— (1996). Quasi-Orderings and Population Ethics. *Social Choice and Welfare*, 13, 129–50.

——— ——— (1997). Birth-Date Dependent Population Ethics: Critical-Level Principles. *Journal of Economic Theory*, 77, 260–84.

——— ——— (2002). Rationalizable Variable-Population Choice Functions. *Economic Theory*, 19, 355–78.

——— ——— ——— (2005*a*). Multi-Profile Welfarism: A Generalization. *Social Choice and Welfare*, 24, 253–67; Erratum 25, 227–8.
——— ——— ——— (2005*b*). *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.
——— ——— ——— (2006*a*). Anonymous Single-Profile Welfarism. *Social Choice and Welfare*, 27, 279–87.
——— ——— ——— (2006*b*). Population Ethics and the Value of Life. In M. McGillivray (ed.), *Inequality, Poverty and Well-being*, 8–21. Basingstoke: Palgrave Macmillan.
——— ——— ——— (2007). Variable-Population Extensions of Social Aggregation Theorems. *Social Choice and Welfare*, 28, 567–89.
——— ——— ——— and Fleurbaey, M. (1998). Critical Levels and the (Reverse) Repugnant Conclusion. *Journal of Economics*, 67, 1–15.
——— Primont, D., and Russell, R. (1978). *Duality, Separability, and Functional Structure: Theory and Economic Applications*. Amsterdam: North-Holland.
Blau, J. (1976). Neutrality, Monotonicity, and the Right of Veto: A Comment. *Econometrica*, 44, 603.
Bossert, W., and Weymark, J. (2004). Utility in Social Choice. In S. Barberà, P. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory, ii: Extensions*, 1099–177. Dordrecht: Kluwer.
Broome, J. (1993). Goodness is Reducible to Betterness: The Evil of Death is the Value of Life. In P. Koslowski (ed.), *The Good and the Economical: Ethical Choices in Economics and Management*, 69–83. Berlin: Springer-Verlag.
——— (2004). *Weighing Lives*. Oxford: Oxford University Press.
Carlson, E. (1998). Mere Addition and the Two Trilemmas of Population Ethics. *Economics and Philosophy*, 14, 283–306.
Dasgupta, P. (1988). Lives and Well-Being. *Social Choice and Welfare*, 5, 103–26.
——— (1993). *An Inquiry into Well-Being and Destitution*. Oxford: Clarendon Press.
d'Aspremont, C., and Gevers, L. (1977). Equity and the Informational Basis of Collective Choice. *Review of Economic Studies*, 44, 199–209.
Debreu, G. (1959). *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.
Gorman, W. (1968). The Structure of Utility Functions. *Review of Economic Studies*, 32, 369–90.
Griffin, J. (1986). *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.
Guha, A. (1972). Neutrality, Monotonicity, and the Right of Veto. *Econometrica*, 40, 821–6.
Hammond, P. (1979). Equity in Two Person Situations: Some Consequences. *Econometrica*, 47, 1127–35.
——— (1988). Consequentialist Demographic Norms and Parenting Rights. *Social Choice and Welfare*, 5, 127–46.
Heyd, D. (1992). *Genethics: Moral Issues in the Creation of People*. Berkeley: University of California Press.
Hurka, T. (1983). Value and Population Size. *Ethics*, 93, 496–507.
——— (2000). Comment on "Population Principles with Number-Dependent Critical Levels", Unpublished MS, University of Calgary, Department of Philosophy.
McMahan, J. (1996). Wrongful Life: Paradoxes in the Morality of Causing People to Exist. Unpublished MS, University of Illinois.

Ng, Y.-K. (1986). Social Criteria for Evaluating Population Change: An Alternative to the Blackorby–Donaldson Criterion. *Journal of Public Economics*, 29, 375–81.

—— (1989). What Should We Do about Future Generations? Impossibility of Parfit's Theory X. *Economics and Philosophy*, 5, 235–53.

Parfit, D. (1976). On Doing the Best for our Children. In M. Bayles (ed.), *Ethics and Population*, 100–2. Cambridge: Schenkman.

—— (1982). Future Generations, Further Problems. *Philosophy and Public Affairs*, 11, 113–72.

—— (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Ryberg, J., and Tännsjö, T. (eds.) (2004). *The Repugnant Conclusion: Essays on Population Ethics*. Dordrecht: Kluwer.

Sen, A. (1977). On Weights and Measures: Informational Constraints in Social Welfare Analysis. *Econometrica*, 45, 1539–72.

—— (1987). *The Standard of Living*. Cambridge: Cambridge University Press.

Sider, T. (1991). Might Theory X be a Theory of Diminishing Marginal Value? *Analysis*, 51, 202–13.

Sikora, R. (1978). Is it Wrong to Prevent the Existence of Future Generations? In R. Sikora and B. Barry (eds.), *Obligations to Future Generations*, 112–66. Philadelphia: Temple University Press.

Sumner, L. (1996). *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press.

Tännsjö, T. (2002). Why we Ought to Accept the Repugnant Conclusion. *Utilitas*, 14, 339–59. Repr. in Ryberg and Tännsjö (2004), 219–37.

CHAPTER 21

DISTRIBUTIVE JUSTICE

AN OVERVIEW OF EXPERIMENTAL EVIDENCE

WULF GAERTNER

21.1 Introduction

In his *Nicomachean Ethics*, Aristotle wrote that "both the unjust man and the unjust act are unfair or unequal, and clearly in each case of inequality there is something intermediate, viz., that which is equal. . . . Then if what is unjust is unequal, what is just is equal." Aristotle continued by saying:

> a just act necessarily involves at least four terms: two persons for whom it is in fact just, and two shares in which its justice is exhibited. And there will be the same equality between the shares as between the persons, because the shares will be in the same ratio to one another as the persons; for if the persons are not equal, they will not have equal shares; and it is when equals have or are assigned unequal shares, or people who are not equal, equal shares, that quarrels and complaints break out. (1976, pp. 177–8)

Helpful suggestions by the editors of this Handbook are gratefully acknowledged.

Aristotle viewed justice as a kind of proportion. "What is just . . . is what is proportional, and what is unjust is what violates the proportion. So one share becomes too large and the other too small. This is exactly what happens in practice: the man who acts unjustly gets too much and the victim of injustice too little of what is good" (p. 179).

We shall see in the following sections that proportionality of some sort between giving and receiving or between contributions and rewards plays a role in formulations of distributive justice, at least in certain situations. Equality of something yet to be defined can be considered as an ideal. However, there are various reasons for a departure from equality. Yaari and Bar-Hillel (1984, p. 7) identify several broad categories, among them (1) differences in needs, (2) differences in tastes, (3) differences in beliefs, and (4) differences in effort and contribution. These and other aspects will be examined in this chapter.

Bar-Hillel and Yaari were probably the first to study the concept of justice or just distribution via "judgments of justice", elicited from hypothetical questions. More concretely, the authors gave students hypothetical distribution problems and asked them "to solve them justly" (Bar-Hillel and Yaari 1993, p. 59). They emphasize that the focus of their research is the ethical notions in people's minds, not their actual behavior, keeping in mind that actual behavior "is inevitably contaminated by political, strategic, and other considerations" (p. 59). They add that "it is people's expressed sentiments (namely what they say ought to be done) rather than their revealed ones (namely what they actually do) that primarily guides the search for a *normative* theory of justice, as well as the rhetoric of public debate on issues of distributive justice" (p. 59). Clearly, intuitions about specific situations can mingle with theoretical conceptions, but at the end of this process of deliberation, there is hopefully some state of equilibrium that Rawls (1971) referred to as a "reflective equilibrium".

The general public has an opinion on issues of distributive justice. This view may sometimes be quite vague; it may depend, as we shall see in this survey, on the particular context in which the actual problem is embedded. It may also be culture-dependent, and may change over time. But it definitely exists and should be taken into consideration in a political democracy. Schokkaert (1999) argues that if normative economics wants its analysis to have real influence on the decisions taken within a political system, it has to consider the opinions and preferences of its citizens. The political system itself has to explain its ideas regarding justice and fair distribution to the members of society. Otherwise, public support for a particular distributive policy and its implementation is doubtful. Empirical research may help to find out what is going on in the minds of people.

Clearly, there are various problems when setting up questionnaire experiments. The phrasing, for example, has to be chosen very carefully. An extra word (e.g. "luckily" or "unfortunately", "healthy" or "unhealthy", "dangerous" or "safe") can easily achieve manipulative force which the experimenter should try to avoid.

Admittedly, sometimes the manipulative power of a word is discovered only *ex post*, too late for this particular investigation. Should respondents (in most experiments they are students) be acquainted with an underlying theory? Perhaps not (and this is the case in our own questionnaire studies), because then there is the danger that the students see the experiment as a test to check whether they really understood the theory behind the questions, with the consequence that the whole study may turn into an IQ test. And this is clearly not what the experimenter was looking for.

In the following sections, we shall discuss, based on questionnaire-experimental investigations, aspects of needs (Section 21.2), aspects of effort and productivity (Section 21.3), the function of the veil of ignorance (Section 21.4), and, in the concluding section (Section 21.5), various other aspects that seem to play a role.

21.2 Needs

We said in our introduction that reasons have to be given for departing from equality. Do differences in needs provide sufficient grounds for a departure from equal division? One may argue that needs bestow on recipients an entitlement that is, whenever feasible, proportionately responsive to their need.

Yaari and Bar-Hillel (1984) and Bar-Hillel and Yaari (1993) studied the following situation, in which they asked students to divide a bundle of goods between two persons in order for the division to be "just".

Q 2.1 A shipment containing 12 grapefruit and 12 avocados is to be distributed between Jones and Smith. The following information is given, and is known also to the two recipients:

— Doctors have determined that Jones's metabolism is such that his body derives 100 milligrammes of vitamin F from each grapefruit consumed, while it derives no vitamin F whatsoever from avocado.

— Doctors have also determined that Smith's metabolism is such that his body derives 50 milligrammes of vitamin F from each grapefruit consumed and also from each avocado consumed.

— Both persons, Jones and Smith, are interested in the consumption of grapefruit and/or avocados only insofar as such consumption provides vitamin F—and the more the better. All the other traits of the two fruits (such as taste, calorie content, etc.) are of no consequence to them.

— No trades can be made after the division takes place.

How should the fruits be divided between Jones and Smith, if the division is to be just?

Table 21.1. Q 2.1, n = 163

Distribution	% of respondents
J: 6, 6; S: 6, 6	8
J: 6, 0; S: 6, 12	0
J: 8, 0; S: 4, 12	82
J: 9, 0; S: 3, 12	8
J: 12, 0; S: 0, 12	2

This problem of dividing grapefruit and avocados can be expressed more succinctly, or more technically, in the following way. Let ω be the bundle of fruits to be divided between Jones and Smith so that we have $\omega = (12, 12)$. Jones and Smith have different abilities to metabolize the fruits into vitamins. Therefore, we shall write $u_J(x, y) = 100x$ for Jones and $u_S(x, y) = 50x + 50y$ for Smith, with x and y being quantities of grapefruit and avocados, respectively. The functions u_J and u_S can be interpreted purely technically. We shall, however, view them as utility functions of the two persons. Moreover, these functions can be interpreted as cardinal utility functions with the property that the units of measurement (milligrammes of vitamin) are comparable across the individuals.

How did the students divide the given bundle of twelve grapefruit and twelve avocados between Jones and Smith? Yaari and Bar-Hillel presented five different distributions to young male and female applicants for admission to Hebrew University of Jerusalem in the years 1978 to 1980. The respondents were confronted with two versions of question Q 2.1. One version asked the students to mark which of the five distributions *they* considered as the most just. The other version asked the respondents to assess how Jones and Smith would divide the shipment "on the assumption that both recipients are committed to looking for a just division" (Yaari and Bar-Hillel 1984, p. 10, n. 10). The authors report that differences between the distributions of responses to these two versions were negligible.

Table 21.1 provides answers where (J: 9, 0; S: 3, 12), for example, means that Jones gets nine grapefruit and no avocados, while Smith receives three grapefruit and twelve avocados. In all tables that follow, n stands for the number of observations.

Strict equality of fruit is supported only by a minority. Equal split of the number of fruit is also Pareto-inefficient. A division which takes account of differing degrees of metabolic efficiency and yields an equal amount of vitamins, if possible, is favored by a large majority. Actually, mechanisms as diverse as Rawlsian maximin (1971) and bargaining from zero according to Kalai and Smorodinsky (1975) "support" the distribution (J: 8, 0; S: 4, 12).

Table 21.2. Q 2.2, $n = 146$

Distribution	% of respondents
J: 6, 6; S: 6, 6	4
J: 4, 0; S: 8, 12	82
J: 6, 0; S: 6, 12	4
J: 8, 0; S: 4, 12	7
J: 12, 0; S: 0, 12	3

We now come to an issue that Yaari and Bar-Hillel have called "tenability". The wording of the authors is extremely cautious.

> We are prepared to interpret the numbers . . . as saying, for example, that the distribution (J: 8, 0; S: 4, 12) is much more in agreement with moral intuition than, say, the distribution (J: 12, 0; S: 0, 12). . . . Indeed, it would be hard to make a case for a distribution mechanism that picks the distribution (J: 12, 0; S: 0, 12) . . . without explaining why this distribution should fare so badly in an experimental setting designed to trace out prevailing moral intuitions. (1984, p. 10)

What happens when metabolic efficiency decreases considerably in one of the two persons? To answer this question, Yaari and Bar-Hillel modified the original situation Q 2.1 in the following way.

Substitute the third paragraph in Q 2.1 by

> Q 2.2 — Doctors have also determined that Smith's metabolism is such that his body derives 20 milligrammes of vitamin F from each grapefruit consumed and also from each avocado consumed.

The "only" change from Q 2.1 is that Smith's metabolism is less effective than originally. In technical terms, the problem now reads:

$$\omega = (12, 12);$$
$$u_J(x, y) = 100x;$$
$$u_S(x, y) = 20x + 20y.$$

Maximin supported as mentioned (J: 8, 0; S: 4, 12) in Q 2.1 and now advocates (J: 4, 0; S: 8, 12), where, again, the vitamin intake of the two persons is equalized. The authors note that maximin compensates Smith for the deterioration in his metabolism. The results for Q 2.2 are shown in Table 21.2 (the respondents were, of course, different from the ones who had answered Q 2.1).

The students' "vote" in favor of a maximin-supported division is very strong, both in absolute terms and in relation to the other proposals. Yaari and Bar-Hillel remark that one might, perhaps, have expected this, given the fact that the problem

presented to the students isolated the issue of needs and, furthermore, needs were readily quantifiable.

What happens when Smith experiences a further deterioration in his metabolism? Would respondents be willing to further compensate Smith as they did in Q 2.2? Bar-Hillel and Yaari found that maximin now loses much of its previous attractiveness. Other proposals become more popular; even the highly insensitive equal-split solution now gets a reasonable amount of support.

We now turn to one of our own investigations. The focus is on an equity axiom that is fundamental for Rawls's second principle of justice. This equity axiom was formulated in different forms by Sen (1973), Hammond (1976), and Deschamps and Gevers (1978), among others. Rawls's second principle requires that economic and social inequalities be arranged such that they are to the greatest benefit of the least advantaged members of society. The equity axiom makes a particular demand for a society of only two individuals or, more generally, for a society where only two individuals are affected by a change from one policy to another. Let there be two policies x and y. We postulate that person 1 prefers x to y, person 2 prefers y to x, and independently of whether x or y will eventually be the social outcome, person 2 is always better off than person 1. We know that in such a situation, the equity axiom requires x to be socially preferred to y.

Is it possible to check whether individuals follow this axiom in their judgments (check in an indirect way, of course; to ask people directly would be rather naïve)? The question I wish to discuss is twofold. First of all, we would like to know whether people's evaluations satisfy the demands of the equity principle. In a second step, we will ask whether those who fulfill this axiom would follow it *unconditionally*, i.e. focus always exclusively on the worst-off members of society. This step is somewhat related to the latter part of the Yaari and Bar-Hillel investigation, where Smith's metabolism became poorer and poorer.

How can we check for a fulfillment of the equity axiom? In Gaertner (1992), I made the following suggestion. Let us consider the subsequent two-person profile of so-called extended orderings $\tilde{R}_i$, $i \in \{1, 2\}$, that I shall denote by E^1.

$$\tilde{R}_1 : \ (y, 2)(x, 2)(x, 1)(y, 1),$$
$$\tilde{R}_2 : \ (y, 2)(x, 2)(x, 1)(y, 1).$$

These lines should be read as follows. Both individuals agree that it is best to be person 2 under policy y. This is deemed better than being person 2 under policy x. This, again, is better than being person 1 under x, which is better than being person 1 under y. The reader should verify that this two-person profile mirrors the structure of the equity axiom just stated. Both persons diverge in their evaluations of policies x and y as far as *their own position* is concerned, but they agree that it is person 2 who is always better off.

According to the equity axiom, x will be declared as preferable to y. We shall now enlarge this basic profile by adding the extended orderings of persons 3, 4, . . . ,

thereby preserving the structure of E^1. E^2, for example, is:

$$\tilde{R}_1 : (y, 3)(x, 3)(y, 2)(x, 2)(x, 1)(y, 1),$$
$$\tilde{R}_2 : (y, 3)(x, 3)(y, 2)(x, 2)(x, 1)(y, 1),$$
$$\tilde{R}_3 : (y, 3)(x, 3)(y, 2)(x, 2)(x, 1)(y, 1).$$

We then ask all members of society how they would wish to resolve the situations E^1, E^2, ... All those individuals who accept the equity axiom will, of course, say that for E^1 alternative x should be the preferred state. For a moment, let us focus on just one member of the society. Will he or she find x also preferable in situation E^2? If "yes", will the same verdict hold in E^3, E^4, ... ? It is entirely possible that at some point in this successive questioning the individual wishes to switch from "x preferable to y" to "now y should be preferred to x socially". It could also be the case, however, that given the size of the society, the evaluating member of society would always want x to be socially preferred to y, and thus follow the equity axiom unconditionally.

The situation that I shall present and discuss now can be found on the Internet[1] together with several other cases. This situation as well as the others was presented to classes of undergraduate students at the University of Osnabrück between 1989 and 2002. All students were enrolled in economics or business administration. At the time of the investigation the students had not yet had a course on welfare economics and theories of distributive justice, such as utilitarianism, Rawlsianism, and game-theoretical solutions.

Here is the situation we wish to focus on.

Q 2.3 (o) A small society has received a certain amount of money which can be used either to provide some help and assistance for a handicapped person or to further the education of an intelligent child. The child could receive a good education in languages and in natural sciences, let's say. Let the handicapped person be person 1; if the sum of money were used for her support (alternative x), she would be able to learn some very basic things, so that at least in certain areas of daily life she would no longer be totally dependent on the assistance from other people. Let the intelligent child be person 2; the investment into its education represents alternative y. The interpersonal welfare ranking reads:

$$(y, 2)(x, 2)(x, 1)(y, 1)$$

Which alternative should be realized in your view, x or y?

[1] The Internet address is <http://nts4.oec.uni-osnabrueck.de/mikro/darp.pdf>. All in all, we gave six different situations to the students. All these situations are fully reproduced in Gaertner and Jungeilges (2002). We should mention that in Osnabrück, we had two versions of our questionnaire, a technical and a non-technical version (the technical version is reproduced here and on the Internet). The non-technical version did not use the specification in terms of extended orderings but provided a somewhat lengthier verbal description of the same "facts" instead. Of course, each student only saw one version. Table 21.3 gives the results from the non-technical version only. The Osnabrück results for the two versions did not show any difference on the basis of a two-sample nonparametric test, given an error probability of 5 percent.

(a) Imagine that the sum of money which could be used to help the handicapped person, is so large that, on the other hand, this amount would suffice for the education of not only person 2 but also a second child (person 3) who is even somewhat more intelligent than person 2. Person 3 would, therefore, benefit even a bit more from the education so that the following interpersonal welfare ranking can be assumed:

$$(y, 3)(y, 2)(x, 3)(x, 2)(x, 1)(y, 1)$$

Would you choose x or y under these conditions?

(b) Imagine that if the money were used to finance alternative y it would be possible to educate still another child (person 4). The reason may simply be "economies of scale" or the fact that a talented teacher will be able to provide a good education for several children simultaneously. Let us assume that all the other characteristics of the situation remain as before. The interpersonal welfare ranking now reads:

$$(y, 4)(y, 3)(y, 2)(x, 4)(x, 3)(x, 2)(x, 1)(y, 1)$$

Which alternative should be picked in your view, x or y?

(c) Add another child to the situation (person 5), who could also receive an instruction in languages and the natural sciences out of the given budget. Everything else remains the same and the interpersonal welfare ranking reads:

$$(y, 5)(y, 4)(y, 3)(y, 2)(x, 5)(x, 4)(x, 3)(x, 2)(x, 1)(y, 1)$$

Would you want x or y to be realized?

The underlying issue apparently is to allocate a certain amount of money to provide some help for a handicapped person (alternative x) or to teach one (or several) intelligent child(ren). Clearly, the intelligent child(ren) is (are) always better off than the handicapped person whatever decision is taken.

Our students most likely played the role of an external judge. In other words, their identification with the position and the circumstances of a particular person was only of an indirect nature. On second thoughts, however, this need not necessarily have been the case. Imagine that a student himself (herself) turned out to be handicapped or that one member of his (her) family or a close friend suffered from a handicap. We do not know this, of course, but had it been the case, it would certainly have mattered.

In Table 21.3, we give the results for the Osnabrück students during the period 1989–2002. In this table, 0 always represents the choice of alternative x, 1 stands for the choice of alternative y. To be more explicit, the sequence 0000, for example, refers to those students who took a decision in favor of x in all cases, i.e. in the basic situation and in all of its variants. The sequences 0001, 0011, and 0111 represent the verdicts of those respondents who decided at one point to revise their original judgment. The numbers in the columns give the percentages of answers within each of the cohorts of undergraduates. Relative frequencies of a revision or "switch" are contained in the lower part of the table. All those sequences which begin with

Table 21.3. Q 2.3

Sequence	Year of investigation				
	1989 *n* = 65	1990 *n* = 93	1993 *n* = 81	1994 *n* = 63	2002 *n* = 86
0 0 0 0	0.723	0.581	0.494	0.603	0.407
0 0 0 1	0.046	0.086	0.062	0.016	0.035
0 0 1 0	0.0	0.0	0.0	0.0	0.0
0 0 1 1	0.077	0.151	0.148	0.095	0.174
0 1 0 0	0.0	0.0	0.0	0.0	0.0
0 1 0 1	0.0	0.0	0.0	0.0	0.0
0 1 1 0	0.0	0.0	0.0	0.0	0.012
0 1 1 1	0.077	0.086	0.173	0.143	0.233
1 0 0 0	0.0	0.0	0.0	0.0	0.0
1 0 0 1	0.0	0.0	0.0	0.0	0.0
1 0 1 0	0.0	0.0	0.0	0.0	0.0
1 0 1 1	0.0	0.0	0.0	0.0	0.0
1 1 0 0	0.0	0.011	0.0	0.0	0.0
1 1 0 1	0.0	0.0	0.0	0.0	0.0
1 1 1 0	0.0	0.0	0.0	0.0	0.0
1 1 1 1	0.077	0.086	0.123	0.143	0.140
% switch	19.8	32.1	38.3	25.4	44.2
% fulfillment of equity axiom	92.3	90.3	87.7	85.7	86.0

0 represent students who satisfied the equity axiom. Correspondingly, all those sequences which start with 1 hint at a violation of the equity axiom. The percentages of students who satisfied the equity axiom are given at the bottom of the table.

Let us try to interpret these findings. We start with the year 1989. The decision to give the money to the handicapped person in all cases, i.e. unconditionally, was very strong indeed (72.3 percent). Only 7.7 percent of the respondents wanted the amount of money to go into the education of the intelligent child(ren) right away. Those who wished to revise their original decision which, at the outset, was in favor of helping the handicapped were 19.8 percent of the students. The percentages of those who wanted to revise their decision after the first or second "round" were equally high (7.7 percent). All in all, the equity axiom was fulfilled by 92.3 percent of the respondents.

When we now examine the following years, we have to state that the percentages for unconditional support of the handicapped have gone down more or less continually. At the same time, unconditional support for an education of the child(ren) as well as the desire to switch already after the first round (the latter from 7.7 percent in 1989 to 23.3 percent in 2002) experienced a steady increase over the years. All these developments are reflected in a steady decline of the fulfillment of the equity axiom

and in a considerable increase of the desire to revise a decision originally made (the latter from 19.8 percent in 1989 to 44.2 percent in 2002).

These tendencies or differences, rather, that evolved over time were checked statistically by using a chi-squared test with the H_o hypothesis of an identical distribution of the responses between any two cohorts (years). The results of these tests are such that the H_o hypothesis was rejected at the 5 percent significance level between the cohorts of 1989 and 1993 and between 1989 and 2002. Furthermore, the H_o hypothesis was rejected at the 10 percent level between the years 1994 and 2002.

The situation depicted above was given to students in other countries. Gaertner *et al.* (2001) ran their questionnaire studies in Austria, the Baltics, Israel, and Slovenia, among other countries. The Israeli results turned out to be quite close to the German figures; the results from the Baltic countries, however, were vastly different. Austria and Slovenia were somewhere in between. This means, but I say this with utmost caution, that the social, political, and historical contexts seem to matter. This should be no surprise at all on second thoughts.

21.3 Effort, Productivity, and Pareto Efficiency

In Section 21.2 the issue was to distribute a given bundle of goods or a certain amount of financial resources among a group of people (two or more than two (groups of) persons). In this section, I wish to consider what happens in the context of production where individuals make contributions and exercise effort to a smaller or higher degree, depending on their physical and mental abilities. So there are perceived inputs and perceived outputs. What would be a "just" or "fair" allocation of produced commodities? This question was also taken up in the recent philosophical literature (see e.g. Arneson 1989, 1990; Cohen 1989, 1990; and Dworkin 1981). Konow (2001) argued that one should distinguish between discretionary and exogenous variables. A discretionary variable affects production and can be controlled or influenced by the person considered (like work effort). An exogenous variable can have an influence on the amount or quality of output but cannot, under normal circumstances, be influenced by the person (e.g. some physical disability). Konow proposed an accountability principle in this context, calling for allocations to be in proportion to volitional contributions, meaning that "a worker who is twice as productive as another should be paid twice as much if the higher productivity is due to greater work effort but not if it is due to innate aptitude" (Konow 2001, p. 138). Thus, Konow argues, "individuals are only held accountable for factors they may reasonably control" (p. 142).

Do respondents share this view? Konow gave the following question plus variations to students at his university.

Q 3.1 Bob and John are identical in terms of physical and mental abilities. They become shipwrecked on an uninhabited island where the only food is bananas. They can collect as many bananas as they want by climbing up a tree, picking them before they fall into the ocean and throwing them into a pile. In this way, Bob picks 12 bananas per day and John picks 8 per day. Bob takes from the pile the 12 bananas he picked leaving John with the 8 that John picked. Please rate this as:

Q 3.1 – $n = 76$

	% of respondents
fair	74 %
unfair	26 %

Q 3.2 Same background. However:
Bob and John are identical in terms of physical and mental abilities except that Bob was born with one hand and John with two. Together they pick a total of 20 bananas per day, but because of his condition Bob picks fewer bananas per day than John. John takes 12 bananas from the pile leaving 8 for Bob. Please rate this as:

Q 3.2 – $n = 78$

	% of respondents
fair	19 %
unfair	81 %

Q 3.3 Same background as in Q 3.2. However: John takes 10 bananas from the pile leaving 10 for Bob. Please rate this as:

Q 3.3 – $n = 78$

	% of respondents
fair	90 %
unfair	10 %

In the initial situation, i.e. Q 3.1, there are no explicit exogenous differences between the two persons; the only difference is of a discretionary type, i.e. harvesting bananas. So, according to accountability, Bob should get twelve, John should receive eight bananas. There is a wide agreement among the respondents to support this view.

In Q 3.2, the greater productivity of two-handed John is not viewed as sufficient ground for granting him twelve bananas from the pile, leaving only eight to disabled Bob. Eighty-one percent of the respondents view the larger share of John as unfair. Since Bob is in no way responsible for his disability, an equal split of the harvest, as suggested in Q 3.3, is seen as fair by an overwhelming majority of the respondents (90 percent). Since in Q 3.2 and Q 3.3, there are no differences in the perceived discretionary variables, an unequal allocation is deemed highly unfair, and an equal split as truly fair.

Let us now introduce a twist towards larger differences in productivity. The background story is the same as above, but the productivity of one of the two

persons is going up sharply, while the second person's productivity is constantly decreasing. This questionnaire study is based on telephone interviews that Konow undertook in the Los Angeles area.

Q 3.4 Bob and John become shipwrecked on an uninhabited island. The only food is bananas which the castaways collect and throw into a pile daily. Bob and John are identical in terms of abilities and work effort except that Bob was born with only one hand and John with two. John picks 14 bananas per day while Bob can pick only 6 because of his condition. John takes 12 bananas from the pile leaving 8 bananas for Bob. Please rate this as:

Q 3.4 – $n = 117$

	% of respondents
fair	17 %
unfair	83 %

Q 3.5 Same as in Q 3.4. However: John picks 16 bananas per day while Bob can pick only 4 because of his condition . . . Please rate this as:

Q 3.5 – $n = 121$

	% of respondents
fair	28 %
unfair	72 %

Q 3.6 Same as in Q 3.4. However: John picks 18 bananas per day while Bob can pick only 2 because of his condition . . . Please rate this as:

Q 3.6 – $n = 109$

	% of respondents
fair	39 %
unfair	61 %

Note that in all three situations, the total size of the allocable resource is the same. The results are in conformity with the accountability principle insofar as a vast majority in Q 3.4 and still a large majority in Q 3.5 and Q 3.6 deemed the exogenous differences between Bob and John as irrelevant. However, with John's productivity rising, the opposition to the unequal allocation of twelve bananas for John and eight for Bob is getting less pronounced. In an indirect sense, the widening productivity gap is being honored. This observation is to some degree reminiscent of the Yaari and Bar-Hillel results, where students were not willing to compensate Smith for his declining metabolism *ad infinitum*.

Pareto efficiency gets more pronounced in the following situation, also considered in Konow (2001).

Q 3.7 A: A small newly independent island nation is considering how to allocate its one banana plantation and its one sugar plantation. There are only two farmers in the island interested in these plantations. The government chooses among the following two plans either of which would result in the same total production of both bananas and sugar.

Plan X: Both farmers receive one-half of each plantation. Each farmer earns an average profit of US$100 per day from bananas and sugar combined. Therefore, the total of both farmers' profits is US$200 per day.

Plan Y: One farmer receives the banana plantation and the other farmer receives the sugar plantation. The average daily profit of the banana farmer is US$150 and that of the sugar farmer is also US$150. At US$300 per day, combined profits are greater under this plan because specialization reduces production costs.

Please circle the plan that you consider more fair:

Q 3.7A – $n = 147$

	% of respondents
Plan X	20 %
Plan Y	80 %

B: A variation of Plan Y. The first sentence is the same. However: The farmers' profits are unequal since the sugar plantation is more profitable than the banana plantation: average daily profit of the banana farmer is US$100 and that of the sugar farmer is US$200. At US$300 per day, combined profits are . . . (same text as under A).

Please circle the plan that you consider more fair:

Q 3.7B – $n = 132$

	% of respondents
Plan X	57 %
Plan Y	43 %

Note that while there is the same total production in variants A and B, the profits of the farmers differ. They differ both between Plan X and Plan Y and between Plan Y in version A and Plan Y in version B.

In version A, there is a clear efficiency gain for farmers due to a specialization in production, and this is widely acknowledged by the respondents, again university students in this part of the investigation. In both Plans X and Y, the farmers seem equally accountable, and since the profits are equal under the Pareto superior Plan Y, the accountability principle wins the day. In version B, the total daily profits under Plan Y are still the same as in version A, but the profits of the two farmers have become unequal, thus violating the accountability principle. So, obviously, a conflict between accountability and Pareto efficiency arises. A majority of the students is now in favor of Plan X, which is Pareto-inefficient but satisfies the accountability principle. The sharp fall in support of Plan Y is stunning. Konow's results show that the rather intuitive (as economists often say) Pareto efficiency criterion is not that cherished after all.

Beckman *et al.* (2002) also investigated to what extent Pareto efficiency is generally accepted. In their experiments, they considered one variant where individuals knew their own income position and another variant where this was not known. The study was run in the USA, Taiwan, Russia, and China. The Pareto improvement that was considered consisted of giving additional income to one out of five

positions, where the position which benefited changed from round to round. Across nations, only 10.1 percent opposed Pareto improvements when positions were not known. Opposition to Pareto gains increased to 18.3 percent when the recipient of this gain was, income-wise, below the person who was asked to give her view. Opposition increased to 28.8 percent when the recipient had a higher position than the evaluating person. There were large differences across countries. In the USA, opposition to a Pareto gain was only 6.4 percent, in China and Russia, opposition to a Pareto improvement went up to roughly 20 percent. When all income positions gained, though four out of five positions only slightly, opposition to a Pareto improvement went down to 10 percent again.

Amiel and Cowell (1994) also found large opposition to Pareto improvements in their questionnaire studies, in which the respondents were acting as impartial observers with no stake in the outcome. Klemisch-Ahlert (1992) has evidence for opposition against Pareto improvements in her bargaining experiments. She concluded that envy plays a role in the distribution of money in her experimental studies; in other words, envy may generate payoff agreements that are not strongly Pareto-optimal.

21.4 Behind the Veil of Ignorance or Outside the Veil

Traub *et al.* (2005) investigated the evaluation of income distributions for two different roles in which the evaluating person may find herself and for two different information scenarios. The two different roles are those of an outside observer with no stakes, called an umpire, and of a person who becomes an income recipient within her most favored income distribution, once the veil of ignorance has been lifted, called the self-concern mode. The two informational scenarios are the case of "ignorance", where it is assumed that only the set of possible incomes is known, while there is absolutely no information on probabilities, and the case of "risk", where agents know both the possible incomes and the probability distribution over these incomes.

Combining the self-concern mode and the umpire mode with the two scenarios of ignorance and risk gives four different situations. The authors wanted the students to participate in all four cases. The experiment lasted for two hours. There were material incentives for the students which, to some degree, depended on their choices in the course of the experiment. Sixty-one students were involved in this study, mostly from economics and business administration. The authors assert that there were no gender differences in the students' responses. The sequence in which

Table 21.4. Ignorance scenario: Q 4.1

No.	Income set
1	{59,000 110,000}
2	{60,000}
3	{40,000 45,000 50,000 55,000 60,000}
4	{30,000 150,000}
5	{30,000 180,000}
6	{20,000 50,000 100,000 150,000 220,000}
7	{20,000 60,000 100,000 160,000 220,000}
8	{0 100,000 220,000 250,000}
9	{10,000 20,000 30,000 40,000 45,000 50,000 55,000 60,000 80,000 90,000 100,000}

the four cases were given to the students was not varied among the participants. So it was not possible to check for order effects—for example, learning.

The ignorance scenario, depicted in Table 21.4, consisted of nine income sets. These sets represented eligible entries in income distributions. Respondents were told that the eventual income distributions were made up only by using components of these sets, so that not all components of these sets necessarily entered the ensuing income distribution. This was said in order to destroy any connotation of probabilities. The ignorance scenario was devised to mimic a Rawlsian setup. The reader will recall that his difference principle was designed without using the concept of probability.

The risk scenario, shown in Table 21.5, consisted of twelve income distributions, each of which contained exactly five entries representing income quintiles. This scenario was designed to mimic a Harsanyi-type (1953, 1955) environment. In all twelve distributions, agents know both the possible incomes and their probability distributions.

The respondents were required to state complete preference orderings over both the nine income sets in the ignorance scenario (Q 4.1) and the twelve income distributions in the risk scenario (Q 4.2). One of the focuses in this paper was to investigate differences in the respondents' behavior under the self-concern mode and under the mode of an external observer. A second focus was whether and to what degree the respondents' orderings came close to one of the standards of behavior, such as the Rawlsian maximin principle or its lexicographic variant, Harsanyi's utilitarian criterion, the Gini ranking, or some other standard of behavior. The authors focus in particular on a hybrid standard which they call Boulding's principle, where the realization of a floor constraint is combined with the maximization of expected utility.

What were the results? The authors found that under the self-concern mode, income set 7 was ranked highest, with income set 6 winning the second position

Table 21.5. Risk scenario: Q 4.2

No.	Income distribution
1	{60,000 60,000 60,000 60,000 60,000}
2	{50,000 55,000 60,000 65,000 70,000}
3	{40,000 50,000 60,000 70,000 80,000}
4	{40,000 40,000 60,000 80,000 80,000}
5	{40,000 60,000 60,000 60,000 80,000}
6	{10,000 20,000 60,000 100,000 110,000}
7	{10,000 60,000 60,000 60,000 110,000}
8	{70,000 70,000 100,000 110,000 120,000}
9	{70,000 70,000 70,000 90,000 180,000}
10	{15,000 15,000 100,000 110,000 120,000}
11	{15,000 15,000 70,000 90,000 180,000}
12	{0 60,000 80,000 250,000 250,000}

(note that set 7 weakly vector-dominates set 6). Income set 8, where one of the incomes is zero, fared very badly. Under the umpire mode, sets 7 and 6 were still at the top, but they lost in terms of mean rank. Interestingly, income set 8 gained five rank positions. Subjects, in their role as outside observer, seem to have thought that the possibility of rather high incomes under set 8 compensates the society for the chance of a zero income. "However, when possibly affected by a zero income under the self-concern mode, they shied away from income set 8" (Traub *et al.*, p. 296). While the first three ranking positions are taken by the same income sets under both modes, sets 5 and 3 lose two ranking positions when going from the first to the second mode. The outside observer apparently now considers set 3 "much worse" than set 2 which contains only one income, namely 60,000, whereas under the self-concern mode, the sets are not "that far apart" both according to the mean rank and the ranking position. Does this mean that in the ignorance scenario, respondents became more inequality-averse as impartial umpires? We admit that a positive answer to this question would be in some contrast to the statement above concerning the possibility of a zero income in income set 8.

Given the standards of behavior to which we referred earlier, the authors computed the theoretical ranking of the nine income sets for each standard of behavior. Then, for every respondent, Spearman's rank correlation between the empirical rank ordering of the nine income sets and any theoretical ranking was computed. This generated sixty-one rank correlation coefficients for each theoretical ranking. Without going too much into technical details, one can say that under the self-concern mode, Boulding's principle and expected utility were among the winning standards of behavior, and both were also in the leading group under the umpire mode. The leximin principle was the big loser under both modes.

In the risk scenario, income distributions 8, 9, and 12, characterized by high payoff, high risk, and high variance, lost significantly in mean rank in favor of distributions 1, 3, 4, and 5, which have low payoffs, low risk, and low variance, when switching from self-concern to external observer. Apparently, the respondents, on average, exhibited more inequality aversion under the umpire role than under self-concern. Income distributions 8 and 9 have lost in terms of mean rank when going from self-concern to umpire; however, they still rank highest in the students' orderings. Interestingly, for income distribution 12, which contains a zero income, a reaction was observed which is in full contrast to the one for income set 8 in the ignorance scenario.

A statistical analysis analogous to the one briefly outlined in connection with the ignorance scenario, shows that, again, expected utility maximization is among the winning standards of behavior. The Boulding principle fares "relatively well" under the umpire mode. Leximin behavior fares much better now than under the ignorance scenario, though it is "quite far away" from expected utility maximization.

There are several other investigations that look at how people judge income distributions when they are either behind a veil of ignorance but involved, or outside the veil, being uninvolved, or when they know their exact position in the distribution (see e.g. Herne and Suojanen 2004, and Amiel *et al.* 2006).

21.5 Other Aspects

Recent debates among philosophers (e.g. Arneson 1989, 1990; Cohen 1989, 1990) and economists, social choice theorists in particular (see Bossert 1995; Fleurbaey 1995, 1998; and Bossert and Fleurbaey 1996), introduced the notion of responsibility into welfare economic reasoning. It was proposed that one should distinguish between characteristics for which individuals are to be held responsible and those for which they should be compensated. Remember that we already discussed this issue in the context of production in Section 21.3 when we introduced Konow's concept of accountability. It was argued that the notion of control plays a major role in such a context. Given that a person is physically fit, her effort is largely under her direct influence. Natural talents are not under a person's control, so people with lower talents should be compensated for their disadvantage. But the dividing line between responsibility (i.e. noncompensation) and compensation is not so clear. Should a person, for example, be held responsible for her preferences?

Schokkaert and Devooght (2003) studied these issues from an empirical perspective by doing questionnaire-experimental investigations in three different countries, namely Belgium, Burkina Faso, and Indonesia. Among other issues, they examined the case of health-care financing where compensation takes place through a vector

of individual subsidies. Full compensation then means that two persons with the same responsibility characteristics should pay the same contribution. All in all, there was some support for the view that individuals are responsible for the preferences with which they identify, and also for the idea that individuals are responsible for "those things" which are under their control (being a heavy smoker, for example). While full egalitarianism was largely rejected, there was a clear support for intermediate compensation which is an inequality-reducing measure that does not go all the way towards egalitarianism. The intercultural differences were much less pronounced than one might have expected.

Gaertner and Schwettmann (2007) looked at the issue of whether responsibility aspects are being considered in cases of basic needs (Q 5.1). We saw in Section 21.2 when we discussed the situation of the handicapped person and the child(ren), that the support for the needy person diminished over the years, while efficiency aspects became stronger. Would responsibility as an additionally included argument further reduce the support for the worst-off person?

In one version, we gave the information that the retarded person was severely handicapped from birth. In a second version, presented to other students, of course, it was said that brain damage was due to an accident from participation in a dangerous sport (paragliding, let's say). Otherwise, the descriptions were exactly the same as before. We found (see Table 21.6) that fulfillment of the equity axiom is weaker for the responsibility case and furthermore, that the relative frequency of revising the initial decision is lower in the responsibility version—from 47.2 percent in the case of handicap from birth to 36.4 percent in the case of a dangerous sport. Surprisingly, unconditional support for the handicapped (i.e. sequence 0000) was higher in the latter case than it was in the former. We used chi-squared tests to check whether there were statistically significant differences between the answering patterns of the two versions. The null hypothesis that the distribution of responses in the "responsibility" variant is identical to the distribution in the "no responsibility" version could not be rejected. Only for the switching behavior (i.e. the sum of the frequencies of the sequences 0001, 0011, and 0111) could we reject the null hypothesis of no differences between the two samples at the 0.05 significance level.

As just mentioned, there was an increase in unconditional support for the handicapped in the case of responsibility. In order to learn more about what was really going on, we did a gender breakdown. We found that in comparison with female answers, the fact that the handicap resulted from a sports accident had a positive and significant (at the 5 percent level) effect on male answers regarding both the fulfillment of the equity axiom and the unconditional support of the worst-off. Having controlled for this effect, there was a highly significant (at the 1 percent level) negative influence on the fulfillment of the equity axiom coming from female respondents. In other words, basic needs are considerably less often supported by females if the suffering person is to be blamed for their own situation (for

Table 21.6. Q 5.1

Sequence	No responsibility 2002 + 2003 $n = 178$	Responsibility 2002 + 2003 $n = 187$
0 0 0 0	0.360	0.412
0 0 0 1	0.022	0.032
0 0 1 0	0.0	0.0
0 0 1 1	0.213	0.139
0 1 0 0	0.0	0.0
0 1 0 1	0.011	0.0
0 1 1 0	0.006	0.0
0 1 1 1	0.236	0.193
1 0 0 0	0.0	0.0
1 0 0 1	0.0	0.0
1 0 1 0	0.0	0.0
1 0 1 1	0.006	0.011
1 1 0 0	0.011	0.005
1 1 0 1	0.0	0.0
1 1 1 0	0.0	0.0
1 1 1 1	0.135	0.209
% of switch	47.2	36.4
% fulfillment of equity axiom	84.8	77.5

further details, see Gaertner and Schwettmann 2007). So gender obviously makes a difference. While, in general, women show more concern for the worse-off in society, in this particular case, they seem to have clear-cut reservations.

This brings us to the gender issue. In the psychological and sociological literature, there are many studies on gender disparities. A general view is that women are more socially oriented, care more about needs, and are better able to take on the perspective of others (Gilligan 1982; Davis 1983; Eisenberg and Lennon 1983; Eagly 1995). Dickinson and Tiefenthaler (2002) and Michelbach *et al.* (2003) found that men are often concerned about efficiency, while women are more likely to prefer equality. Croson and Gneezy (2004) argue that men tend to decide with less regard to context and more in compliance with abstract rules, so to speak, trying to detect a common structure in different situations, while women consider each case separately, being sensitive to small variations. Also, according to Croson and Gneezy, women tend to be significantly more risk-averse than men, which could explain, at least to some extent, why male students in Osnabrück seemed to honor the risky sport of the paraglider, while women were inclined to punish this activity.

But, of course, gender does not always matter. In some of the investigations we discussed in previous sections, the authors asserted that they did not see any gender

effect. In one of our own studies that we did not cover in this survey, there clearly were gender effects, but not in a systematic way. This shows that both analysis and interpretation of gender differences have to be done with great care.

Are goods that are priced for the pleasure one derives from their consumption, their hedonic value so to speak, viewed differently from goods valued for their importance to one's health? Yaari and Bar-Hillel (1984) rewrote the situation that we discussed in the first half of Section 21.2 in such a way that the underlying concern was not needs but tastes (the issue was no longer metabolism but willingness to pay). They found that the distribution of answers to the new problem was quite different from the one in terms of needs (the authors mention that under a chi-squared test, the difference between the distributions was significant at the 1 percent level). The distribution (*J*: 8, 0; *S*: 4, 12) still received a relatively high percentage of support (much less, however, than under Q 2.1), but it was surpassed by (*J*: 12, 0; *S*: 0, 12), which is supported by utilitarianism. When the authors considerably reduced the willingness to pay on the part of Smith, the latter distribution gained further support, so that, apparently, the respondents wanted to penalize Smith for the drop in his willingness to pay.

Another twist in the Yaari and Bar-Hillel investigations was to introduce beliefs. Returning to their very first situation, at the beginning of Section 21.2, it was now said that Jones *believed* that each grapefruit contained 100 milligrammes of vitamin F and that avocados did not contain vitamin F at all, and that Smith *believed* that a grapefruit and an avocado each contained 50 milligrammes of vitamin F. The authors found that there was fairly strong support for an equal split in this case, and little support for the so-called utilitarian solution. Fifty-one percent of the respondents wanted to honor the beliefs of the two persons, and clearly wished that an equal amount of vitamin F be obtained by the two, each according to his own beliefs.

Tversky and Kahneman (1981) were among the first to point out that different framings of the same problem matter. If a certain phenomenon has positive and negative features, then framing in positive terms will elicit different reactions or verdicts from framing in negative terms, though the underlying situation is exactly the same under both variants.

Do similar phenomena exist in the context of justice evaluations? This was the object of an investigation by Gamliel and Peer (2006), and the two authors provide evidence that the answer apparently is "yes". The hypothesis is that "positive framing of a resource allocation should lead to a more favorable association, which will lead to a more favorable judgment of the allocation situation and the principles used to accomplish the allocation" (2006, p. 312). One of the authors' experiments examined selection procedures, viz. accepting/rejecting students who applied to higher education institutions and accepting/rejecting potential personnel who applied to prospective employers. The distributive principle which was tested was the rule of merit (a combination of an individual's ability and effort). The situations were described in a positive way (to accept half of the applicants) or—relative

to the very same situation—in a negative way (to reject half of the applicants). The experimental results are that allocation by merit was preferred more under positive framing (acceptance) than under negative framing (rejection). A different framing apparently generates a shift in people's evaluations due only to different descriptions of exactly the same situation. In other words, the positive framing of an allocation problem seems to lead to a different encoding of the information than the negative framing of an identical situation.

I wish to end this essay by mentioning at least two aspects that I find particularly important. One aspect is the time dimension, the other the inter-country dimension. In order to arrive at statements or verdicts which express more than a whim or a quirk, one has to reiterate experiments. One can then see whether the results that one has obtained are stable over time or, if they are not, whether some time trend is visible. A one-time experiment done at a particular location cannot provide much insight. Also, by running experiments at different places, one can (try to) find an answer to the question of whether the cultural and/or historical background matters. All this is highly relevant for a refinement of the underlying theory. But it is also important insofar as empirical social choice may eventually offer advice and recommendations to decision-makers, politicians in particular.

References

Amiel, Y., and Cowell, F. A. (1994). Income Inequality and Social Welfare. In J. Creedy (ed.), *Taxation, Poverty and Income Distribution*, 193–219. Brookfield: E. Elgar.

——— and Gaertner, W. (2006). To Be or Not to Be Involved: A Questionnaire-Experimental View on Harsanyi's Utilitarian Ethics. STICERD Discussion Paper, London School of Economics. Forthcoming in *Social Choice and Welfare*.

Aristotle (1976). *The Nicomachean Ethics*, trans. J. A. K. Thomson, 2nd edn. Harmondsworth: Penguin.

Arneson, R. J. (1989). Equality and Equal Opportunity for Welfare. *Philosophical Studies*, 56, 77–93.

——— (1990). Liberalism, Distributive Subjectivism, and Equal Opportunity for Welfare. *Philosophy and Public Affairs*, 19, 158–94.

Bar-Hillel, M., and Yaari, M. (1993). Judgments of Distributive Justice. In B. A. Mellers and J. Baron (eds.), *Psychological Perspectives on Justice*, 55–84. Cambridge: Cambridge University Press.

Beckman, St. R., Formby, P. P., Smith, W. J., and Zheng, B. (2002). Envy, Malice and Pareto Efficiency: An Experimental Examination. *Social Choice and Welfare*, 19, 349–67.

Bossert, W. (1995). Redistribution Mechanism Based on Individual Characteristics. *Mathematical Social Sciences*, 29, 1–17.

——— and Fleurbaey, M. (1996). Redistribution and Compensation. *Social Choice and Welfare*, 13, 343–55.

Cohen, G. (1989). On the Currency of Egalitarian Justice. *Ethics*, 99, 906–44.
—— (1990). Equality of What? On Welfare, Goods, and Capabilities. *Recherches Economiques de Louvain*, 56, 357–82.
Croson, R., and Gneezy, U. (2004). Gender Differences in Preferences. Mimeo, Graduate School of Business, University of Chicago.
Davis, M. (1983). The Effect of Dispositional Empathy on Emotional Reactions and Helping: A Multi-Dimensional Approach. *Journal of Personality*, 51, 167–84.
Deschamps, R., and Gevers, L. (1978). Leximin and Utilitarian Rules: A Joint Characterization. *Journal of Economic Theory*, 17, 143–63.
Dickinson, D. L., and Tiefenthaler, J. (2002). What is Fair? Experimental Evidence. *Southern Economic Journal*, 69, 414–28.
Dworkin, R. (1981). What Is Equality? Part 1: Equality of Welfare; Part 2: Equality of Resources. *Philosophy and Public Affairs*, 10, 185–246, 283–345.
Eagly, A. H. (1995). The Science and Politics of Comparing Women and Men. *American Psychologist*, 50, 145–58.
Eisenberg, N., and Lennon, R. (1983). Sex Differences in Empathy and Related Capacities. *Psychological Bulletin*, 94, 100–31.
Fleurbaey, M. (1995). Three Solutions for the Compensation Problem. *Journal of Economic Theory*, 65, 505–21.
—— (1998). Equality among Responsible Individuals. In J.-Fr. Laslier *et al.* (eds.), *Freedom in Economics*, 206–34. London: Routledge.
Gaertner, W. (1992). Distributive Judgements. In W. Gaertner and M. Klemisch-Ahlert (eds.), *Social Choice and Bargaining Perspectives on Distributive Justice*, ch. 2. Heidelberg, Berlin, and New York: Springer Verlag.
—— and Jungeilges, J. (2002). Evaluations via Extended Orderings: Empirical Findings from West and East. *Social Choice and Welfare*, 19, 29–55.
—— and Schwettmann, L. (2007). Equity, Responsibility and the Cultural Dimension. *Economica*, 74, 627–49.
—— Jungeilges, J., and Neck, J. (2001). Cross-Cultural Equity Evaluations: A Questionnaire-Experimental Approach. *European Economic Review*, 45, 953–63.
Gamliel, E., and Peer, E. (2006). Positive versus Negative Framing Affects Justice Judgments. *Social Justice Research*, 19, 307–22.
Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.
Hammond, P. J. (1976). Equity, Arrow's Conditions, and Rawls' Difference Principle. *Econometrica*, 44, 793–804.
Harsanyi, J. C. (1953). Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking. *Journal of Political Economy*, 61, 434–5.
—— (1955). Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility. *Journal of Political Economy*, 63, 309–21.
Herne, K., and Suojanen, M. (2004). The Role of Information in Choices over Income Distributions. *Journal of Conflict Resolution*, 48, 173–93.
Kalai, E., and Smorodinsky, M. (1975). Other Solutions to Nash's Bargaining Problem. *Econometrica*, 43, 155–62.
Klemisch-Ahlert, M. (1992). Distributive Results in Bargaining Experiments. In W. Gaertner and M. Klemisch-Ahlert (eds.), *Social Choice and Bargaining Perspectives on Distributive Justice*, ch. 5. Heidelberg, Berlin, New York: Springer Verlag.

Konow, J. (2001). Fair and Square: The Four Sides of Distributive Justice. *Journal of Economic Behavior and Organization*, 46, 137–64.

Michelbach, Ph. A., Scott, J. T., Matland, R. E., and Bornstein, B. H. (2003). Doing Rawls Justice: An Experimental Study of Income Distribution Norms. *American Journal of Political Science*, 47, 523–39.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Schokkaert, E. (1999). Tout-le-monde est "Post-Welfariste": Opinions sur la Justice Redistributive. *Revue Economique*, 50, 811–31.

—— and Devooght, K. (2003). Responsibility-Sensitive Fair Compensation in Different Cultures. *Social Choice and Welfare*, 21, 207–42.

Sen, A. (1973). *On Economic Inequality*. Oxford: Clarendon Press.

Traub, S., Seidl, C., Schmidt, U., and Levati, M. V. (2005). Friedman, Harsanyi, Rawls, Boulding—or Somebody Else? An Experimental Investigation of Distributive Justice. *Social Choice and Welfare*, 24, 283–309.

Tversky, A., and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211, 453–8.

Yaari, M., and Bar-Hillel, M. (1984). On Dividing Justly. *Social Choice and Welfare*, 1, 1–24.

CHAPTER 22

SOCIAL CHOICE IN HEALTH AND HEALTH CARE

AKI TSUCHIYA
JOHN MIYAMOTO

22.1 Introduction

While a reasonable level of health is essential for individuals to flourish, necessary information for rational decision-making regarding health can be seriously lacking to individual consumers. The individual consumer, or patient, typically does not know well enough about their state of health and prognosis, available interventions, or expected outcomes of these. At the extreme, the consumer cannot tell whether or not consuming health care would promote their well-being. On the other hand, the providers, i.e. the medical profession, may know more; but even then, most complex biological processes and medical interventions involve some element of uncertainty, so they cannot always predict the outcome with certainty. Furthermore, a typical consultation process means that the supplier of the service (i.e. the doctor) acts at the same time as the agent of the consumer (i.e. patient),

An earlier version of this chapter was presented at the Health Economists' Study Group meeting at the University of York, UK, July 2006; we would like to thank all those who participated in the debate. We would also like to thank Jan Bleichrodt for comments and Paul Anand for encouragement. The usual disclaimer applies.

and this imperfect agency relationship compromises the standard market model (Arrow 1963 is a classic addressing the various issues of the health-care market).

Where this understanding that a market-based model of health-care provision is unlikely to achieve an efficient outcome is coupled with the recognition that there are concerns about equity in health policy, these often lead to a justification for some form of government intervention in the allocation of health-care resources. Roughly speaking, there are at least three ways of operationalizing concerns for equity in this area. The first is equity defined as equal access to health-care services for equal need, regardless of ability to pay. This was a central objective of the National Health Service when it was set up in the UK in 1948, prompted by the *Beveridge Report* (Beveridge 1942; see also McGuire *et al.* 1988; Mooney 1992). The second is equity defined as equal health across individuals, or population subgroups. This is in line with, for example, the World Health Organization (1981) declaration for "health for all by the year 2000", or the 1999 UK Government White Paper (Department of Health 1999) in which reducing "avoidable inequalities in health" was a key theme (see also Culyer and Wagstaff 1993; Williams and Cookson 2000). In addition to these, the concept of equity has been explored in terms of procedural justice in health-care decision-making, as opposed to distributional justice; more so in the academic literature than in policy documents (see e.g. Daniels and Sabin 1997; Tsuchiya *et al.* 2005; Wailoo and Anand 2005).

Whichever way equity is conceptualized, instead of individual consumers deciding (with the help of their doctor, the supplier) how much of what health care to consume, given their state of health, their preferences, their income, and going prices, now the public health-care system will prescribe what can be provided to which patients under what circumstances, often at no, or heavily subsidized, charge to the patient at the point of consumption. Individuals maintain the choice not to utilize the services offered or to refuse specific treatment, but the scope for their exercising their choice to consume will be highly restricted. As a result, a large part of the discipline of health economics today, especially in countries where there is a large, publicly funded health-care system, is devoted to the development of methodologies of economic evaluation of health-care interventions, and the actual execution of these evaluations (see Maynard and Kanavos 2000 for a brief overview). In other words, within the discipline of health economics, there has been a major shift from a positive study of how individuals make consumption decisions for themselves, to a normative study of how society should make resource allocation decisions within the remit of a publicly funded health-care system. While there remains a significant portion of health economics that is about positive explorations of individual behavior and choice, this chapter will concentrate on the approaches in health economics regarding normative social choice.

This move, from the individual level to the social level, may also involve a widening of the context, from clinical medicine to public health, social care, and beyond. Clinical medicine can have major impacts on individual health once illness

or injury has happened. On the other hand, factors such as clean air and water, sufficient nutrition, warm and dry housing, basic education, friendly and trustworthy neighbors—in other words, factors beyond health care—can have a much larger long-term impact on the health of a population (see e.g. Fuchs 1974; McKeown 1976). However, while there are emerging moves to expand health economics to cover these topics, the current core business of heath economics is in essence the economics of health-care services, in particular, economic evaluation of medical interventions.

Economic evaluation of medical interventions may take many different forms (see Drummond *et al.* 2005). The first choice for a (non-health) economist is likely to be cost–benefit analysis, where both the costs and the benefits of the intervention are evaluated in monetary terms, and the net present value of the intervention is calculated. However, a large part of economic evaluations carried out in the health-care area does not use cost–benefit analyses. Part of the reason for this is undoubtedly due to the difficulties associated with assigning monetary values to nonmarket goods in general; both revealed preference studies and expressed preference studies are known to have problems. In addition to these, health economists are faced with strong resistance from the medical community (without whose collaboration it is impossible to carry out an economic evaluation of a medical intervention) regarding monetary valuations of people's health and life. Consequently, the most commonly used form of economic evaluation relies on the concept of the "quality-adjusted life-year (QALY) gained" to represent the benefit of health care, and compares this with the costs of producing it. The QALY is a composite measure of the duration of survival and the health-related quality of life (HRQOL) associated with each period of survival (Culyer *et al.* 1971). This form of economic evaluation is referred to in the literature as "cost–utility analysis", "cost-effectiveness analysis using QALYs as the outcome measure", and "cost per QALY analysis". The last term will be used in this chapter for reasons that will become clear below.

Against this background, this chapter will break the problem into two parts. The first is about what it is that should be maximized, or distributed. People have referred to this as the maximand or distribuendum. In this chapter, we will use the term "desideratum" to mean what it is that is treated as the good in question so that efficiency requires producing more of it and equity requires distributing it fairly. By narrowing down the key topics in this way, the main concept of equity addressed in this chapter becomes the equality of health across individuals and/or population subgroups. At a practical level, for those engaged in cost per QALY analyses, the desideratum is the net number of QALYs gained, achieved across a given population. So the issue addressed here is what QALYs represent, and why they are regarded as the desideratum. The second part is about how to aggregate. The assumption is that each individual has some level of the desideratum, and that the desirability of health policies or the running of health-care systems is a function

of the desideratum to individuals: the question is how the individual shares are added up across the population. The chapter does not cover topics on HRQOL valuation in much detail, and does not address any of the issues of actual economic evaluation, which include costs, discounting, and the treatment of uncertainty. The interested reader is referred to Drummond *et al.* (2005) and/or Brazier *et al.* (2007).

22.2 The Policy Desideratum

This section addresses two different approaches within health economics to the justification of the QALY as the desideratum in cost per QALY analyses. The first approach, "welfarism", holds that the QALY is the desideratum because it represents individual utility, whereas the second approach, "non-welfarism", holds that the QALY is the desideratum because that is what the relevant stakeholders want in the context of a publicly funded health-care system.

22.2.1 Welfarism and the QALY as the Desideratum

Utilitarian economic theory is welfarist. Welfarism holds that the only information that is relevant to assessing social welfare is the level of personal utility achieved by each composite member of society (see e.g. Sen 1979). In this chapter, the term "personal utility" is used to indicate the level of utility that is perceived by the individual as the level of their own well-being. This is not necessarily the same as "individual utility", which may be affected by those things that the individual appreciates and abhors, but where their own state of well-being is not affected in any direct sense. In many situations, personal utility and individual utility coincide. However, for example, preferences over possible outcomes for other people or preferences over distributions where one's own absolute and relative well-being are not affected can be part of individual utility but not personal utility.

Now, the welfarist desideratum in the context of health and health care is personal utility of the individuals affected by the health-care intervention in question. In other words, the QALY is the desideratum because it represents the level of utility that individuals will experience in association with their own level of health. There is a growing literature dedicated to the identification of the set of conditions that need to be met for this interpretation to hold.

The most basic, or "linear" QALY model asserts that

$$U(Y, Q) = Y \cdot H^I(Q),$$

where U is the personal utility function over life-years (Y) and HRQOL (Q) that the individual in question enjoys, and H^I is a utility function defined on each individual's own health states. This linear QALY model is the standard QALY model—when people refer to "the" QALY model, they almost always mean the model above. From the expected utility perspective, the claim that personal utility can be represented by the product of life-years and the utility over states of health depends on the validity of a series of assumptions. Pliskin, Shepard, and Weinstein (1980) proved that the model holds if duration and health quality are mutually utility-independent, constant proportional time tradeoff holds, and there is risk neutrality on life-years. Subsequent research showed that these assumptions could be substantially simplified (Bleichrodt, Wakker, and Johannesson 1997; Miyamoto 1999; see Bleichrodt and Pinto 2006 for a review).

Empirical research shows that actual people do not value the survival duration linearly (Miyamoto and Eraker 1985, 1989), so one may question the empirical validity of the standard linear QALY model (see also Tsuchiya and Dolan 2005 for a review). However, it may be argued that it is a convenient and reasonable approximation (see Miyamoto and Eraker 1985; Doctor *et al.* 2004).

Of interest is how to assess the quality-adjustment weights to represent HRQOL. The welfarist answer is to ask individuals as the recipients of health. This can be justified in two ways: because the individual himself is the best judge of his own welfare, or in other words, of how much utility the expected health improvement will yield for him, and because it is his own health. Since the issue is knowledge and self-reflection, in certain health conditions that affect mental capacities, it might be better to rely on carers' or experts' views than on patients' views. So let us call the *compos mentis* patient the "capable patient". The capable patient is assumed to satisfy certain criteria. First, he knows what it feels like to have different health problems. Second, he is assumed to be rational in the practical sense, which we define here to mean having preferences that satisfy requirements of internal consistency required by the method of assessment (viz. if the results are to be interpreted using expected utility theory, then the preferences must be rational as defined by this theory). And third, he is selfish as in the economic sense; i.e. he takes into account only his own personal utility derived from the intervention.

Capable patients who suffer a specific health problem (call it "health state X") are interviewed with respect to their preferences for health state X. Other capable patients who are suffering from health state Y are interviewed with respect to Y. Utilities for states X and Y are placed on the same scale by having reference points, full health and dead, included among the health states that both groups evaluate. One such method is the standard gamble, which operationalizes expected utility as proposed by von Neuman and Morgenstern (1944). Standard gambles in the health application ask the interviewee to compare two scenarios, one in which he survives in his current state X (or Y) for certain for a fixed period, say, 10 years, and then dies, and another scenario in which he faces a gamble between survival in full health

for the same fixed period (followed by death) with probability p, or immediate death. The utility associated with state X can be inferred from the level of p that makes the respondent indifferent between the two scenarios. Another method of HRQOL valuation is the time tradeoff method (Torrance *et al.* 1972), where the respondent is asked to choose between two scenarios, one in which he will live in his current state X for a fixed period, say 10 years, and then die, and another in which he will live in full health for a shorter duration t years and then die. The value of living in state X can be inferred from the level of t that makes the respondent indifferent between the two scenarios.

If we set

$$H^I(\text{full health}) = 1$$

and

$$H^I(\text{dead}) = 0,$$

so that the utility of dead and full health are assumed to be identical across individuals, then the ratios

$$[H^I(\text{full health}) - H^I(X)]/[H^I(\text{full health}) - H^I(\text{dead})]$$

and

$$[H^I(\text{full health}) - H^I(Y)]/[H^I(\text{full health}) - H^I(\text{dead})]$$

become straightforward to interpret (Fryback and Lawrence 1997; see also Bleichrodt 1997).

The QALY is a cardinal and interpersonally comparable measure of utility. Since it is the product of the number of life-years and HRQOL weights derived in the manner above, the resulting number of QALYs is measurable on a cardinal scale. Interpersonal comparability of QALYs relies, first, on imposing common fixed utility values for full health and dead across individuals, and secondly on assuming that a year of life in full health yields equal personal utility to all. One may find these problematic: how can we ascertain whether or not different people appreciate the difference between full health and dead for themselves equally, and whether or not different people achieve the same level of utility from a year of life in full health, without assuming interpersonal comparison of utility in the first place?

An alternative to surveying capable patients is to ask non-patients to take part in a similar exercise, but instead of asking them to evaluate the health state they are currently in, asking them to evaluate hypothetical health states against full health and dead, imagining they were to find themselves in these states. The obvious advantage of this design is that valuations for more than one state such as X and Y can be obtained from the same respondent. However, since non-patients do not know at first hand what it is like to be in hypothetical states, it is crucial that they be given enough information to carry out the evaluation task. Let us call

such respondents "informed non-patients". The informed non-patient is assumed to know about ill health states, but he himself is not ill; he is assumed to be rational and selfish.

It may be feasible to try to contrast the capable patient values and the informed non-patient values in a manner parallel to the contrast between experienced utility and decision utility (Kahneman *et al.* 1997). While all four concepts are about how states of ill health are perceived, capable patient values and experienced utility are concerned with how things actually feel in real time when the individuals are living with the condition in question, whereas informed non-patient values and decision utility capture how people think they would feel were they to experience these states. In other words, the informed non-patient corresponds to the consumer before consumption, contemplating consumption of a good, and the competent patient corresponds to the consumer after consumption, having purchased the good. While using the values obtained from informed non-patients is closer to the framework of consumer theory based on decision utility, in a move parallel to the emergence of interest in experienced utility in the recent economics literature, there is an emerging interest in values obtained from competent patients in the health economics literature (Brazier *et al.* 2005).

22.2.2 Non-Welfarism and the QALY as the Desideratum

On the other hand, there is an alternative approach within health economics which holds the QALY as the social desideratum not because it is valued by individuals as patients or consumers (although it may well be), but because it is valued by the public at large or the relevant decision-makers (e.g. policymakers in the National Health Service or the Department of Health). This approach has been referred to as "the decision-makers' approach" (Sugden and Williams 1978), or "extra-welfarism" (Culyer 1989). While some authors distinguish between the two, here they are not distinguished from each other and are collectively referred to as "non-welfarism" (Tsuchiya and Williams 2001). The common tenet of these is that social welfare in the context of public policy decision-making is not a function of the utility enjoyed by constituent individuals of society as judged by themselves, but a function of social desiderata, dictated by the relevant policy context. In the context of public health policy, the desideratum is population health, as operationalized by the QALY. This is because the objective of the heath-care system is to make people healthier, not to make people happier (see Feldstein 1963). This is in contrast to welfarism (Section 22.2.1 above), where health was the desideratum precisely because and to the extent that individuals as patients or consumers appreciate it (see Pauly 1994). On the other hand, the non-welfarist interpretation of the QALY has been compared to Sen's concept of capabilities (Cookson 2005). Since the QALY does not represent individual utility, the term "cost–utility analysis" becomes unsuitable.

The non-welfarist approach also follows the linear QALY model:

$$W(Y, Q) = Y \cdot H^S(Q),$$

where W is the social welfare function over life-years (Y) and HRQOL (Q), and H^S is a function reflecting the societal value of different health states. This linear QALY model corresponds to the standard QALY model under welfarism above. Therefore, the same set of conditions as applies to H^I also applies to H^S; i.e. mutually independent social value of duration and health quality, constant proportional time tradeoff, and risk neutrality with respect to duration. Similarly, the non-welfarist QALY is cardinal and interpersonally comparable. Interpersonal comparability is less problematic under non-welfarism than under welfarism, since it is down to what the relevant decision-maker wants. It is a matter of choice for the decision-maker to set equal values across individuals for full health and for dead, and to set an equal value for a year of survival in full health.

Regarding the method of assessing HRQOL, non-welfarism argues that, since in a publicly funded health-care system it is ultimately about how to use tax moneys, it should be based on what members of the tax-paying citizenry think about different health outcomes across society; in other words, the judgment should come from a citizen, or societal perspective.[1] Let us call the individual in this context the "informed citizen". The informed citizen is assumed to know what it feels like to have different health problems, to be rational, and to be selfless in the sense that she will not make judgments in order to forward her own case, or to advance the case of one particular health problem over another (see e.g. Gold *et al.* 1996). In other words, the informed citizen corresponds to a variant of the planner or the ethical impartial observer.

Note that informed citizens are in effect the same people as competent patients or informed non-patients, but they assess health states from different perspectives. The capable patient can also be the informed citizen by adopting the appropriate perspective. Consequently, the same valuation methods, such as standard gamble and time tradeoff, can be used to elicit values from the societal perspective. For instance, participants can be asked to imagine a group of a certain number of unnamed individuals and asked to make a choice for them between survival in state X for certain and a gamble between survival in full health and death; or between survival in state X for a fixed number of years and survival in full health for a shorter duration.

[1] Note that some authors use the term "societal perspective" to mean the remit of an economic analysis. For example, an analysis carried out from the perspective of a specific health-care institution (say the national health service or a hospital) will be different from one carried out from the societal perspective, covering all costs and all benefits regardless of to whom they accrue. Here, the term "societal perspective" is used to mean the citizen perspective, as opposed to the individual, consumer perspective.

Some authors have associated cost–benefit analysis with welfarism, and cost per QALY analysis with non- (or extra-)welfarism (Culyer 1989; Hurley 2000). However, and first, as we saw in Section 22.2.1 above, there is a thriving literature on the welfarist conceptualization of the QALY, which implies that cost per QALY analysis is compatible with welfarism. Second, there are not many actual HRQOL valuation studies that use the citizen perspective. For example, the time tradeoff exercise used in the well-established health state classification instrument EQ-5D asked whether the *respondents themselves* preferred to live life A (a longer life in less than full health) or life B (a shorter life in full health) (Dolan 1997), which assumes that the respondents are informed non-patients, and therefore welfarist. The only set of HRQOL weights that currently exists that is non-welfarist is likely to be the Disability Weights, developed for use in the calculation of the global burden of disease (World Bank 1993), based on the person tradeoff method (Nord 1995). The person tradeoff method asks respondents to choose between two groups of people, where she does not belong to either group. For example, one group may consist of 1000 people who, if chosen, will live for a fixed period of time in state X and then die, whereas the other group consists of n people who, if chosen, will live for the same duration in full health and then die. Those in the unchosen group will all die within a few days. The value associated with X relative to full health and dead can be inferred from the level of the number of people, n, that makes the respondent indifferent between choosing either group.

The welfarist's concern over non-welfarism is the legitimacy of the informed citizen as a source of value. If the personal preferences of the capable patient were to clash with the judgments of the informed citizen, why should the latter view be given any more weight than the former view? For example, if people with chronic health problems learn to adapt to their state, then their valuation of their own health state as capable patients may be much higher than how an informed citizen with no direct long-term experience of the health problem may value the same state. Who is to say the informed citizen's value is "correct" and the competent patient's value is "wrong"? The non-welfarist's reply to this is likely to be along the following lines. First, it is not an issue of which values are "correct" and which ones "wrong". The two parties have different values and preferences. The issue is which value is the more appropriate to use. Second, if the debate is set against a freely competitive health-care market, where the competent patient is paying out of their own pocket for their own health care, then their own marginal utility should be a key variable determining consumption. However, third, if the debate is set in the context of a publicly funded health-care system, this brings in two additional considerations. The main objective of a publicly funded health-care system is not merely to pursue the most efficient ways in which to facilitate individuals maximizing their own personal utility. As was noted above, publicly funded health-care systems are typically concerned with improving population health as opposed to personal utility, and with improving equity as well as efficiency. Governments are

concerned not only about the inefficiency of a possible health-care market. This requires a perspective that is detached from the individual as the selfish utility maximizer.

22.3 The Aggregation Rule

The key issues addressed in this section are: what the aggregation rule should be, and, if nonuniform weights are involved, how are they to be justified, and how are they to be determined. Aggregation rules do not preclude any particular desideratum, but may have higher affinity with either welfarism or non-welfarism.

22.3.1 The Simplest Aggregation Rule: Total Sum with Uniform Weights

The simplest aggregation rule[2] is to add up the changes in the desideratum across individuals without any weights (or, equivalently, with uniform weights) so that the outcome with the largest total is recognized as the best outcome. This aggregation rule is applicable to both welfarism and non-welfarism, and in effect this is how the benefits are calculated in cost per QALY analyses. Although the use of uniform weights is the simplest approach to aggregation, and it may seem to be the obvious default choice, this does not mean that it needs no justification. Under non-welfarism, equal weights can be justified with reference to what the relevant policymakers think or what members of the public as informed citizens support from the societal perspective. A less simple issue is how to justify the use of uniform weights under welfarism.

Equal weighting can be inferred from a "permutation axiom" that was proposed by Camacho (1979, 1980) in his repetitions approach to the foundations of cardinal utility. In the repetitions approach, the individual is asked to state preferences for arbitrarily many repetitions of the same riskless choice, e.g. the choice of wine at a given restaurant, at the same table, with the same menu, with the same company, etc. The permutation axiom states that "the satisfaction derived from a finite sequence of choices depends *only* on the choices entering the sequence and not on the order in which they appear" (Camacho 1980, p. 364; emphasis original)

[2] By far the least restrictive social decision rule says that nobody should lose or be made worse off (viz. the Pareto criterion), and it is applicable under either welfarism or non-welfarism. However, it is not the most useful rule, since it is highly incomplete (i.e. there will be multiple outcomes that cannot be rank-ordered against each other). It is rare that actual policy decisions can be justified with reference to the Pareto criterion.

As Wakker pointed out (personal communication to JM), the choices could be interpreted as the outcomes for different individuals in a society rather than as outcomes of a series of repetitions of the same choice. The social welfare version of the permutation axiom would then state that the level of social welfare derived from a finite set of outcomes across individuals who are equal in all relevant respects depends only on the outcomes in the set and not on the particular pattern of distribution of these outcomes across the individuals. The sameness of the choice in the original axiom will translate into the individuals being equal in all relevant aspects in this social version.

So, for example, the aggregation process should be indifferent between an outcome where you are very sick and I am healthy, and an outcome where I am very sick and you are healthy. Thus, equal weighting in aggregation can be linked to more basic preference assumptions. The next challenge for welfarism then becomes who decides, and how, whether or not different people are equal in all relevant respects.

22.3.2 The Introduction of Inequality Aversion, or Distributional Weights

If there is aversion to unequal distributions of the desideratum across people who are equal in all relevant respects, then the aggregation rule can incorporate inequality aversion so that the marginal societal value of increased desideratum is greatest when it goes to the worst-off individuals. This aggregation rule is blind to the characteristics of the individuals, and simply has the effect of equalizing the distribution of outcomes. Under non-welfarism, the degree of inequality aversion can be derived from the informed citizen or policymakers, by using valuation methods that trade off benefits. For example, they will present two or more groups of patients and contrast outcomes that have larger total health (in terms of unweighted QALYs) but with less equal distribution of this, and those that have smaller total health but with more equal distribution of this.

To illustrate, suppose two groups of equal size: A and B. In outcome 1, those in A can expect to live 70 QALYs and those in B can expect to live 80 QALYs. In outcome 2, those in A can expect to live 73 QALYs and those in B can expect to live 74 QALYs. If efficiency is measured by the sum of the levels of the desideratum across the two groups, and if equality is measured by the difference in the levels of the desideratum across the two groups, then outcome 1 is relatively more efficient and relatively less equal, whereas outcome 2 is less efficient and more equal. The aim would be to present different combinations of levels of the desideratum, in order to ascertain the amount of efficiency that people are willing to forgo to obtain an equal distribution of this. More specifically, suppose the median individual (or mean preference) is indifferent between the two outcomes above. Then by specifying an objective function (e.g. one with a constant elasticity of substitution between

marginal health improvements between the two groups), the implied degree of inequality aversion can be derived corresponding to this particular individual, or preference (Williams *et al.* 2005). This will allow the identification of the implied equally distributed health equivalent, and the calculation of relative weights that should be applied to marginal health changes to different people based on the marginal rate of substitution between the health of the two groups. An alternative approach is to base the social objective function on the rank-dependent utility model (Bleichrodt *et al.* 2004; Bleichrodt *et al.* 2005).

The measurement of inequality aversion above is similar to the way that risk aversion is measured, with probabilities associated with different outcomes replaced with the proportion of people associated with different outcomes. In the context of personal utility, a risk-averse individual will feel safer in a world with less inequality than more, because this suggests less variability in possible outcomes for herself. For this reason, various mechanisms have been proposed under which preferences of selfish individuals faced with uncertainty over their future prospects are interpreted as representing aversion to inequality across different individuals within the society. However, when individuals have personal utility as consumers over possible outcomes for themselves, this represents the level of risk aversion of the personal utility function, which is distinct from societal preferences that individuals as citizens may have over possible distributions across different individuals in society. And it is this latter preference that represents the level of inequality aversion of the social objective function.

To illustrate the distinction, think of the case where there is a disease with an incidence rate, p. Those individuals who are hit by the disease will be in poor health, and those who are not affected will be in good health. In this case, individual risk can be translated into a distribution at the population level. However, think of another case where there are two states of the world; endemic and no endemic. If endemic happens with probability p, then everybody will be in poor health, and if no endemic happens, then everybody will be in good health. In this case, the risk to the individual may not translate into distribution at the population level. Let us assume, for the sake of the argument, that the impact of poor health is short-lasting, that people achieve full recovery within a couple of days, and that overall it has no long-term impact on the economy. From the point of view of an entirely selfish consumer, the individual incidence case and the endemic case are the same; they will be in poor health with probability p, and otherwise in good health. At the social level, whereas a risk-averse and distribution-neutral social objective function will be indifferent between the two cases, a risk-neutral and inequality-averse social objective function will rank the second case higher.

Since the capable patient and the informed non-patient are selfish, although they may well be risk-averse, they are less suited to be a source for determining the level of inequality aversion to use in aggregation. As such, this aggregation rule, which incorporates weights to reflect aversion to inequality, has higher affinity with

non-welfarism (which is based on the societal perspective and the social welfare function) than welfarism (which is based on the individual consumer perspective and the personal utility function), not because welfarism is incompatible with unequal weights, but because welfarism cannot determine the level of inequality aversion beyond individual risk aversion.

As an alternative approach, by rephrasing risk aversion as diminishing marginal utility of income, or of QALYs, at the individual level, and by assuming an inequality-neutral aggregation rule, social welfare will be improved more by allocating additional income to the poor, or additional QALYs to the poorly, so that the effects of inequality aversion are achieved. Nevertheless, while the effects are similar, the underlying reasons are completely different. With this approach, equality is achieved as a side product of efficiency. There have been further attempts to incorporate inequality aversion into welfarism: for instance, by defining distributional weights so that the marginal social value of personal utility is decreasing in own income, or health. The difficulty is, as long as one stays within a welfarist framework, it is not obvious who determines this weight, and how.

22.3.3 The Introduction of Equity Weights, and Efficiency Weights

Within non-welfarism, if there is some notion of equity or justice that the informed citizen, or policymaker, supports, then the aggregation rule can include "equity weights". For example, *ceteris paribus*, if a severe health problem is regarded as deserving of higher priority than mild health problems, then this can be incorporated. Other candidate considerations may include expected health outcome with treatment, age, cause of the ill health, etc. (See Dolan *et al.* 2005 for a review of empirical studies, and Dolan and Tsuchiya 2006 for an overview.)

Elicitation of equity preferences is an area where it is important to probe the reasons why people support differential treatment of fellow citizens depending on their characteristics. Contrast this with typical welfarist utility assessments by standard gamble or time tradeoff methods—rarely are respondents asked *why* they give the responses that they give. When eliciting equity weights, researchers need to distinguish between justifiable societal preferences and unacceptable views based on prejudices (e.g. differential treatment by "irrelevant" characteristics such as race, sexual orientation, or religion). This has led to the use of qualitative methods, typically in discussion group settings, where participants are invited to exchange views and explain why people with one characteristic should be given higher or lower priority than others. This process is useful in ascertaining that the citizen perspective, as opposed to the consumer perspective, is being used by participants.

Since qualitative studies do not generate specific, quantitative values for weights, they must be followed by quantitative elicitation exercises. These have often used the person tradeoff method explained above, or the benefit tradeoff method (see Tsuchiya *et al.* 2003 for an example). Benefit tradeoff questions are similar to person tradeoff questions, but vary the size of the benefit instead of the number of people so as to reach a point of indifference between the two groups; obviously, it cannot be used for HRQOL valuation, but it can be used to elicit the relative values of different population characteristics. Moreover, adaptations of conventional methods (e.g. standard gamble and time tradeoff, using scenarios for groups of patients as opposed to individual respondents themselves) are possible. The key in all such cases would be to adopt a societal perspective. Respondents would be asked to behave as informed citizens who disregard information relevant to their own situation and personal preferences, but retain general understanding of the ways and the values of their society, as in the "thin" veil of ignorance. Alternatively, they could be asked to imagine themselves as committee members with the task of making the best decision for society, detaching themselves from their own personal interests. The obvious issue regarding this exercise is the extent to which actual elicitation exercises can be made genuinely disinterested and fair.

Treating some people rather than others can also have knock-on effects in terms of efficiency. For instance, in a serious crisis it makes more sense to save the life of a self-supporting adult than that of an elderly person or a young child who will need support from others to survive further. The above non-welfarist framework for deriving equity weights could also be used for deriving "efficiency weights", where a QALY accruing to individuals of a more "important" group within the population is given a larger weight than the rest. As with the elicitation of equity preferences, there will be concerns over the process of weighing the importance of the health and survival of various people and trying to attach relative efficiency weights to them. If such weights are to be incorporated in cost per QALY analyses, they also need to be based on a non-welfarist approach, where values of the informed citizen are elicited from an impartial and detached perspective.

Again, since welfarism is embedded in the selfish consumer's utility, it is difficult to see how equity weights or efficiency weights can be set in a fair manner. A possible challenge to this might be that there is nothing intrinsically wrong with using selfish consumers' utilities as one input, but not necessarily the sole input, to the analysis; and this may well be the case. The issue then is, if personal utility is not the sole input, then where are the other inputs to come from? At some point in the process of deriving these weights, there needs to be consideration of whose well-being should count more or less compared with others, and by how much, and if this judgment is to be fair, then it cannot be left to selfish agents. The judgment needs to be made by disinterested parties—in other words, from the non-welfarist perspective.

22.4 Concluding Remarks

The intellectual backdrop against which a large part of economic evaluation of health-care interventions is carried out is one where putting a monetary price tag on human life and health is often categorically regarded as immoral and unacceptable. Economists could go into long lectures beginning with the concept of opportunity cost, followed by how introducing a monetary value of health is unavoidable, and how failing to do so will lead to wasteful use of limited resources and thus to fewer lives saved and less health recovered, which, presumably, will also be immoral and unacceptable, if not more so. In the real world, practicing health economists have instead introduced a form of economic evaluation that does not explicitly introduce monetary values of health within the analysis. This, obviously, does not avoid the issue of the monetary value of health altogether, since it comes back in the form of a threshold cost per QALY amount, beyond which an intervention will be regarded as not cost-effective enough to be funded.

For non-welfarist health economists, this is largely an acceptable state of affairs, since it is relatively straightforward to accept the number of QALYs gained as the distribuendum. Welfarist health economists, on the other hand, have two choices. One is to argue for cost–benefit analyses, which have a more solid theoretical foundation in welfare economics, and are less restrictive in many respects. The other is to explore a welfarist interpretation of cost per QALY analyses, and of the concept of the QALY. But why bother with the QALY in the first place? What is the attraction of the QALY to a welfarist, when the restrictions imposed on personal utility functions are more severe compared with representing the value of health in monetary terms?

There may be two possible motivations for this second enterprise, exploring a welfarist foundation for QALYs. One, the more likely of the two, is the practical element, that most economic evaluations of health-care interventions are carried out in the form of cost per QALY analyses, and that there are few cost–benefit analyses by comparison. Given that the health-care sector is a significant player in the economy, there should be a way to understand how choices are made in this sector from a welfarist perspective. The other, more speculative motivation could be the very way in which the QALY requires restrictive assumptions. While welfarism may struggle to incorporate inequality aversion and equity weights into the health-related social welfare function, standard unweighted cost per QALY analyses imply one important condition: viz. that everybody's QALY counts the same, regardless of who they are. In this respect, it is highly egalitarian compared with, for example, a compensating variation for changes in one's own health, which will most certainly be a function of disposable income. Pure welfarism has low affinity with equity, but welfarists need not be anti-equity. The welfarist QALY may be an attractive concept in this regard.

In the meantime, non-welfarist health economists have other concerns regarding equity, and as we saw above, the debate has moved on from the mantra: a QALY is a QALY is a QALY. The new mantra seems to claim: a QALY is not a QALY, depending on how much health you already have, who you are, and why you are ill. The interest is in identifying the relevant characteristics that makes your QALY different from mine, and developing a method to quantify by how much they should differ.

References

Arrow. K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*, 53, 941–73.

Beveridge, W. (1942). *Social Insurance and Allied Services*. London: HMSO.

Bleichrodt, H. (1997). Health Utility Indices and Equity Considerations. *Journal of Health Economics*, 16, 65–91.

—— and Pinto, J. L. (2006). Conceptual Foundations for Health Utility Measurement. In A. M. Jones (ed.), *The Elgar Companion to Health Economics*, 347–58. Cheltenham: Edward Elgar.

—— Diecidue, E., and Quiggin, J. (2004). Equity Weights in the Allocation of Health Care: The Rank-Dependent QALY Model. *Journal of Health Economics*, 23, 157–71.

—— doctor, J., and Stolk, E. (2005). A Nonparametric Elicitation of the Equity–Efficiency Trade-off in Cost–Utility Analysis. *Journal of Health Economics*, 24, 655–78.

—— Wakker, P., and Johannesson, M. (1997). Characterizing QALYs by Risk Neutrality. *Journal of Risk and Uncertainty*, 15, 107–14.

Brazier, J., Akehurst, R., Brennan, A., Dolan, P., Claxton, K., McCabe, C., Sculpher, M., and Tsuchiya, A. (2005). Should Patients Have a Greater Role in Valuing Health States? *Applied Health Economics and Health Policy*, 4, 201–8.

—— Ratcliffe, J., Salomon. J. A., and Tsuchiya, A. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford: Oxford University Press.

Camacho, A. (1979). On Cardinal Utility. *Theory and Decision*, 10, 131–45.

—— (1980). Approaches to Cardinal Utility. *Theory and Decision*, 12, 359–79.

Cookson, R. (2005). QALYs and the Capabilities Approach. *Health Economics*, 14, 817–29.

Culyer, A. J. (1989). The Normative Economics of Health Care Finance and Provision. *Oxford Review of Economic Policy*, 5/1, 34–58.

—— Wagstaff, A. (1993). Equity and Equality in Health and Health Care. *Journal of Health Economics*, 12, 431–57.

—— Lavers, R. J., and Williams, A. (1971). Social Indicators: Health. *Social Trends*, 2, 31–42.

Daniels, N., and Sabin, J. (1997). Limits to Health Care: Fair Procedures, Democratic Deliberation, and the Legitimacy Problem for Insurers. *Philosophy and Public Affairs*, 26, 303–50.

Department of Health. (1999). *Saving Lives: Our Healthier Nation*. London: HMSO.

Doctor, J. N., Bleichrodt, H., Miyamoto, J., Temkin, N. R., and Dikmen, S. (2004). A New and More Robust Test of QALYs. *Journal of Health Economics*, 23, 353–67.

Dolan, P. (1997). Modeling Valuations for EuroQol Health States. *Medical Care*, 35/11, 1095–1108.

—— and Tsuchiya, A. (2006). The Elicitation of Distributional Judgements in the Context of Economic Evaluation. In A. M. Jones (ed.), *The Elgar Companion to Health Economics*, 382–91. Cheltenham: Edward Elgar.

—— Shaw, R., Tsuchiya, A., and Williams, A. (2005). QALY Maximisation and People's Preferences: A Methodological Review of the Literature. *Health Economics*, 14, 197–208.

Drummond, M., Sculpher, M., Torrance, G. W., O'Brien, B. J., and Stoddart, G. L. (2005). *Methods for the Economic Evaluation of Health Care Programmes,* 3rd edn. Oxford: Oxford University Press.

Feldstein, M. S. (1963). Economic Analysis, Operational Research, and the National Health Service. *Oxford Economic Papers*, 15/1, 19–31.

Fryback, D. G., and Lawrence, W. F. Jr (1997). Dollars may not Buy as Many QALYs as we Think: A Problem with Defining Quality-of-Life Adjustments. *Medical Decision Making*, 17, 276–84.

Fuchs, V. R. (1974). *Who Shall Live?* New York: Basic Books.

Gold, M. R., Siegel, J. E., Russell, L. B., and Weinstein, M. C. (eds.) (1996). *Cost-Effectiveness in Health and Medicine*. Oxford: Oxford University Press.

Hurley, J. (2000). An Overview of the Normative Economics of the Health Sector. In A. J. Culyer and J. P. Newhouse (eds.), *Handbook of Health Economics*, iA: 55–118. Elsevier Amsterdam: Science.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to Bentham? Exploration of Experienced Utility. *Quarterly Journal of Economics*, 112/2, 375–405.

Maynard, A., and Kanavos, P. (2000). Health Economics: An Evolving Paradigm. *Health Economics*, 9, 183–90.

McGuire, A., Henderson, J., and Mooney, G. (1988). *The Economics of Health Care: An Introductory Text*. London: Routledge.

McKeown, T. (1976). *The Modern Rise of Population*. New York: Academic Press.

Miyamoto, J. M. (1999). Quality-Adjusted Life Years (QALY) Utility Models under Expected Utility and Rank-Dependent Utility Assumptions. *Journal of Mathematical Psychology*, 43, 201–37.

—— and Eraker, S. A. (1985). Parameter Estimates for a QALY Utility Model. *Medical Decision Making*, 5, 191–213.

—— —— (1989). Parametric Models of the Utility of Survival Duration: Tests of Axioms in a Generic Utility Framework. *Organizational Behavior and Human Decision Processes*, 44, 166–202.

Mooney, G. (1992). *Economics, Medicine and Health Care*. Brighton: Harvester Wheatsheaf.

Nord, E. (1995). The Person Trade-Off Approach to Valuing Health Care Programmes. *Medical Decision Making*, 15, 201–8.

Pauly, M. V. (1994). Editorial: A Re-examination of the Meaning and Importance of Supplier-Induced Demand. *Journal of Health Economics*, 13, 369–72.

Pliskin, J. S., Shepard, D. S., and Weinstein, M. C. (1980). Utility Functions for Life Years and Health Status. *Operations Research*, 28, 206–24.

Sen, A. (1979). Personal Utilities and Public Judgements: Or What's Wrong with Welfare Economics. *Economic Journal*, 89, 537–58.

Sugden, R., and Williams, A. (1978). *The Principles of Practical Cost–Benefit Analysis*. Oxford: Oxford University Press.

Torrance, G. W., Thomas, W., and Sackett, D. (1972). A Utility Maximization Model for Evaluation of Health Care Programs. *Health Services Research*, 7, 118–33.

Tsuchiya, A., and Dolan, P. (2005). The QALY Model and Individual Preferences for Health States and Health Profiles over Time: A Systematic Review of the Literature. *Medical Decision Making*, 25/4, 460–7.

—— and Williams, A. (2001). Welfare Economics and Economic Evaluation. In M. Drummond, and A. McGuire (eds.), *Theory and Practice of Economic Evaluation in Health Care*, 22–45. Oxford: Oxford University Press.

—— Dolan, P., and Shaw, R. (2003). Measuring People's Preferences Regarding Ageism and Health: Some Methodological Issues and Some Fresh Evidence. *Social Science & Medicine*, 57/4, 687–96.

—— Miguel, L. S., Edlin, R., Wailoo, A., and Dolan, P. (2005). Procedural Justice in Public Healthcare Resource Allocation. *Applied Health Economics and Health Policy*, 4, 119–27.

von Neuman, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wailoo, A., and Anand, P. (2005). The Nature of Procedural Preferences for Rational Health Care Decisions. *Social Science & Medicine*, 60, 223–36.

Williams, A., and Cookson, R. (2000). Equity in Health. In A. J. Culyer and J. P. Newhouse (eds.), *Handbook of Health Economics*, iB: 1863–1910. Amsterdam: Elsevier Science.

—— Tsuchiya, A., and Dolan, P. (2005). Eliciting Equity–Efficiency Tradeoffs in Health. In P. C. Smith, L. Ginnelly, and M. Sculpher (eds.), *Health Policy and Economics: Opportunities and Challenges*, 64–87. Buckingham: Open University Press.

World Bank. (1993). *World Development Report 1993: Investing in Health*. Oxford: Oxford University Press.

World Health Organization (1981). *Global Strategy for Health for All by the Year 2000*. Geneva: WHO.

CHAPTER 23

THE CAPABILITIES APPROACH

ERIK SCHOKKAERT

23.1 Introduction

The origins of the capabilities approach within welfare economics are to be found in a series of influential papers and monographs, written by Amartya Sen in the early 1980s (Sen 1980, 1985*a*; Sen *et al.* 1987). He developed and discussed the approach further in some widely read books (Sen 1992, 1999; Nussbaum and Sen 1993). The basic purpose of the approach is neatly summarized in the preface to the seminal monograph *Commodities and Capabilities*: "to present a set of interrelated theses concerning the foundations of welfare economics, and in particular about *the assessment of personal well-being and advantage*" (Sen 1985*a*; my italics). At first sight, this may seem a purely descriptive exercise. However, normative considerations were crucial from the very beginning. The introduction of the capabilities idea was meant to be an answer to the question "Equality of what?" (Sen 1980). The basic idea is to find a definition of personal well-being and advantage that can be used in a meaningful way as the equalisandum for an egalitarian policy (or, in a less egalitarian approach, as the basic concern for policymakers).

Sen's answer to the question "Equality of what?" introduces two basic notions. What matters to define *well-being* are the *functionings* of a person, i.e. her achievements: what she manages to do or to be (well-nourished, well-clothed, mobile, taking part in the life of the community). According to him, however, more important

than well-being is the *advantage* of the person, i.e. her real opportunities. These are called *capabilities*. These ideas were not new.[1] Moreover, the basic intuitions captured by the ideas of functionings and capabilities are closely related to the multidimensional approaches to the quality of life and to deprivation which were prominent in the social sciences long before Sen introduced his concepts in the early 1980s.[2] Yet, it is undoubtedly true that the growing acceptance of these ideas within (welfare) economics started with the seminal contributions of Sen. He was the first to translate the intuitions about multidimensional measurement of quality of life into the language of welfare economics, comparing them explicitly with traditional economic concepts such as income and utility. Moreover, he related the discussion about "Equality of what?" in a coherent way to the informational approach to social welfare functions and to the growing discussions about the limitations of welfarism.

The influence of the idea of capabilities soon went far beyond welfare economics, and even far beyond economics. It became the inspiration for a large multidisciplinary effort to understand better the ideas of "well-being" and freedom and their relation to development.[3] This growing popularity has (unavoidably) gone together with a proliferation of the number of possible interpretations. The discussion now brings together analytical welfare economists, exploring more deeply the framework introduced by Sen (1985*a*), as well as critical scientists who identify themselves as heterodox economists and are keen to reject mathematical or even analytical approaches as being overly restrictive. One strand of the empirical work aims at developing quantitative techniques to measure functionings and capabilities; another strand advocates the implementation through participative focus groups. In fact, the whole framework is often presented as a broad framework of thought, rather than as a sharp analytical tool (Robeyns 2006*a*). It is difficult to evaluate this whole movement. I will therefore be much less ambitious and go back to the starting point: how do the ideas of functionings and capabilities contribute to the welfare economic debate about "equality of what"? How does the growing experience with empirical applications contribute to a better understanding of the basic methodological issues? And what questions have remained open until now? After a brief overview of the main concepts in Section 23.2, I will discuss various methodological issues in Sections 23.3–23.6. In each case I will try to confront the theoretical challenges with the available empirical experience. Section 23.7 concludes.

[1] Basu and Lopez-Calva (forthcoming) give a brief sketch of the history of the ideas, linking it to Aristotle, Marx, Berlin, Smith, and Rawls.

[2] Cummins (1996) covers 1500 articles related to multidimensional approaches to quality of life, in an attempt to check his own definition of relevant domains.

[3] There is now even a successful Human Development and Capability Association. Launched in September 2004, it aims at "promoting research from many disciplines on key problems including poverty, justice, well-being, and economics". The Association has its own journal (*Journal of Human Development*) and an already impressive membership.

23.2 Equality of What? Capabilities as a Way of Assessing Individual Advantage

The *well-being* of a person has to be evaluated on the basis of what he or she manages to do or to be. These "functionings" have to be distinguished from the commodities which are used to achieve them, because personal features matter a lot in the transformation from objective characteristics of the commodities to functionings. The nutritional value of food depends on the biological characteristics of the body; books do not contribute much to the personal development of persons who were never taught to read. Because a focus on the possession of material commodities neglects these crucial inter-individual differences, it is not acceptable as a description of well-being.

Sen (1985*a*) gives a first and very useful formalization of these concepts. The achieved functionings vector b_i of individual i can be written as

$$b_i = f_i(c(x_i)) \tag{1}$$

where x_i is the vector of commodities possessed by person i, $c(.)$ is the function converting the commodity vector into a vector of objective characteristics, and $f_i(.)$ is a personal utilization function of i reflecting one pattern of use that i can actually make. While the $c(.)$ function has to be interpreted in the Gorman (1956)–Lancaster (1966) tradition and is independent of the individual concerned, the transformation of these characteristics into functionings is individual-specific. The well-being of person i can then be seen as the valuation of the vector of functionings b_i:

$$v_i = v_i(f_i(c(x_i))) \tag{2}$$

Sen emphasizes that the valuation function $v_i(.)$ can represent a partial ordering.

The interpretation of $v_i(.)$ is crucial. If we interpret the valuation exercise as objective and as the same for all individuals, we could drop the individual subscript. If we introduce the possibility of inter-individual differences and therefore keep the subscript, $v_i(.)$ is formally similar to a utility function $u_i(x_i)$, since it can also be seen as the representation of a (possibly partial) ordering of commodity bundles x_i. However, in Sen's view, it is necessary to distinguish the valuation of functionings vectors from the utility derived from it. He distinguishes different possible interpretations of utility.

(a) The first defines utility on the basis of "revealed preference" and choice. This is the most popular approach in modern welfare economics, but it is really a nonstarter. The assumption that choices are motivated only by personal well-being is heroic. Moreover, as is well known, the revealed preference approach

cannot easily accommodate interpersonal comparisons of well-being. Yet such interpersonal comparisons are indispensable for the purpose of defining an acceptable equalisandum.

(b) The second and the third interpretations are closely related and are situated in the traditional utilitarian interpretation: one interprets utility as subjective happiness (pleasure and pain), the other as the extent to which desires are fulfilled. As representations of well-being, they both entail similar problems. The first problem is what Sen calls "physical condition neglect": utility is grounded only on the mental attitude of the person, and does not sufficiently take into account the real physical conditions of the person. This has two aspects. One is the issue of expensive tastes; the other is that persons may adapt to their objective circumstances or realistic expectations: "A person who is ill-fed, undernourished, unsheltered and ill can still be high up in the scale of happiness or desire-fulfillment if he or she has learned to have "realistic" desires and to take pleasure in small mercies" (Sen 1985*a*, p. 21). The second problem is "valuation neglect". Valuing a life is a reflective activity in a way that "being happy" or "desiring" need not be (Sen 1985*a*, p. 29). An acceptable approach to well-being should explicitly take into account this valuational activity by the persons themselves. This is not to say that "happiness" or "desire-fulfillment" cannot be important components of well-being. But they are only part of the story. The most adequate way of taking them into account is to see them as elements of the vector b_i.

In a further step, Sen claims that a description of individual living standards in terms of achievements is not sufficient, because one has also to introduce the notion of freedom. He therefore proposes the concept of the *advantage* of a person, i.e. his or her real opportunities. The person can choose the utilization function $f_i(.)$ from an individual-specific set F_i. If we assume, moreover, that his choice of commodity vectors is restricted to his "entitlements" X_i, we can represent his real freedom by the set of feasible functioning vectors

$$Q_i(X_i) = [b_i \mid b_i = f_i(c(x_i)), \text{ for some } f_i \in F_i \text{ and for some } x_i \in X_i] \qquad (3)$$

Q_i can then be called the "capabilities" of person i. Sen is quite explicit about the importance of the move from functionings to capabilities. The typical example is the comparison between two individuals who are both undernourished. For the first individual, the undernourishment is the result of his material deprivation. The second individual is wealthy, but freely decides to fast for religious reasons. While their achievements in terms of nutritional functioning are identical, it seems clear that their situations are not equivalent from an egalitarian point of view.

Equalization of capabilities goes beyond equalization of opportunities in the narrow sense of the word, and also beyond removal of discrimination, although both are important elements of it. Capabilities are a reflection of the real (positive) freedom of individuals, and should not be restricted to the securing of negative

freedoms. Persons should not only have the legal right to provide themselves with food; they should also have the economic possibilities to do so. Although Sen emphasizes the importance of freedom, his approach is definitely not contractarian, but remains firmly consequentialist (Sugden 1993).

The capabilities approach is not a complete theory of justice. Although the writings of people using it have an outspoken egalitarian flavor, in principle it can be integrated into many different theories. One can formulate a concave social welfare function in terms of individual capabilities levels. But functionings can also be the outcome measure used in the theory of equality of opportunity as introduced by Roemer (1998) or in theories of responsibility-sensitive egalitarianism (Fleurbaey 2008). It is possible to trade off considerations of well-being or advantage for other dimensions (such as respect for political rights or for property rights). In all these cases, the specific application will depend on the exact content given to the functionings or capabilities themselves, which remains very open. Rather than a theory of justice, the capabilities approach is a proposal for the evaluative space which should be used for policy purposes.

It may still have some direct implications for policy issues. Take the issue of relative versus absolute poverty (Sen 1983).[4] Introducing the idea of capabilities suggests an approach in which poverty is absolute in the space of capabilities, but relative in the space of resources and commodities. While a functioning such as social integration (being able to appear in public without shame) has an absolute core, because it is important in all societies and at all times, at the same time the commodities needed to realize this functioning will be widely different in different societies and at different times. Relative deprivation in the space of commodities can go together with absolute deprivation in the space of capabilities (or functionings).

Since the mid 1980s there have been many empirical studies trying to implement this theoretical framework. Some of them are at the country level. Already in Sen (1985*a*) one important example was a comparison of the performance of India, Sri Lanka, and China. Since then, the most influential application is undoubtedly the Human Development Index, used by the UNDP to measure the well-being of countries in terms of adjusted GDP per capita, life expectancy at birth, and educational performance. Closer to the original intuition of the approach—which is to measure poverty and well-being at the individual level—are studies with individual data. This is a booming domain of research, and the number of such studies is rapidly growing.[5]

[4] The somewhat personalized exchange of ideas between Townsend (1985) and Sen (1985*b*) following this article (Sen 1983) shows that the approach is definitely not empty.

[5] These include among others (and in chronological order) Schokkaert and Van Ootegem (1990); Lovell *et al.* (1994); Balestrino (1996); Ruggeri Laderchi (1997); Brandolini and D'Alessio (1998); Chiappero Martinetti (2000); Klasen (2000); Phipps (2002); Qizilbash (2002); Anand *et al.* (2005, 2009); Kuklys (2005); Lelli (2005); Qizilbash and Clark (2005); Ramos and Silber (2005);

Although these studies are very diverse, two important conclusions can nevertheless be drawn. The first is a positive one. There is by now overwhelming empirical evidence that the multidimensional approach adds something to the traditional approaches in terms of GDP or income. The ranking of countries on the basis of multidimensional well-being is strikingly different from the ranking on the basis of GDP per capita. The identification of poor individuals and groups changes when one introduces more dimensions. The second conclusion is more tentative. Many of the earlier empirical studies were only loosely connected to the theoretical framework, and only recently has empirical work sought to operationalize directly some of the key distinctive parts of the approach. Furthermore, a number of lists of functionings and capabilities have been proposed, and this has made it difficult for researchers to settle on a particular set of dimensions with which to measure welfare or deprivation. In any case, more work is needed to bridge the gap between the theory and the empirical applications. More specifically, it is striking that there is almost no empirical research using a full explanatory model, which specifies the relationship between achieved functionings and capabilities and explores how achievements are influenced by psychological characteristics and by features of the external and social environment.[6] Estimation of such a structural model of behavior would make it possible to go beyond a mere descriptive exercise and could give a better insight into the (perceived) tradeoffs between the different capabilities.

Recent advances on the theoretical side suggest that there are no straightforward answers to the methodological challenges raised by empirical work. The analytical questions raised in Sen (1985*a*) are not yet answered in a fully satisfactory way, and important new questions have arisen.[7] In Section 23.3 I will discuss the issues of freedom, responsibility, functionings, and capabilities. In Section 23.4 I will describe and compare different ways of selecting the elements of the functionings vector. Section 23.5 will be devoted to the indexing problem, i.e. the operationalization of the function $v_i(.)$ and the differences with a utility function. Finally, in Section 23.6, I summarize some recent findings concerning the aggregation problem, i.e. the measurement of capabilities at the country level. The need to think in terms of a complete structural model will be a thread throughout my discussion.

Zaidi and Burchardt (2005); Anand and van Hees (2006); Robeyns (2006*b*); Anand and Santos (2007); and Lelli (2008). The website of the Human Development and Capability Association (<www.capabilityapproach.com>) contains a regularly updated overview of the empirical work.

[6] Kuklys (2005) contains a structural model, estimated with individual data. Krishnakumar (2007) and Anand, Santos, and Smith (2008) combine a latent class approach with data at the country level.

[7] I do not claim any originality for this list of issues. They were present in the debate about the capabilities approach from the very start. See also Robeyns (2006*a*).

23.3 Capabilities, Functionings, and Responsibility

Fasting voluntarily and starving as a result of economic deprivation are obviously not equivalent from a policy point of view. Sen's argument that we should go beyond functionings to introduce considerations of freedom is a strong one. The real question is how to do this. In this respect it is important to note that Sen from the very beginning and throughout his work has pointed out that defining capabilities in terms of opportunity sets is not the only possible way to incorporate freedom into the analysis. An alternative is to work with what he calls "refined functionings" or comprehensive outcomes, where the "refinement" refers to the operation of taking note of the alternatives available or of the process of choice itself. Consider again the fasting–starving example. The one fasting is choosing to eat less; the poor starving person is exercising no choice at all. These can be seen as two different "refined" functionings—choosing A when B is also available is a different refined functioning from choosing A when B is not available (Sen *et al*. 1987, pp. 36–7). Or one could consider, in addition to the functioning of being well-nourished or not, another functioning: "exercising choice with respect to what one eats". Again, the description of the situation of the one who fasts and the one who starves would be different. In this section I will compare these two approaches: on the one hand the "opportunity set" approach, on the other hand the "refined functionings" approach. I will first discuss three conceptual and ethical points raised by the move from functionings to opportunity sets (as in Eq. 3) and then turn to the issue of application.

The *first* point relates to the fact that to evaluate capabilities as in (3), it is necessary to evaluate *sets*. One possible approach would be to define the value of a set of functioning vectors by the value of the best element in that set. Sen (1985*a*) calls this method "elementary evaluation", but immediately acknowledges that it does not do justice to the basic idea of freedom—indeed, removing from a set all but the best alternatives would in this case not reduce its value. Moreover, how define the "best" element?

Moving beyond elementary evaluation, however, raises some tricky issues, which were illustrated in a striking way in an influential article by Pattanaik and Xu (1990). They show that acceptance of a set of reasonable looking axioms implies the so-called cardinality-based ordering, which simply ranks two sets on the basis of the number of elements in the sets. In his reply, Sen (1990) pointed out that this disappointing result is due to the fact that the axioms imposed exclude the possibility of taking into account the "quality" of the alternatives in the set. To give an example, one of the axioms states that two opportunity sets with no choice (i.e. containing only one element) are equally valuable from the point of view of freedom. Sen then considers the situation of a person who has two alternatives in

going home from the office by taking a short walk: (1) she can hop on one leg to home, but she is not permitted to walk; (2) she can walk normally to home, but she is not permitted to hop on one leg. Given that she vastly prefers to walk, it is strange to claim that she has no less freedom when she is forced to hop. To integrate such considerations, it seems necessary to introduce preferences into the analysis. Yet, as soon as one introduces preferences, a new series of questions pops up: do we take into account actual individual preferences or the preferences that may emerge at some point in the future or the preferences that a reasonable person in that situation can possibly have? And should we in this case consider subjective feelings or cognitive valuations? These philosophical issues are not yet settled in the growing literature on the topic.[8] The problem of the evaluation of opportunity sets remains open.

Can we avoid it through the use of (refined) functionings? I suggested already that the famous fasting–starving distinction can be taken care of in a satisfactory way. Fleurbaey (2005) extends this idea and argues that all the relevant aspects of freedom can be captured through functionings. Basic freedoms of thought, speech, political activity, travel, etc. are obviously part of the functioning vector, and the same is true for the freedom to engage in economic activities. The (crucial) distinction between formal and real freedoms can be made operational by considering individual achievements in terms of education, income, and social relations. The freedom from avoidable disease can be approximated in terms of the achieved health functioning, of the accessibility of the health-care system, and of the environmental and social factors influenced by public health policy. The examples immediately show that the refined functionings approach also raises difficult challenges. Understanding the "process of choosing" is not straightforward. As soon as one has to resort to indirect indicators (such as education, income, social relations, accessibility of the health-care system), one has to think carefully about the specific social, environmental, and individual variables that determine the influence of these indicators. In moving from "capability sets" to "refined functionings", we move from the problem of evaluating sets to the problem of investigating carefully the process of "producing" refined functionings. To make progress on these issues, the construction of better structural models of choice behavior is badly needed.

The *second* question is whether it is sufficient to look at capability sets or whether, on the contrary, we also have to consider achievements in addition to capabilities (Fleurbaey 2005, 2006). Consider the situation of two persons with the same opportunity set, i.e. the same capabilities. However, the first ends up with an achieved functionings vector, which is dominating the achievements of the other. Would we claim that from an ethical point of view their situations are equivalent? The answer can only be yes if one holds the persons fully responsible for the choices they make

[8] See Barberà *et al.* (1998) for an overview of the literature.

within their opportunity set. This can be a very harsh position, given the well-documented limitations of individual decision-making capacities. Freedom implies individual responsibility, but we have to face the question of how to define the ethical limits of this responsibility. The problem becomes even more difficult when we introduce the time dimension into the analysis: the opportunity sets of older people are heavily influenced by decisions they took when they were young. How long do individuals have to remain responsible for "mistakes" committed earlier in life?

Again there are two possible paths to take. Sen (2002, p. 83) proposes to focus in the opportunity set approach on "the *actual* ability to achieve". This means that limited decision-making capacities, e.g. due to social background, should be integrated into the definition of the opportunity sets or into the procedure for evaluating them. This is not trivial, given the present state of our knowledge about evaluation of sets. The other path is again to broaden the description of achievements to "comprehensive outcomes", including the process of choice. Here also, as noted already, there are difficult questions to be answered. However, it seems that the notion of refined functionings is better suited to the careful empirical analysis which is needed to begin to answer these questions about choice, well-being, and differences in opportunities.[9]

The *third* issue was already raised by Basu (1987), and is taken up again in Basu and Lopez-Calva (forthcoming).[10] In general, the achieved functionings of any person do not depend only on the choices made by that individual, but also depend on actions taken by other individuals. Take two game situations in which one person has an identical strategy set in each game, but the strategy sets of other players differ. How to compare the capabilities of that person in these two games? This conceptual issue of *defining* the capabilities in a setting of social interdependencies goes much deeper than the obvious point that individual well-being and advantage depend on the social environment of the persons.

How do "refined functionings" fare in this regard? Again, it seems that they may offer promising prospects, precisely because the concept is less ambitious and does not necessitate the full description of the opportunity sets from which different persons can choose. In fact, the discussion about capabilities sets and refined functionings shows some similarity to the discussion about modeling individual rights: the problems involved in defining opportunity sets in a setting with social interdependencies are related to the problems with the definition of rights in terms of social choice (originally introduced by Sen, 1970); the approach in terms of refined functionings bears some similarity to the procedure of modeling rights in terms of game forms (Gaertner, Pattanaik, and Suzumura 1992).

[9] Alkire (2005) gives an interesting overview of direct questionnaire approaches to measuring human agency (autonomy and self-determination).

[10] They illustrate the point in the Edgeworth box of a two-person two-good exchange economy.

Until now I have focused mainly on conceptual questions related to the choice between opportunity sets and refined functionings. From the point of view of application, there is the additional issue of observability. What we observe are achieved functionings, because these can be derived from the actual (observable) way of living of the person. We can also derive from observations some direct or indirect indicators of the degree to which the individual had the freedom to choose. Again, the example of the person starving because of deprivation and the person voluntarily fasting for religious reasons illustrates the point, as it essentially takes for granted that the environment contains sufficient observable clues to distinguish between the two situations in a reliable way. Things are very different with respect to the concept of opportunity sets: opportunities that are not chosen are not realized. Therefore describing opportunities requires consideration of counterfactual states which cannot be directly observed (Fleurbaey 2005).

These remarks seem to suggest that the perspectives for interesting empirical work on capabilities look bleak. However, recent work (involving researchers from economics, philosophy, psychology, and politics), has shown that conventional survey methods can be very useful for assessing the extent of a person's capability set. Initially, this work focused on a distinction between achievement and scope, in a small number of life domains (Anand and van Hees 2006). Subsequently, a range of standard household surveys were examined, and it was concluded that some of the secondary datasets widely used by social scientists do in fact contain information on what people *can* do, what they have access to, as well as on the degree and source of the constraints they face. Variations in these variables provide indicators of variations in people's capability sets. However, typically, the extant capability indicators in secondary datasets cover only a fraction of the dimensions that quality of life and poverty researchers might be interested in. Therefore it proved necessary to develop a survey instrument, including specfic indicators of capabilities. Such an instrument, based on over sixty indicators across a wide range of life domains (Anand *et al.* 2005*a*), has been used as the basis for two national surveys (in the UK and Argentina), is now being developed into a short-form questionnaire by public health researchers in Glasgow, and is being incorporated into a project on mental health and coercion by researchers in Oxford.

This work yields some important insights about the scope for empirical progress in this area. First and foremost, the researchers point out that while direct option enumeration (measurement) is probably not usually feasible, the availability and use of self-report data, including information on opportunities, abilities, and constraints (indicators) relating to particular life dimensions is in fact widespread. As an example, they note that virtually all "income" data in household surveys, often used to provide an indicator of consumption opportunities, is based on self-report. Furthermore, non-income-based capability indicators may be superior for some purposes, as they can be less susceptible to high-powered incentives to

misrepresent. Second, while subjectivity of data sources is inevitable, this is not a problem *per se*, so long as the implications for appropriate research methods and questions are carefully understood. One concern about subjectivity within regression models surrounds endogeneity due to omitted variables but this is something that can be tested for and instrumented (Anand *et al.* 2009) or addressed by incorporating data on personality within single wave surveys (Anand *et al.* 2008), by merging datasets with national data on regional variations in a variety of opportunity related variables (Anand and Santos 2007) or by moving to panel data.

It is perhaps too early to provide a definitive assessment of the impact of this latter empirical work, but the production of new data, the analysis of associated econometric issues, and the discussion of methodological issues concerning the production of welfare statistics do seem to open up a broad field of potentially fruitful and innovative empirical research.

That said, the rest of the chapter will revert to using the term "capabilities" in a looser way, which can capture both the approach in terms of opportunity sets and the approach in terms of refined functionings.

23.4 How to Select the Relevant Functionings or Capabilities?

While the available evidence clearly shows that the move to a multidimensional framework is a considerable enrichment for policy analysis, there is no consensus about how to define the most adequate multidimensional space. Should one include all capabilities in the list, some of them possibly of minor importance? Or should one focus on a limited and abstract list of essential capabilities? How to set that list?

There are two "extreme" approaches to this problem. The first one is exemplified in the work of the philosopher Martha Nussbaum (2000, 2006). Inspired by Aristotle, she starts from an openly normative (or "objective") view about what constitutes human flourishing and defines a list of abstract essential capabilities on the basis of this a priori view. Of course, the translation of these abstract capabilities into implementable terms will depend on the specific social, cultural, and economic context, but it remains true that such essentially perfectionist approaches leave little room for inter-individual differences in opinions about what constitutes a good life. Consensus seems to be within reach when one remains at the level of abstract formulations, but soon crumbles down when one turns to more specific applications. A priori defined lists of capabilities are useful, because they provoke

debate and discussion, but they do not seem to offer a solid foundation for scientific analysis.[11]

Amartya Sen is the exponent of the alternative approach, in which the definition of the list of capabilities is deliberately left open, and has to be settled in a democratic process through public reasoning (see e.g. Sen 2004).[12] This dynamic process creates room for participation of the people concerned—on its own already a crucial capability. Yet, from an analytical point of view, it is not much of a help. First, when one makes the definition of capabilities itself dependent on the social and economic context, the whole approach becomes in some sense relative. One then loses one of the main advantages of the capabilities approach: that it reconciles an absolute view of well-being and poverty in the space of capabilities with a relative view in the space of economic resources.[13] Second, the real scientific challenge is to understand why some capabilities are more prominent in some situations than in others, on what basis people make decisions, how views about capabilities develop over time. For such an analysis, one needs at least some general frame of reference.

Although these two approaches seem to be at opposite sides of the spectrum, one should not exaggerate the differences. Philosophers in the first approach acknowledge, and even stress, that the specific content of the abstract capabilities has to be decided through a participatory process. And within the second approach, the process of participation and deliberation will usually start from some first a priori proposal. Yet, the main emphasis of the two approaches remains different. And, from an analytical perspective, neither of the two is very helpful.

The problem is well illustrated by the work of the empirical researchers who have had to soil their hands with defining specific lists of capabilities and functionings. In the empirical work based on surveys, the definition of the dimensions is largely data-driven.[14] Often the first problem is the reduction of a long and overlapping list of very specific indicators to some more basic underlying dimensions. Factor

[11] This is perhaps the right place to restate the point that this is not the main purpose of these authors. Alkire (2002, p. 194) sees the set of dimensions as "a nonpaternalistic and useful tool in addressing a number of knotty development problems—from participatory exercises to data collection drives, from national policy making initiatives to public debates—in a multidimensional fashion".

[12] Alkire (2001) has applied this approach in a participatory process for the evaluation of three small-scale development projects.

[13] The following example given by Sen (2004, p. 79) illustrates my point: "Given the nature of poverty in India as well as the nature of available technology, it was not unreasonable in 1947 to concentrate on elementary education, basic health, and so on, and not worry too much about whether everyone can effectively communicate across the country and beyond. However, with the development of the Internet and its wide-ranging applications, and the advance made in information technology (not least in India), access to the web and the freedom of general communication are now parts of a very important capability that is of interest and relevance to all Indians."

[14] Robeyns (2005) has proposed a procedure for selecting the list of capabilities in empirical work. Her procedural criteria are not based on a theoretical approach, however, but boil down to a list of checks to correct for the potential personal biases of the researcher (as she herself acknowledges).

analysis (Schokkaert and Van Ootegem 1990) and fuzzy set theory (Chiappero Martinetti 2000) have been proposed as possible tools. Lelli (2008) compares the two approaches on the same data set and finds that the results are not very different. This should not hide, however, that the two approaches reflect very different conceptions. One view sees the definition of the underlying dimensions as a measurement issue. There is one "true" value of the functioning, and each of the different specific indicators is approximating that true functioning with some measurement error. The other view interprets the definition of the underlying dimensions as a normative weighting problem. The indicators are informative in their own right, but the question is how important they are; i.e. what weight they should get in the construction of the more encompassing dimension. Factor analysis is meaningful only in the first perspective. It is a valuable measurement tool, but the statistical correlations between the specific items do not give any indication about their relative importance from a normative or substantial perspective. Fuzzy set theory is more difficult to locate in one of the two views. However, it fits better in the second than in the first. I will therefore return to it in the next section, in which I discuss the indexing problem.

Empirical work within the capabilities approach has led to a large variation in "lists" of capabilities, heavily dependent on the specific problem (which may already be problematic) and on the availability of data (which is worse). It is not surprising that a list of functionings relevant for the long-term unemployed (Schokkaert and Van Ootegem 1990) is very different from a list of functionings used to describe the well-being of children in different countries (Phipps 2002).[15] For specific policy purposes (improving the living standard of the unemployed or the well-being of children) this variation might even be desirable. Moreover, as suggested by Ramos and Silber (2005), the policy conclusions following from different lists may not be very different. But if we want to develop a convincing theory of well-being that can be used to analyze differences between different countries or social groups and (possibly long-run) historical developments, that would be helpful in formulating clearly the tradeoffs between different policy issues, and that could be integrated in a second-best analysis of policy measures in a world of asymmetric information, we should be more ambitious.

Some authors have tried to go further than the simple exploitation of existing data. I give two examples. Anand *et al.* (2005*a*) explicitly tried to operationalize Nussbaum's list of capabilities with survey data from the British Household Panel Survey. As noted, they point out that this survey does in fact contain some information on aspects of freedom from questions to do with how capable people

[15] For Schokkaert and Van Ootegem (1990) the list of refined functionings consists of social isolation, happiness, physical functioning, microsocial contact, degree of activity, and financial situation; Phipps (2002) works with the functionings birth weight, asthma, accidents, activity limitation, trouble concentrating, disobedience at school, bullying, anxiety, lying, hyperactivity. But activity means something quite different in the two lists.

feel, whether they have access to certain forms of transport when needed, and so on. For some of Nussbaum's capabilities, however, it is impossible to find a suitable indicator. For other capabilities only an indirect indicator is found—e.g. the capabilities related to senses, imagination, and thought are approximated by educational level. At the same time, some mental health and psychological locus of control questions do appear to be quite close in terms of meaning to theoretical issues of autonomy that have interested many researchers in this field.

Clark (2005) investigated through a small number of high-quality interviews how the South African poor perceive "development" (a good form of life). He concludes that space must be made for utility (defined broadly to include all valuable mental states) and for the intrinsic value of material things. A challenging example is Coca-Cola, which turns out to be very important to many poor respondents. While the nutritional value of Coca-Cola is low, it is "perceived as a superior first world product" (Clark 2005, p. 1353) and it is important "to achieve other important functionings such as relaxing, facilitating social life and enhancing friendships" (Clark 2005, p. 1354). But is "having the opportunity to drink Coca-Cola" really a crucial capability?

How to proceed from here? In my view, it is necessary to raise explicitly a series of conceptual questions—and then to try to get better insights through the estimation of structural explanatory models. First, how "subjective" should our concept of well-being be? Or, formulated somewhat differently, what is the place of psychological functionings? The larger the number of psychological functionings included in the list (or the greater the weight given to them), the larger the risk that the problem of "physical condition" neglect will reappear, and the more difficult the issue of "valuation neglect" will become. I give two examples. Social status may be a crucial functioning, but in most societies it depends on relative consumption levels, and in a certain sense even reintroduces the problem of expensive tastes (Robeyns 2006*a*): the CEO of a large firm may "need" a certain material life-style to be respected in his group of peers; a university professor in a philosophy or welfare economics department may perhaps earn more prestige through a sober life-style. Do we accept these "needs" in our definition of well-being? To give a second example, what about feelings of depression that are not obviously linked to physical conditions? Where to draw the line between real psychiatric problems (which most observers would include in the definition of well-being) and overly subjective reactions, which can be easily manipulated and are well within the sphere of private information? These questions are related to, but do not coincide with, the role of personal preferences in the definition of capabilities, to which I will return in the next section.

Secondly, how should we treat so-called social capabilities, which cannot be reduced to narrow individualistic considerations? Take the examples of "living in a just society" or "having the capability to engage in meaningful social relations".

Not only can it be argued that these capabilities should be part of an Aristotelian conception of the good life (as in Nussbaum's list), but they also turn out to be important from a psychological point of view.[16] Yet they are essentially dependent on the whole social environment. I do not achieve the functioning of "living in a just society" if I am treated in a just way myself: it is equally important that other persons in society are treated equally justly. This suggests that these capabilities can be evaluated only at the aggregate level. But different individuals may have widely different opinions about what constitutes a just society or about what are meaningful social relations. Perhaps the best we can do in these cases is to shift the focus away from (individual or aggregate) functionings and/or capabilities to the necessary social and political institutions which create room for different kinds of social relations or for an open and democratic debate about the content of justice.

Thirdly, and most importantly, the capabilities idea has been introduced as an answer to the normative question, equality of what? Ethical considerations are essential in the delineation of relevant capabilities (or refined functionings). More specifically, given that the ultimate purpose is *not* simply to derive the best possible descriptive measure of subjective well-being, it is impossible to avoid the question of individual responsibility. This question has different dimensions. While I argued before that holding persons responsible for all their choices would be a very harsh position, some responsibility for choice is unavoidably linked to the introduction of freedom. This means that at least some achievements should *not* be taken up in a concept of well-being that is meant to be used in an egalitarian perspective. The problem of responsibility for choices is a very tricky one from a philosophical point of view, but cannot be neglected from a policy perspective.[17] A different but equally important issue is the delineation of a personal sphere, in which government decisions should not intrude out of respect for privacy and for personal integrity. Some of the psychological capabilities appearing in Nussbaum's list (and in other lists) definitely seem to belong in this category. There are then two possible approaches. One is to make explicit that taking up some capabilities in the definition of advantage does *not* necessarily imply that there is need for direct government intervention if some individuals lack these capabilities. What the government has to do is to set the environmental and social conditions under which individuals can take up their own responsibility. This is basically Nussbaum's position. It requires a deep empirical analysis of the influence of the social environment on these "private" capabilities. The second is to include only

[16] Remember Lerner's (1980) hypothesis of the need to believe in a just world.

[17] In fact, while the philosophical question of responsibility for choices is conceptually very different from the incentive problem in a second-best world, the two are closely linked in the policy debate—and the concern for "responsibility" in public opinion and among decision-makers often is the translation of second-best intuitions.

those refined functionings in the definition of well-being which are part of social responsibility, i.e. to "carve out" room for individual responsibility by disregarding explicitly some functionings (Fleurbaey 1995). This latter position implies that the definition of the relevant functionings is not an empirical but a purely normative question.

The three questions raised are essentially of a conceptual nature. Yet empirical research can make a useful (and perhaps even necessary) contribution to answering them. First, unless one takes an extreme objective (perfectionist) approach to the definition of a good life, the opinions of the people concerned should matter in the definition of the relevant refined functionings and in the delineation of the sphere of personal responsibility. Structured empirical research about values in society may then be an interesting complement to participatory focus groups. Second, even if one rejects the idea that normative questions can be settled by empirical research, there are many empirical issues underlying the normative discussion. What is the empirical relevance of psychological functionings and of social pressure in consumption behavior? How do opinions about a just society differ? What are the most important features of the social environment that may help stimulate meaningful social relations? How can the government create conditions to help persons take autonomous decisions in their own private spheres? Building and estimating good structural models may give a better insight into the empirical relationship between abstract capabilities and specific indicators, and may show how individual achievements are linked, on the one hand, to the socioeconomic and environmental background, and on the other hand, to the psychological features of the process of choice and decision-making.

23.5 The Indexing Problem

Let us now take the following step. Suppose agreement is reached about the list of (refined) functionings or capabilities. Suppose too that we have perfect information about the level of functionings for all persons in society; i.e. we know all the vectors b_i (see Eq. 1). Is it then possible to construct with this information a one-dimensional indicator of the well-being or advantage of person i, as in Eq. 2?[18] Note that the construction of such an indicator is not necessary if the only purpose of the exercise is to get a richer description of the well-being of individuals than is possible with a purely monetary approach. In fact, this is the position implicitly taken by

[18] In Herrero (1996) the existence of "capability indices" is assumed.

the bulk of the empirical work on capabilities, which does not go further than the computation of the refined functionings vectors for the individuals (or countries) in the sample.

As soon as we want to go further than mere description, however, and use the capabilities approach to derive overall conclusions about welfare, poverty, and inequality in a given society, the possible tradeoffs between the different dimensions can no longer be neglected.[19] I mentioned already that Sen has always emphasized the difficulties involved in defining an overall index of advantage, and has suggested that the best one can hope for is to find a partial ordering. It is useful to distinguish two possible sources of these difficulties (Sugden 1993). One is to say that well-being and advantage are *objective* concepts, but that it is intrinsically difficult to define what is a good life. An alternative approach accepts that the valuation of functionings bundles should be at least partly based on the valuations or *preferences* of the persons themselves. In this view the difficulty of defining an index reflects the fact that it is not straightforward to find a kind of "overlapping consensus". As emphasized before, the whole idea of defining capabilities is in the first place a normative question; but this does not detract from the idea that in a society with different cultures and subcultures, it seems hard to defend a purely perfectionist approach. Therefore individual valuations of their living standard by the persons themselves should most probably play at least some role in the indexing exercise. This is the least one can expect in an approach emphasizing the importance of freedom.

Introducing individual valuations raises difficult problems, however. These are clearly illustrated by looking at the vector dominance relation, which was already proposed by Sen (1985*a*) as a good starting point for constructing a partial ordering. If person i is better off than person j for all functionings, it seems indeed natural to state that the advantage of i is not smaller than the advantage of j. Of course, the resulting partial ordering may be extremely incomplete. More important for our purposes, though, recent work has shown that the dominance relation soon comes into conflict with even a minimal respect for people's own opinions about what constitutes a good life (Brun and Tungodden 2004; Fleurbaey 2007; Pattanaik and Xu 2007). This reinforces the urgency of the question as to how to construct an indicator of individual well-being without returning to simple welfarism?

The fairness approach proposed by Fleurbaey (2005, 2007) is an ambitious and attractive framework for reconciling respect for individual preferences, ordinal noncomparability of preferences, and a maximal application of the dominance principle that is compatible with respect for individual preferences. But its empirical application is not straightforward, and some difficult philosophical issues

[19] Even the refusal to make any tradeoffs is a well-defined position about these tradeoffs.

remain unsolved at this stage.[20] Other theoretical approaches to the indexing problem do not really tackle the issue. Brun and Tungodden (2004) explicitly stick to the dominance principle, even when it comes into conflict with individual opinions. Gaertner and Xu (2006) work with star-shaped capability sets, and then define the standard of living as the distance between the frontier of these sets and a reference functioning vector. Given the state of the theoretical literature, it is not surprising that the empirical work also largely neglects the problem of how to treat differences in individual opinions about what constitutes a good life and usually assumes preference homogeneity (implicitly or explicitly).

There has recently been an upsurge in the methodological literature on multidimensional inequality and poverty measurement.[21] How should we reformulate the traditional Pigou–Dalton criterion in a multidimensional setting? And what are the implications of different reformulations? How complete is the ordering of social states that one can derive by introducing more and more requirements on the individual advantage functions without going so far as to prescribe a specific functional form? And, finally, what are the ethical features of some explicitly specified multidimensional inequality measures? How should we define the poor? Is someone poor when she is deprived on one dimension, or only when she is deprived on all dimensions? Or is it perhaps meaningful to count the number of dimensions on which she is deprived? What generic assumptions for the advantage function are implied by these different poverty definitions? It is in fact somewhat surprising that this literature on multidimensional inequality measurement has until now not had a larger influence on the empirical work within the capabilities approach *stricto sensu*.

I mentioned already that a large number of empirical applications content themselves with a mere description of functionings vectors. At the other extreme, there are also some examples in which one overall index value is constructed in an explicit way. Klasen (2000) calculates a deprivation index as the average score of all individual components. A similar method is followed in the well-known Human Development Index (HDI), which computes the well-being of a country as the simple (equally weighted) sum of the (transformed) scores on the three dimensions (log GDP per capita, education, life expectancy). Such an explicit weighting procedure has the advantage of being transparent and open for discussion. Of course, its weaknesses then become immediately clear. More specifically, the use of a simple sum implies perfect substitutability between the different dimensions, which

[20] Measurement of individual willingness to pay plays an important role in empirical applications of this approach. See Fleurbaey and Gaulier (2006) for a first application to international welfare comparisons.

[21] Weymark (2006) gives a survey of the normative approach to the measurement of multidimensional inequality; Trannoy (2006) summarizes multidimensional dominance approaches. An application to poverty with an interesting discussion of statistical aspects and some nice empirical illustrations can be found in Duclos *et al.* (2006).

strongly contradicts the proclaimed philosophy of the HDI, as stated, for example, in a recent Human Development Report: "Losses in human welfare linked to life expectancy, for example, cannot be compensated for by gains in other areas such as income or education" (UNDP 2005).

Other approaches in the literature have derived the weights on the basis of a statistical technique like principal components analysis (e.g. Klasen 2000),[22] have estimated output distance functions (Lovell *et al.* 1994; Ramos and Silber 2005), or have applied the Borda count (Dasgupta and Weale 1992; Qizilbash 1997). Recently, the fuzzy set methodology has attracted some interest (see e.g. Chiappero Martinetti 2000; Lelli 2008). Individuals who have a score below a lower threshold or above an upper threshold are classified unambigously as being deprived or nondeprived, respectively. For values between the two thresholds a membership function is specified to indicate the degree of "partial" deprivation. In some applications survey data are used to define these upper and lower cutoff points (Qizilbash and Clark 2005). In a next step, different operators (union, intersection, or averaging) are introduced to aggregate the different dimensions. While undoubtedly more attractive than the simple *ad hoc* approaches, the fuzzy set approach is less general than it may look at first. At the end, it boils down to applying specific hypotheses about (more or less attractive) functional forms for the membership functions and for the aggregation operators. The questions raised by this procedure are then very similar to the questions analyzed in the literature on multidimensional poverty measurement. An interesting procedure, which is not restricted to the fuzzy methodology, is the use of frequency-based weights to construct the overall index (see also Desai and Shah 1988). This captures the idea that the lower the proportion of people with a certain deprivation, the larger the weight assigned to that specific deprivation should be. It would be useful to get a better understanding of the theoretical underpinnings of this weighting scheme.

Once one has calculated an index of the living standard, one can use it to calculate "equivalent incomes", i.e. the income that persons with different characteristics need in order to reach a given level of living standard. These equivalent incomes can then be compared with poverty lines. They can also be compared with the equivalence scales as calculated with traditional economic methods. Recent papers which have pursued the idea of "functioning equivalence scales" (Zaidi and Burchardt 2005; Lelli 2005) have not solved the indexing problem, however, but have worked instead with equivalence scales computed for one individual functioning (having any savings in the former case, shelter in the latter).

An intriguing possibility relates to the use of "overall satisfaction" measures as aggregators. If much of the information used to estimate the functionings

[22] I mentioned already that the usefulness of these statistical techniques is doubtful, as is also acknowledged by Klasen (2000) himself. The "statistical" weights reflect only the correlation between the different dimensions; their relative importance in explaining the variation for the original items as such does not contain any useful normative information.

(or capabilities) is derived from questionnaire studies, then why not ask the respondents directly about their "valuation" of these capabilities and use the answers on this question as a measure of v_i? Some suggestions along this line are made by Anand and van Hees (2006). But there remains the problem of distinguishing clearly between "subjective happiness" (as one specific functioning) and "overall satisfaction" (as an aggregator). In the latter case, care must be taken to avoid the problems of "physical condition neglect" and "valuation neglect", if one does not want to fall back on simple welfarism. And these considerations confront us again head-on with the crucial questions raised earlier concerning the place to be given to individual valuations.

23.6 Macro versus Micro Studies: The Aggregation Problem

Given that the main focus of the capabilities approach is undoubtedly the individual, it is perhaps somewhat surprising that from the very beginning the most popular applications have been at the macro level, the best known being the Human Development Index. How to interpret these indices at the country level?

The most natural approach would be to construct the indices at, e.g., the national level as an aggregate of the living standards of individuals. If we have solved the problems of the previous section, and we have been able to define a measure of individual well-being, we can then write social welfare as $W(v_1(b_1), \ldots, v_n(b_n))$, where v_i is the valuation function defined in Eq. (2). Provided that the necessary measurability and comparability assumptions are satisfied, one could pick many possible specifications for $W(\cdot)$, going from the simple sum of capability index values to a leximin criterion, with different concave functions between these two extremes. Note again that in this setting it is not necessary (or desirable) to interpret v_i as a utility value: it should be seen as the value attached to one's life in a broader sense. This changes considerably the interpretation of the "comparability" of such values—and also suggests that the function $W(.)$ could be interpreted as the outcome of an (ideal?) political decision-making process.

This natural approach is *not* the one underlying the HDI and other similar country indices, however. These popular indices first aggregate over the different dimensions (e.g. by computing an average value for each country) and then aggregate these values for the different dimensions into one overall index. One possible interpretation of this approach is to look at the countries as if they were individuals and to apply the whole idea of "well-being" and "advantage" at the country level. But this is an unattractive approach because it completely neglects

the distribution of the different functionings within the countries. If we reject this interpretation of countries as representative individuals, we have to face the crucial question: does the "country" approach give reasonable approximations to the ethically preferable approach of first computing the indices of individual advantage and then aggregating these individual indices? Dutta, Pattanaik, and Xu (2003) have shown that this is not the case in general. The two approaches will yield the same results only if the aggregation functions have trivial and unattractive forms, boiling down basically to linearity.[23] This result is easy to understand, since the dimension-by-dimension approach completely discards all substitution and complementary relationships between the different dimensions at the level of the individuals.

The Dutta *et al.* (2003) result shows that the popular shortcut of working with country aggregates is apparently not very sensible if we are ultimately interested in the well-being (or deprivation) of individual persons. Since this seems indeed the dominant concern in the capabilities approach, the conclusion must be that we cannot avoid the task of collecting adequate data at the individual level.

23.7 Conclusion

The popularity of the capabilities approach has grown rapidly in recent years. The "capabilities movement" has even become very successful outside academia. This is good news for those who think that a sound policy analysis should look further than simple monetary measures of the living standard while at the same time not going the whole way towards subjective welfarism. A focus on individual human development with special emphasis on positive freedoms is indeed very attractive from an ethical point of view.

From an analytical perspective, the picture is perhaps more open. It is undoubtedly true that a lot of useful empirical research has now shown convincingly that a multidimensional approach offers rich insights for evaluating well-being and deprivation. Difficult methodological questions have remained unsolved, however. The recent theoretical literature has made it possible to formulate them in a sharp way. How should we evaluate opportunity sets? How can we introduce considerations of freedom into a "refined" functionings approach? How should we formulate a list of capabilities which can be used to analyze changes over time and differences between different societies without being open to manipulation? How can we construct an overall index of well-being, and what should be the relative role of a priori

[23] One can turn this negative conclusion on its head and argue that it offers some justification for the simple functional form of the HDI.

ethical valuations and of the opinions of the individuals themselves? Researchers developing capability indicators have recently opened up some interesting and novel lines of inquiry and I think that we will continue to progress these questions if there is in the future more extensive interaction between philosophers and social scientists, and between theory and empirical work. More specifically, it is crucial, *first*, to estimate structural models with individual data, analyzing the link between individual achievements, the socioeconomic and environmental background of the persons concerned, and the specific features of the individual processes of choice and decision-making; and, *second*, to integrate the insights from these models into applied ethical thinking.

References

Alkire, S. (2001). *Valuing Freedoms: Sen's Capability Approach and Poverty Reduction.* Oxford: Oxford University Press.

—— (2002). Dimensions of Human Development. *World Development*, 30/2, 181–205.

—— (2005). Subjective Quantitative Studies of Human Agency. *Social Indicators Research*, 74, 217–60.

Anand, P. and Santos, C. (2007). Violence, Gender Inequalities and Life Satisfaction. *Revue d'economie politique*, 117, 135–60.

—— and van Hees, M. (2006). Capabilities and Achievements: An Empirical Study. *Journal of Socio-Economics*, 35, 268–84.

—— Hunter, G., and Smith, R. (2009). Capabilities and Well-Being: Evidence Based on the Sen–Nussbaum Approach to Welfare. *Social Indicators Research*, 74, 9–55.

—— —— Carter, I., Dowding, K., Guala, F., and van Hees, M. (2005). The Development of Capability Indicators. *Journal of Human Development*, forthcoming.

—— Santos, C., and Smith, R. (2008). The Measurement of Capabilities. In *Essays in Honour of Amartya Sen* (2008). Oxford: Oxford University Press.

Balestrino, A. (1996). A Note on Functioning-Poverty in Affluent Societies. *Notizie di Politeia*, 12/43–4, 97–105.

Barbera, S., Bossert, W., and Pattanaik, P. (1998). Ranking Sets of Objects. In S. Barberà, P. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory*, ch. 17. Dordrecht and New York: Kluwer Academic Press.

Basu, K. (1987). Achievements, Capabilities and the Concept of Well-Being. *Social Choice and Welfare*, 4, 69–76.

—— and Lopez-Calva, L. (forthcoming). Functionings and Capabilities. In K. Arrow, A. Sen, and K. Suzumura (eds.), *Handbook of Social Choice and Welfare.* Amsterdam: Elsevier.

Brandolini, A., and D'Alessio, G. (1998). *Measuring Well-Being in the Functioning Space.* Mimeo, Research Department Banca d'Italia.

Brun, B., and Tungodden, B. (2004). Non-Welfaristic Theories of Justice: Is the "Intersection Approach" a Solution to the Indexing Impasse? *Social Choice and Welfare*, 22/1, 49–60.

Chiappero Martinetti, E. (2000). A Multidimensional Assessment of Well-Being Based on Sen's Functioning Approach. *Rivista Internazionale di Scienze Sociali*, 58, 207–39.

Clark, D. (2005). Sen's Capability Approach and the Many Spaces of Human Well-Being. *Journal of Development Studies*, 41/8, 1339–68.

Cummins, R. (1996). Domains of Life Satisfaction: An Attempt to Order Chaos. *Social Indicators Research*, 38/3, 303–28.

Dasgupta, P., and Weale, M. (1992). On Measuring the Quality of Life. *World Development*, 20, 119–31.

Desai, M., and Shah, A. (1988). An Econometric Approach to the Measurement of Poverty. *Oxford Economic Papers*, 40, 505–22.

Duclos, J.-Y., Sahn, D., and Younger, S. (2006). Robust Multidimensional Poverty Comparisons. *Economic Journal*, 116, 943–68.

Dutta, I., Pattanaik, P., and Xu, Y. (2003). On Measuring Deprivation and the Standard of Living in a Multidimensional Framework on the Basis of Aggregate Data. *Economica*, 70, 197–221.

Fleurbaey, M. (1995). Equal Opportunity or Equal Social Outcome? *Economics and Philosophy*, 11, 25–55.

—— (2005). *Equality of Functionings.* Mimeo.

—— (2006). Capabilities, Functionings and Refined Functionings. *Journal of Human Development*, 7/3, 299–310.

Fleurbaey, M. (2007). Social Choice and the Indexing Dilemma. *Social Choice and Welfare*, 29, 633–48.

—— (2008), *Fairness, Responsibility and Welfare*. Oxford: Oxford University Press.

—— and Gaulier, G. (2006). *International Comparisons of Living Standards by Equivalent Incomes.* Mimeo, CERSES, Paris.

Gaertner, W., and Xu, Y. (2006). Capability Sets as the Basis of a New Measure of Human Development. *Journal of Human Development*, 7/3, 311–22.

—— Pattanaik, P., and Suzumura, K. (1992). Individual Rights Revisited. *Economica*, 59, 161–77.

Gorman, W. (1956). *The Demand for Related Goods.* Iowa Experimental Station, Paper J3129.

Herrero, C. (1996). Capabilities and Utilities. *Review of Economic Design*, 2/1, 69–88.

Klasen, S. (2000). Measuring Poverty and Deprivation in South Africa. *Review of Income and Wealth*, 46/1, 33–58.

Krishnakumar, J. (2007). Going Beyond Functionings to Capabilities: An Econometric Model to Explain and Estimate Capabilities. *Journal of Human Development*, 8/1, 39–63.

Kuklys, W. (2005). *Amartya Sen's Capability Approach: Theoretical Insights and Empirical Applications.* Berlin: Springer Verlag.

Lancaster, K. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74/2, 132–57.

Lelli, S. (2005). Using Functionings to Estimate Equivalence Scales. *Review of Income and Wealth* 51/2, 255–84.

—— (2008). Operationalising Sen's Capability Approach: The Influence of the Selected Technique. In S. Alkire, F. Comim, and M. Qizilbash (eds.), *Capability and Justice: Concepts, Measures and Applications*, 310–51. Cambridge: Cambridge University Press.

Lerner, M. (1980). *The Belief in a Just World.* New York: Plenum Press.

Lovell, K., Richardson, S., Travers, P., and Wood, L. (1994). Resources and Functionings—A New View of Inequality in Australia. In W. Eichhorn (ed.), *Models and Measurement of Welfare and Inequality*, 787–807. Berlin: Springer Verlag.

Nussbaum, M. (2000). *Women and Human Development: The Capabilities Approach.* Cambridge: Cambridge University Press.

—— (2006). *Frontiers of Justice: Disability, Nationality, Species Membership.* Cambridge, MA: Harvard University Press.

—— and Sen, A. (1993). *The Quality of Life.* Oxford: Clarendon Press.

Pattanaik, P., and Xu, Y. (1990). On Ranking Opportunity Sets in Terms of Freedom of Choice. *Recherches Economiques de Louvain*, 56/3–4, 383–90.

—— —— (2007). Minimal Relativism, Dominance, and Standard of Living Comparisons Based on Functionings. *Oxford Economic Papers*, 59, 354–74.

Phipps, S. (2002). The Well-Being of Young Canadian Children in International Perspective: A Functionings Approach. *Review of Income and Wealth*, 48/4, 493–515.

Qizilbash, M. (1997). Pluralism and Well-Being Indices. *World Development*, 25/12, 2009–26.

—— (2002). A Note on the Measurement of Poverty and Vulnerability in the South African Context. *Journal of International Development*, 14, 757–72.

—— and Clark, D. (2005). The Capability Approach and Fuzzy Poverty Measures: An Application to the South African Context. *Social Indicators Research*, 74, 103–39.

Ramos, X., and Silber, J. (2005). On the Application of Efficiency Analysis to the Study of the Dimensions of Human Development. *Review of Income and Wealth*, 51/2, 285–309.

Robeyns, I. (2005). Selecting Capabilities for Quality of Life Measurement. *Social Indicators Research*, 74, 191–215.

—— (2006*a*). The Capability Theory in Practice. *Journal of Political Philosophy*, 14/3, 351–76.

—— (2006*b*). Measuring Gender Inequality in Functionings and Capabilities: Findings from the British Household Panel Survey. In P. Bharati and M. Pal (eds.), *Gender Disparity: Manifestations, Causes, and Implications*, 236–77. New Delhi: Anmol Publishers.

Roemer, J. (1998). *Equality of Opportunity.* Cambridge, MA: Harvard University Press.

Ruggeri Laderchi, C. (1997). Poverty and its Many Dimensions: The Role of Income as an Indicator. *Oxford Development Studies*, 25, 345–60.

Schokkaert, E., and Van Ootegem, L. (1990). Sen's Concept of the Living Standard Applied to the Belgian Unemployed. *Recherches Economiques de Louvain*, 56/3–4, 429–50.

Sen, A. (1970). The Impossibility of a Paretian Liberal. *Journal of Political Economy*, 78, 152–7.

—— (1980). Equality of what? In A. Sen (ed.), *Choice, Welfare and Measurement*, 353–69. Oxford: Blackwell.

—— (1983). Poor, Relatively Speaking. *Oxford Economic Papers*, 35, 153–69.

—— (1985*a*). *Commodities and Capabilities.* Amsterdam: North-Holland.

—— (1985*b*). A Sociological Approach to the Measurement of Poverty: A Reply to Professor Peter Townsend. *Oxford Economic Papers*, 37, 669–76.

—— (1990). Welfare, Freedom and Social Choice: A Reply. *Recherches Economiques de Louvain*, 56/3–4, 451–85.

—— (1992). *Inequality Re-examined.* Oxford: Clarendon Press.

—— (1999). *Development as Freedom.* New York: Knopf.

—— (2002). Response to Commentaries. *Studies in Comparative International Development*, 37/2, 78–86.

Sen, A. (2004). Capabilities, Lists, and Public Reason: Continuing the Conversation. *Feminist Economics*, 10/3, 77–80.

—— Muellbauer, J., Kanbur, R., Hart, K., and Williams, B. (1987). *The Standard of Living.* Cambridge: Cambridge University Press.

Sugden, R. (1993). Welfare, Resources and Capabilities: A Review of "Inequality Reexamined" by Amartya Sen. *Journal of Economic Literature*, 31, 1947–62.

Townsend, P. (1985). A Sociological Approach to the Measurement of Poverty: A Rejoinder to Professor Amartya Sen. *Oxford Economic Papers*, 37, 659–68.

Trannoy, A. (2006). Multidimensional Egalitarianism and the Dominance Approach: A Lost Paradise? In F. Farina and E. Savaglio (eds.), *Inequality and Economic Integration*, 284–302. London: Routledge.

UNDP (2005). *International Cooperation at a Crossroads: Aid, Trade and Security in an Unequal World.* Human Development Report 2005. New York: UNDP.

Weymark, J. (2006). The Normative Approach to the Measurement of Multidimensional Inequality. In F. Farina and E. Savaglio (eds.), *Inequality and Economic Integration*, 303–28. London: Routledge.

Zaidi, A., and Burchardt, T. (2005). Comparing Incomes when Needs Differ: Equivalization for the Extra Costs of Disability in the UK. *Review of Income and Wealth*, 51/1, 89–114.

Index