JUSTICE, POLITICAL LIBERALISM, and UTILITARIANISM

Themes from Harsanyi and Rawls

Edited by

MARC FLEURBAEY

MAURICE SALLES

JOHN A. WEYMARK

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521640930

This page intentionally left blank

JUSTICE, POLITICAL LIBERALISM, AND UTILITARIANISM

The utilitarian economist and Nobel Laureate John Harsanyi and the liberal egalitarian philosopher John Rawls were two of the most eminent scholars writing on problems of social justice in the last century. This volume pays tribute to Harsanyi and Rawls by investigating themes that figure prominently in their work. In some cases, the contributors explore issues considered by Harsanyi and Rawls in more depth and from novel perspectives. In others, the contributors use the work of Harsanyi and Rawls as points of departure for pursuing the construction of new theories for the evaluation of social justice. A lengthy introductory essay by the editors provides background information on the relevant economics, game theory, philosophy, and social choice theory, as well as readers' guides to the individual contributions, so as to make this volume accessible to scholars in a wide range of disciplines.

Marc Fleurbaey is Research Director at the Centre de Recherche Sens, Éthique, Société (CNRS and Université Paris Descartes), France, and, for 2006–08, Lachmann Fellow at the London School of Economics. He is a Managing Editor of the journal *Social Choice and Welfare*, and he coedited the Cambridge University Press journal *Economics and Philosophy* from 2000 to 2004.

Maurice Salles is Professor of Economics at the University of Caen, France, where he has taught since 1982. Coordinating Editor of the journal *Social Choice and Welfare* since 1984, he coedited the volume *Social Choice, Welfare, and Ethics* (Cambridge University Press, 1995) and has served as an Honorary Research Associate at the London School of Economics since 2002.

John A. Weymark is Professor of Economics at Vanderbilt University, Nashville, Tennessee. He is President of the Society for Social Choice and Welfare for 2008–09 and is a Managing Editor of the journal *Social Choice and Welfare.*

Justice, Political Liberalism, and Utilitarianism

Themes from Harsanyi and Rawls

Edited by

MARC FLEURBAEY

MAURICE SALLES

JOHN A. WEYMARK

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521640930

First published in print format 2007

ISBN-13	978-0-511-39461-4	eBook (NetLibrary)
ISBN-13	978-0-521-64093-0	hardback

Contents

PART ONE THEMES FROM RAWLS

PART TWO HARSANYI'S IMPARTIAL OBSERVER AND SOCIAL AGGREGATION THEOREMS

PART THREE GOODNESS AND WELL-BEING

PART FOUR SHARING THE GAINS FROM SOCIAL COOPERATION

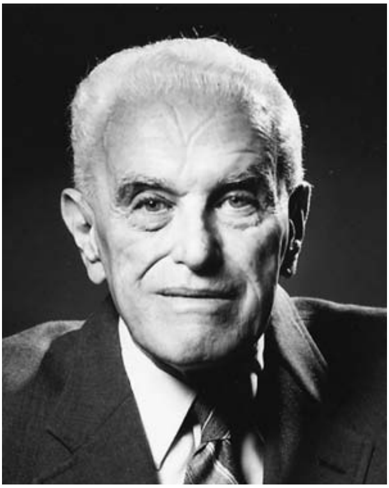
PART FIVE RIGHTS AND LIBERTIES

Preface

This volume has its origins in a conference on “Justice, Political Liberalism, and Utilitarianism in Honour of John Harsanyi and John Rawls” that was held at the Université de Caen in June 1996. This conference received generous financial support from the Association pour le Développement de la Recherche en Économie et en Statistique, the Centre National de la Recherche Scientifique, the Centre de Recherche en Économie et Management (UMR-CNRS 6211), the Université de Caen, and the Ville de Caen.

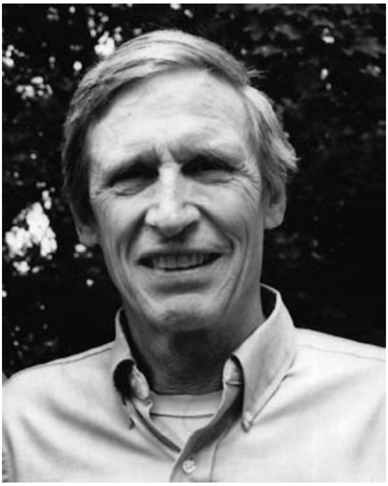
We are grateful to the scholars who served as referees of the individual chapters, to Pascal Engel for his help in the organization of the Caen conference, and to Cathy Zebron who assisted with the preparation of the final manuscript. At Cambridge University Press, thanks are due to Patrick McCartan, who responded so favorably when this volume was first proposed to him, and to Scott Parris, who has been our editor from the outset. We are particularly grateful to our contributors and to Scott for their patience and encouragement during the inordinately long time it has taken to bring this project to fruition.

Marc Fleurbaey
Maurice Salles
John A. Weymark



John Harsanyi

Photograph courtesy of Anne Harsanyi



John Rawls

Photograph courtesy of Mardy Rawls

List of Contributors

Richard Arneson Department of Philosophy, University of California, San Diego

Brian Barry Department of Political Science, Columbia University, and Department of Government, London School of Economics

Ken Binmore Department of Economics, University College London

Charles Blackorby Department of Economics, University of Warwick, and Groupement de Recherche en Économie Quantitative d'Aix-Marseille

John Broome Corpus Christi College, University of Oxford

Claude d'Aspremont Center for Operations Research and Econometrics and Département des sciences économiques, Université Catholique de Louvain

David Donaldson Department of Economics, University of British Columbia

Marc Fleurbaey Centre National de la Recherche Scientifique; Centre de Recherche Sens, Éthique, Société, Université Paris 5; London School of Economics; and Institut d'Économie Publique

James Griffin Corpus Christi College, University of Oxford

John C. Harsanyi* Haas School of Business, University of California, Berkeley

Matthias Hild Darden School of Business Administration, University of Virgina

Richard Jeffrey* Department of Philosophy, Princeton University

François Maniquet Center for Operations Research and Econometrics, Université Catholique de Louvain

Edward F. McClennen Department of Philosophy, Syracuse University

Philippe Mongin Centre National de la Recherche Scientifique and École des Hautes Études Commerciales

Philip Pettit Department of Politics, Princeton University

Mathias Risse Kennedy School of Government, Harvard University

Jonathan Riley Department of Philosophy and Murphy Institute of Political Economy, Tulane University

John E. Roemer Departments of Political Science and Economics, Yale University

Maurice Salles Centre de Recherche en Économie et Management, Université de Caen

Brian Skyrms Department of Logic and Philosophy of Science, University of California, Irvine

Robert Sugden School of Economics, University of East Anglia

John A. Weymark Department of Economics, Vanderbilt University

*Deceased

1

An Introduction to *Justice, Political Liberalism, and Utilitarianism*

Marc Fleurbaey, Maurice Salles, and John A. Weymark

The opposition between utilitarianism and liberal egalitarianism has triggered the most important developments in political philosophy in the twentieth century and has had a considerable effect on other subjects as well, such as law and economics. The turn of the new century has witnessed the death of two prominent scholars in these debates, John Harsanyi and John Rawls. Harsanyi and Rawls have undoubtedly been the leading figures in each of these schools of thought in recent decades. Building on the work of classical utilitarians such as Jeremy Bentham and John Stuart Mill, Harsanyi has provided decision-theoretic foundations for utilitarianism that have served as the touchstone for Rawls's own critique of utilitarianism. Rawls believes that utilitarianism fails to satisfy Immanuel Kant's maxim that individuals should be treated as ends in and of themselves, not just as means for promoting the social good. Drawing inspiration from the writings of social contract theorists such as John Locke and Jean-Jacques Rousseau, Rawls has fashioned a modern statement of liberal egalitarian principles for the design of the basic institutions of society that respect Kant's maxim.

The writings of Harsanyi and Rawls offer vigorous defenses of their theories, which their lively exchanges have done much to illuminate. Their theories draw on and provide support for widely shared values. Their contributions have been, and will continue to be, inspirational for scholars and others who seek to understand what social justice and ethical behavior require. The voluminous literature that has responded to Harsanyi's and Rawls's writings has drawn out many of the implications of their theories, has clarified and refined their most convincing arguments, and has pointed out ambiguities and weaknesses in their reasoning. Whether the divide between

We are grateful to Peter Vallentyne for his comments on an earlier version of this chapter.

these two schools of thought will eventually yield a convergence toward a consensual theory or whether it will be agreed that these theories are based on irreducibly opposed principles is premature to say. But the opposition itself is most useful in order to help us appreciate the difficulties of formulating a coherent account of social justice and social welfare. Furthermore, the ongoing debate between utilitarians and liberal egalitarians may simply reflect a deeper truth that it is utopian to believe that consensus can ever be reached on such fundamental social issues. As Alan Dershowitz has so elequently said, "There are few, if any, moral truths (beyond meaningless platitudes) that have been accepted in all times and places. The active and never-ending processes of moralizing, truth searching, and justice seeking are far superior to the passive acceptance of one truth. . . . Conflicting moralities serve as checks against the tyranny of singular truths" (Dershowitz, 2006, p. 94).

The general principles of utilitarianism and liberal egalitarianism are not the only source of inspiration for recent reflections that have drawn on the work of Harsanyi and Rawls. The particulars of their theories have been challenging and illuminating in different ways. For example, Harsanyi's theorems about utilitarianism have raised new questions about the relevance of individual preferences under uncertainty, and their representation with particular utility functions, for the evaluation of social states. Similarly, Rawls's contractarian defense of his version of liberal egalitarianism and his frequent reference to the mutual gains of social cooperation has engaged theorists attracted by contractarian approaches to further explore what constitutes social justice, even though they have often ended up far from liberal egalitarian conclusions.

Harsanyi and Rawls deserve our most respectful tribute for their fundamental contributions to utilitarianism and liberal egalitarianism, respectively, and, more generally, for helping to bring questions of social justice to the fore after many years of relative neglect. The chapters in this volume honor Harsanyi and Rawls by investigating themes that figure prominently in their writings. In some cases, the chapters that follow take stock of what has been learned by exploring issues considered by Harsanyi and Rawls in more depth and from novel perspectives. However, much as it is valuable to understand and compare their theories, new approaches to dealing with problems of social justice have commanded attention in recent years. Many of the contributors to this volume use the work of Harsanyi and Rawls as points of departure for pursuing the construction of new theories for the evaluation of social justice and injustice.

In this chapter, we introduce the chapters in this collection and place them in the context of the literature. To address the questions considered in subsequent chapters, some authors have employed mathematical arguments, which may be a formidable barrier for some readers with limited mathematical training. Nevertheless, the lessons learned from these "mathematical" articles are important both for evaluating what Harsanyi and Rawls have accomplished and for understanding some of the new approaches that their writings have inspired. Accordingly, we have attempted to provide relatively nontechnical (but not completely nonmathematical) summaries of them here, even though this results in the mathematical chapters receiving more attention than some of the more widely accessible chapters.

1.1 Themes from Rawls

Rawls's theory of social justice is multifaceted, with different elements of the theory cohering in a complex way. His ideas have been challenged in many ways, even by those who are in broad agreement with his approach to questions of justice. The chapters in Part 1 of this volume consider three issues related to the work of Rawls: (i) his use of absolute priority rules, (ii) the role that merit and personal responsibility play in his theory of justice, and (iii) the role that moral intuitions should play in justifying ethical beliefs.[1] Echoing the vast literature devoted to analyzing Rawls's writings, these contributions range from critiques of some fundamental features of his theory to proposals for amending his ideas so as to overcome some of its shortcomings.

The main features of Rawls's theory of justice, first set out in detail in *A Theory of Justice* (Rawls, 1971) and further developed in *Political Liberalism* (Rawls, 1993) and *Justice as Fairness: A Restatement* (Rawls, 2001), can be briefly summarized. Rawls's theory of justice focuses on the basic structure of social institutions, requiring them to be organized in such a way as to favor the most destitute members of society, that is, those individuals who have the fewest basic social goods (what Rawls calls primary goods) such as rights and liberties, power and opportunities, self-respect, and income and wealth. Moreover, Rawls considers equal liberties and a fair equality of opportunities (in particular, for access to careers and related advantages) to be of paramount importance, so that the right to equal liberties and equal opportunities for all should have priority over the advancement of the socio-economic status of the poor, with priority in turn given to liberties over equality of opportunity. These principles were originally developed for a society that shares similar moral values, but in his later

[1] Additional Rawlsian themes are considered elsewhere in this volume.

writings, Rawls has defended them as a political conception of justice (hence the name, "political liberalism") for a pluralistic society whose members can endorse his principles even though they differ fundamentally in their religious, philosophical, and moral beliefs. In this revised account of his theory of justice, Rawls places greater emphasis on distinguishing political liberties from other kinds of liberties and requires that they be of comparable worth in the sense that everyone has an equal opportunity to hold political office and to influence the decisions made in the political sphere.[2]

Rawls's focus on the basic structure of society and on primary goods is related to the liberal features of his theory. Specifically, Rawls regards society as having the duty to provide everyone with a fair share of resources and opportunities. However, society does not have the right to interfere with private uses of these resources that result from personal conceptions of what constitutes a good life, which each member of society is free to develop and revise as he wishes. A liberal society should not dictate to its members how life must be lived, and from this, Rawls derives the far-reaching implication that social institutions should not in any way refer to particular conceptions of the good life and, therefore, should not try to compare individual success by any metric of the good that would involve such a conception.

1.1.1 Harsanyi on Rawls

In Chapter 2, the late John Harsanyi, pursues and, sadly, closes a debate with Rawls that began in Harsanyi (1975) and Rawls (1974). The initial debate focused on the appropriate specification of an impartial perspective, what Rawls calls an original position, from which an impartial observer (to use Harsanyi's terminology) identifies the basic features of their theories.[3] The use of this device plays a substantial justificatory role in the work of both Harsanyi and Rawls, and it raises fundamental issues about what constitutes rational criteria for impartial decision making. More will be said about original position arguments in the next section.

Here, Harsanyi focuses on two other issues. Echoing the earlier exchange, he first objects to the absolute priority accorded to the worst-off social group, or maximin principle, that is embraced by Rawls. With this principle, one social arrangement is judged to be better than a second if the situation of the

[2] For precise statements of Rawls's principles, see Rawls (1971, section 46) and Rawls (1993, pp. 5–6).

[3] Strictly speaking, Rawls views his original position as a forum for multiperson bargaining, but, as we shall see in Section 1.2.2, it can also be viewed in terms of a single impartial observer choosing principles of justice.

worst off in the first alternative is better than the situation of the worst off in the second. For Rawls, an individual's situation is assessed by the expected value of an index of primary goods holdings over his lifetime.[4] More so than in his earlier critique of Rawls, Harsanyi questions the policy implications at the social level of the maximin criterion. According to Harsanyi, such an extreme criterion can only have extreme social consequences in terms of redistributive policies, possibly triggering a civil war. Although Harsanyi does not mention it, the maximin criterion has often been employed in the analysis of optimal redistributive taxation (see, e.g., Atkinson, 1973, 1995; Choné and Laroque, 2005) without inducing extreme consequences because incentive constraints preserve the interests of the wealthy members of society better than an army. This example, however, does not rule out the possibility of extreme consequences with other policies.

Harsanyi also critically examines other absolute priorities that are granted, in Rawls's approach, to certain primary goods over others, in particular liberties over socioeconomic advantage. Harsanyi argues that, in general, individuals acting collectively employ finite trade-offs, whether between social groups or between goods, although he does allow for the absolute preeminence of certain values over others (e.g., moral duties over nonmoral interests). How reasonable it is to give absolute priority to the worse off over the better off, or to liberties over socioeconomic status, is likely to remain a contested topic for some time. Following the publication of Rawls's *A Theory of Justice*, the maximin criterion attracted considerable interest and, in some cases, support from welfare economists. Interest in giving priority to the worst off has ebbed and flowed over the years, but it appears to have enjoyed a recent resurgence (see, e.g., Maniquet and Sprumont, 2004; Tungodden, 2000). Nevertheless, Harsanyi's skepticism about absolute priorities and his preference for middle-ground criteria are very natural and widely shared.

[4] Rawls refers to this principle as the "difference principle," but the term "maximin" is both more standard and a more transparent label for a principle that seeks to maximize the smallest value of some attribute. For example, the maximin principle applied to utility first identifies the utility of the person with the smallest utility in each of a set of distributions of individual utilities and then chooses a distribution for which this utility value is largest. With the lexicographic version of Rawls's maximin principle, leximin for short, the ranking of two alternatives is determined by the value of the index of primary goods of the worst-off group for which the value of this index is not the same in the two alternatives. With both the maximin and leximin principles, groups are defined by their ranks (in terms of primary goods holdings), not by the names of the individuals that comprise them. Although it is in this lexicographic form that the maximin principle is typically employed, for simplicity, we ignore this refinement in the subsequent discussion.

The second set of considerations raised by Harsanyi has to do with Rawls's complex treatment of the notions of merit and personal responsibility in his theory of justice. As previously discussed, Rawls advocates a division of labor between social institutions and individuals. The former provide resources and opportunities, the latter are responsible for how these resources and opportunities are used to pursue their own conceptions of what makes a life valuable. It is not the job of social institutions to track individual merit and responsibility and to reward them accordingly because institutions only have to take care of the distribution of fair shares of resources. One then obtains the somewhat paradoxical situation in Rawls's theory of a concept of responsibility that plays a key role in delineating the limited role of social distribution, while being totally absent from the principles of distribution themselves. Only incentive considerations can justify differential rewards in this view. Rawls (1971, section 48) famously argues that even effort cannot be the moral basis for superior claims over resources because the propensity to work hard is largely inherited and is nurtured in a favorable family environment. Both of these contingencies are morally arbitrary.

In a discussion that appears to be based on commonsense morality as much as on utilitarian maximization of the social good, Harsanyi opposes this view. He argues that it is essential that society publicly recognizes the intrinsic and social value of (i) moral behavior that results from a good character and (ii) the development and employment of talents for the common good even if these talents and characters are partly inherited or nurtured by a caring family. Otherwise, human excellence cannot flourish. Harsanyi abhors the vision of a society in which all kinds of moral characters would be considered equally nonpraiseworthy. He understands Rawls's view as connected to hard determinism, a doctrine that denies the existence of free will and responsibility on the assumption that all causal laws are deterministic. Harsanyi defends a compatibilist approach to the problem of free will, that is, a view that accepts physical determinism but nonetheless carves a place for personal responsibility and moral praise and blame. In this view, while an individual's moral attitudes are heredity and environment dependent, nevertheless, they are subject to choice and, therefore, his actions are subject to moral commendation or discredit.

1.1.2 Liberal Egalitarian Approaches to Personal Responsibility

The issue of individual responsibility has been the subject of much attention in the philosophical literature since the publication of *A Theory of Justice*, most notably by Arneson (1989), Cohen (1989), and Dworkin (1981, 2000).

Each of these scholars has proposed variants of liberal egalitarianism that put personal reponsibility at the core of the definition of individual advantage.

Dworkin has not diverged much from Rawls's view that individuals should be held responsible for their conceptions of the good life and their ambitions, but, unlike Rawls, he believes that individuals are responsible for certain kinds of unlucky outcomes. Specifically, Dworkin distinguishes between option luck, which is concerned with the outcomes of deliberate risky choices, and brute luck, which is not. He holds individuals responsible for a bad outcome in the former (given fair initial conditions), but not the latter, case because the adverse consequences of a deliberate risky choice may be mitigated by purchasing insurance, and, if this is not possible, individuals could have refrained from making a risky decision in the first place.

Arneson and Cohen fully embrace the commonsense view that individuals can be held responsible only for what lies within their genuine control, and define social justice in terms of equal opportunities in a radical sense – equality of what is under the control of individuals to achieve.[5] They differ on how the metric of achievement should be defined for comparisons across individuals, but these differences need not concern us here.[6]

These developments have had a substantial impact on welfare economics, where studies of freedom and opportunities have flourished in recent years. An important strand of this literature is concerned with the fair distribution of resources and opportunities when account is taken of the responsibility individuals have for making choices. Fleurbaey (1998), for example, distinguishes between the objective of neutralizing the effects of factors outside of an individual's control and the various possible objectives that may be adopted to reward an individual's exercise of responsibility. See also Roemer (1998) and the contributions in Laslier, Fleurbaey, Gravel, and Trannoy (1998) for further explorations of these and related issues.

1.1.3 Arneson on Personal Responsibility

In Chapter 3, Richard Arneson undertakes a detailed analysis of the role that personal responsibility plays in Rawls's theory of justice in light of

[5] Rawls's notion of equal opportunities is closer to the ordinary sense of the term, namely, nondiscrimination in the access to positions of authority and responsibility.

[6] The way that Arneson and Cohen view achievement has a close affinity to Sen's theory of capabilities (see Sen, 1985, 1992). While Sen's theory is primarily defended by him in terms of the freedom to choose between alternative options (e.g., lives), factors that individuals cannot be held accountable for also play a fundamental role in assessing their circumstances. See, for example, Sen (1992, section 5.3).

these more recent developments. This analysis illuminates the evolution of Arneson's thought from his theory of equal opportunities (Arneson, 1989) to his desert-sensitive theory of justice (Arneson, 2000).

Arneson observes that Rawls walks on a tightrope because he denies that social institutions should be devoted to rewarding the deserving, while retaining a key role in his theory for individual agency and responsibility. Unlike Harsanyi, Arneson does not interpret Rawls as endorsing hard determinism, nor does he believe that it is necessary to determine the extent to which a person has a free will or is responsible for the outcomes of his actions in real-world situations before any conclusions can be reached on principles of justice. Arneson takes as his starting point "the limiting principle that we should be held responsible at most for what lies within our power to control" (p. 98) and argues that independently of all conceptual and practical difficulties in the definition and measurement of control, it is worth pursuing the ethical analysis of ideal principles in order to derive their consequences before any consideration of practical implementation is raised. In this view, the free will problem is concerned with implementation, not principles, because society often lacks the information needed to assess personal responsibility.

Arneson identifies two main shortcomings with Rawls's account of the role that merit and deservingness play in the design of institutions that shape the distribution of resources. First, he argues that Rawls's use of the maximin principle, with its emphasis on the maximization of the level of primary goods that the worst-off group in society can be expected to enjoy over their lifetimes, fails to distinguish correctly between inequalities that are a matter of choice and those that are not. For example, future wealth (one of the primary goods) depends on one's choice of employment. Second, he argues that Rawls's dismissal of the view that benefits and burdens should be distributed in proportion to moral worth because moral worth cannot be defined independently of the content of the norms of justice is based on a false premise. For Arneson, desert does have an independent specification.

Arneson is not content to show the failings of Rawls's theory compared with a more refined desert-based principle of responsibility. He also suggests a possible amendment to Rawls's theory in which the maximin principle is expressed in terms of the expected potential lifetime holdings of primary goods, not the expected value of their actual holdings, at the onset of adulthood. If individuals do not achieve their potentials, that is their responsibility. As Arneson notes, this proposal has much in common with Dworkin's views discussed earlier. However, he ultimately concludes that this move is not successful because it fails to consider that the ability to make good choices and stick to them is an unchosen characteristic.

After Arneson considers and rebuts some variants of the thesis that the choice of ends individuals make and their consequences are not matters of justice, provided some threshold of rationality is attained, he critically examines the responsibility principle to determine how one could obtain a reasonably acceptable notion of justice that provides a role for both individual and social responsibility in the choice of ends. Arneson, like Rawls, believes that individuals should be responsible for their freely chosen ends, but he also believes that society should undertake any measures that can help improve the quality of people's responsible choices. He concludes that justice requires allocating resources at the onset of adulthood to maximize the effective opportunities for well-being of the most disadvantaged, where effective well-being is measured in terms of the well-being that could be obtained if an individual acts as prudently as one could reasonably expect. Justice does not demand compensation for bad consequences of a rational choice, but it may be required if the individual making a choice is not completely responsible, which would be the case if society failed to provide an adequate environment to nurture the ability to make prudent choices.

Arneson also examines whether there is a case for assigning liabilities for an adverse outcome in a way that diverges from the costs that individuals are responsible for and that may depend on factors outside their control. In other words, are there circumstances in which individuals should be asked to share in the costs of events or decisions for which they are not strictly responsible? Such sharing is a common response to a natural catastrophe and can be justified by the obligation we all have to compensate the affected individuals for events outside their control. It can also be envisaged for the costs of responsible decisions made by individuals if pooling the costs promotes general fairness and efficiency better than a fine-grained sorting out of personal responsibilities.

In going beyond the pure opportunity-based approach to resource allocation by, in some circumstances, considering the consequences of poor decision making when determining whether compensation is merited, Arneson seeks to avoid the common criticism that the pure opportunity approach to compensation can be too unforgiving to individuals who suffer misfortunes apparently as a result of their own choices. This line of reasoning opens the way for the desert-catering prioritarian theory that he develops in Arneson (2000). In this theory, a failure to seize opportunities reduces the moral value of providing a compensatory benefit to the concerned individual, but does not necessarily nullify it.

1.1.4 Griffin on Moral Intuition

Turning now from issues of priority and responsibility, James Griffin, in Chapter 4, raises a third set of issues dealing with the role that moral intuitions play in justifying a set of ethical beliefs. Rawls's arguments in defense of his principles of justice often refer to intuition about what seems reasonable. He refrains from using the common strategy employed in political philosophy of testing general principles by artfully conceived, but sometimes contrived, examples that appeal to our intuitions. Instead, he proposes the concept of a reflective equilibrium (Rawls, 1971, sections 4 and 9) as the archetype of the support that a normative theory should seek to obtain from its double confrontation with reasoning and intuition. A reflective equilibrium in favor of normative principles (of justice or of morality) occurs when abstract analysis (for instance, a description of an impartial observer's reasoning) yields conclusions that fully agree with one's well-considered judgments, that is, with the normative beliefs that one would hold once one's initial beliefs have been revised after having considered alternative normative principles. In Rawls's case, his objective is to justify his three principles of justice and the priority accorded by them of first securing equal liberties for all, secondly providing fair equality of opportunity, and, finally, maximizing the expected holdings of primary goods of the least advantaged. Rawls's hope was, of course, that his theory of justice, with its combination of reasoning about what principles of justice would be agreed to in his original position and its appeal to the intuitive reasonableness of these principles and their priority ranking, would produce such an equilibrium.

Griffin is concerned with the justification of ethical beliefs in general, not simply the justification of normative principles of justice. He is refreshingly critical about the piecemeal approach to ethical reasoning based on intuitive consideration of hypothetical examples, arguing that this procedure gives too much weight to intuition. He also dismisses the other extreme of deriving substantive moral principles without any appeal to intuition as being unsuccessful.

Griffin considers the lessons to be learned for moral reasoning from the coherence theory of justification found in the natural sciences in some detail. As in an ethical theory, a scientific theory distinguishes good from bad beliefs and tries to make them cohere. In the natural sciences, it is empirical observations, the inferences that can be made from them, and how well these observations and inferences describe how the world works that separate credible beliefs from ones that are not.

The analogue of these highly credible perceptual beliefs in ethics are the core values that are part of what makes us human, but, in Griffin's view, neither these nor a somewhat enlarged set of beliefs are extensive enough to imply much about the substantive content of ethical principles. Griffin also considers whether the kind of considered judgments obtained using Rawls's concept of a reflective equilibrium might provide the basis for a coherent, or at least a partially coherent, set of ethical beliefs. He does not have much hope that judgments obtained in this way will place strong constraints on substantive moral principles. He is also skeptical that a reflective equilibrium in terms of considered ethical beliefs would necessarily weed out morally objectional views.

In ethics, unlike in the natural sciences, explanatory power is not a concern and the systematization of beliefs need not be essential, or so Griffin argues. Rather, one needs to develop prudential and moral standards for deciding how to live. These standards may arise in an unsystematic way and prove to be satisfactory in practice.

As Griffin notes, his analysis raises difficult metaethical concerns. For example, metaethical issues are raised when developing criteria for determining whether a revision of beliefs is an improvement and when asking whether ethics can reasonably aspire to being a system of coherent beliefs. Although Griffin does not propose solutions to these metaethical questions here, his observations nevertheless suggest that Rawls's position that one can practice ethics without worrying too much about metaethical foundations deserves closer scrutiny. As Griffin aptly argues, progress in ethics and in metaethics will have to be simultaneous because we should seek a better ethical theory that provides not only better principles, but also sounder foundations for such principles.

1.2 Harsanyi's Impartial Observer and Social Aggregation Theorems

Harsanyi's decision-theoretic defenses of utilitarianism have been the subject of much debate. The contributions in Part 2 of this collection continue this discussion and extend it in a number of ways.

1.2.1 Ordinal and Cardinal Utility

Utilitarianism, in its classical formulation, regards utility as being a measure of an individual's well-being and ranks social alternative x as better than social alternative y if the sum of the individual utilities in x is larger than the sum of the individual utilities in y. Equivalently, if we divide the population into those who gain and those who lose in a move from y to x, then for x to be socially better than y, the sum of the utility gains must exceed the sum of the utility losses. Hence, in order for utilitarianism to be a meaningful doctrine, individual well-being must be measurable by a utility function that permits interpersonal comparisons of utility gains and losses.

Following the ordinalist revolution in utility theory in the 1930s, which has been illuminatingly surveyed by Mandler (1999), economists came to doubt whether such functions existed. For an ordinalist, a utility function is simply a convenient way of summarizing an individual's preferences by assigning numbers to alternatives in a way that preserves the order of preference, and nothing more. For example, if an individual prefers x to y, then x is assigned a larger utility number than y. More generally, if the set of alternatives is X, then for each alternative x in X, a utility function U assigns a number $U(x)$ to x in such a way that $U(x) > U(y)$ if and only if x is preferred to y. In particular, alternatives that are indifferent to each other (i.e., are on the same indifference curve) have the same utility number. As a consequence, if one utility function represents a preference, then so does any other utility function that is obtained from the former by renumbering the indifference curves in a way that preserves the order of preference. This renumbering is formally known as taking an ordinal transform of the original utility function.

However, if utility is ordinal in this sense, then utilitarianism is not meaningful, at least if well-being is identified with preference. To see why, suppose that there are only two individuals, Antoinette and Bernard, with Antoinette preferring x to y and Bernard having the reverse preference. If Antoinette assigns the utility numbers 2 and 0 to x and y, respectively, whereas Bernard assigns them the numbers 1 and 2, respectively, then the sum of the utilities for x is 3 and the sum of utilities for y is 2. A utilitarian would then declare x to be better than y. However, Bernard could equally well use the utility number 5 for y, with x assigned 1 as before, which increases the utility sum for y to 5. A utilitarian would now regard y as better than x even though nobody's preference has changed.

A challenge to this ordinalist perspective was provided by the expected utility theory developed by von Neumann and Morgenstern (1944). In their theory, individual behavior in risky situations provides cardinal information about preferences, or so they argued. In the standard expositions of von Neumann–Morgenstern utility theory, uncertainty is modeled using lotteries over a finite set of sure outcomes. With a lottery, once the uncertainty has been resolved, the outcome will be some alternative from the set $X = \{x_1, \ldots, x_M\}$. A lottery specifies the probability with which each of these outcomes will occur. Thus, with the lottery $p = (p_1, \ldots, p_M)$, the probability of obtaining outcome x_m is p_m. Of course, these probabilities sum to 1. The set of all such lotteries is denoted by $\mathcal{L}$. A decision maker in the von Neumann–Morgenstern theory must act before the uncertainty is resolved, and he does this by choosing a lottery from this set. As in all rational choice models, this choice is governed by preferences, which in this case are preferences over the set of lotteries $\mathcal{L}$.

Von Neumann and Morgenstern argued that the preferences of a rational individual faced with this kind of uncertainty should conform to a set of properties that imply that there exists a utility function U representing the preferences on the set of lotteries $\mathcal{L}$ with the property that the utility $U(p)$ from any lottery p is the expected value (using these probabilities) of the utilities obtained from the sure outcomes in X. We can think of this individual as having two utility functions – U on the set of lotteries $\mathcal{L}$ and V on the set of sure outcomes X. Letting e^m denote the lottery in which outcome x_m is obtained for certain, we must have $V(x_m) = U(e^m)$. Thus, lottery p is preferred to lottery q if and only if the expected utility $U(p) = \sum_{m=1}^{M} p_m V(x_m)$ from p is larger than the expected utility $U(q) = \sum_{m=1}^{M} q_m V(x_m)$ from q. Rather confusingly, both the utility function U on $\mathcal{L}$ and the utility function V on X are referred to as being a von Neumann–Morgenstern utility function.

In order for the preferences over lotteries to be described in terms of expected utilities, it is not possible to use an arbitrary ordinal transform of V in this calculation; rather, only increasing affine transforms are permissible. This means that if $\bar{V}$ is related to V by the equation $\bar{V}(x) = a + bV(x)$ for all x in X, where a and b are numbers with $b > 0$, then $\bar{V}$ can be used instead of V when computing expected utilities without altering the underlying preference over lotteries. A utility function that is uniquely defined up to an increasing affine transform is said to be cardinal. As can be easily verified, in determining whether the utility gain (or loss) from outcome w to y exceeds the utility gain (or loss) from y to z, it does not matter if the utility function V or any increasing affine transform of V is used to evaluate these differences; all of these functions compare utility differences in the same way. For this reason, a utility gain or loss is meaningful with a cardinal utility function.

However, even if one accepts the von Neumann and Morgenstern premises, their theory only justifies comparing utility gains and losses intrapersonally, not interpersonally, as is required by utilitarianism. As discussed in Weymark (2005), for this, among other, reasons, for much of the decade following the first publication of von Neumann and Morgenstern's expected utility theory, most commentators argued that this theory had little or no relevance for welfare economics, let alone providing the kind of utility functions needed for utilitarian calculations.

1.2.2 Harsanyi's Impartial Observer and Social Aggregation Theorems

In Harsanyi (1953), John Harsanyi set out to refute this claim. In doing so, he laid the foundations for a rational choice–theoretic defense of utilitarianism. For Harsanyi, welfare judgments are the impersonal preferences expressed by an impartial observer who ranks social alternatives based on a sympathetic, but impartial, concern for the interests of everyone in society. Specifically, the impartial observer engages in a thought experiment in which he imagines having an equal chance of being any of the n members of society, complete with that person's preferences and objective circumstances. Following von Neumann and Morgenstern, Harsanyi supposed that the set of social alternatives is the set of lotteries $\mathcal{L}$ on a finite set of sure outcomes X and that each individual has preferences over these lotteries that satisfy the axioms of expected utility theory. The impartial observer also faces a lottery, but it is a lottery in which both his identity and the outcome of the actual lottery are uncertain, and for this reason it is called an extended lottery.

In this framework, the observer is sympathetic to the interests of the individuals if his ordering of extended lotteries in which he is some particular person for certain coincides with that individual's ranking of the underlying lotteries in $\mathcal{L}$, what Harsanyi (1977c) calls the Principle of Acceptance. Provided that the observer is sympathetic in this sense and that his preferences satisfy the expected utility axioms, then for each individual i and each sure outcome x in X, the observer's von Neumann–Morgenstern utility function assigns a utility number to the pair (i, x). This number is interpreted as being person i's utility from x. Furthermore, extended lotteries are ranked by the observer according to the expected values of these utilities. By then restricting attention to extended lotteries in which there is an equal chance of being anyone in society and in which the lottery over the alternatives in X is the same for everyone, the observer has implicitly ranked lotteries in $\mathcal{L}$ according to their average utility, that is, by an average utilitarian rule.[7] In this way, Harsanyi has reduced the problem of ranking social alternatives to one of individual decision making under risk, thereby showing the relevance of von Neumann–Morgenstern utility theory for welfare economics

[7] Because the number of individuals is fixed in Harsanyi's argument, there is no distinction between total and average utilitarianism.

in general, and to utilitarian welfare economics in particular. In Weymark (1991), this result is referred to as Harsanyi's *impartial observer theorem*.

In Harsanyi's thought experiment, interpersonal utility comparisons are implicit in the impartial observer's ordering of extended lotteries. For the observer, the analogue of a sure outcome in the von Neumann–Morgenstern theory is being some person i with outcome x, and, as we have noted, the utility assigned to this outcome is interpreted as being person i's utility from x. If one accepts, for the reasons given above, that comparisons of utility gains and losses are meaningful in the von Neumann–Morgenstern theory, it follows that by using a von Neumann–Morgenstern utility function to characterize the observer's preferences over extended lotteries that interpersonal comparisons of utility gains and losses are meaningful, as required by utilitarianism. For example, to determine whether the difference in utility for person i in going from w to x exceeds the utility difference for person j in going from y to z, one simply checks to see whether the difference in utility that the impartial observer attributes to the move from (i, w) to (i, x) exceeds the utility difference in going from (j, y) to (j, z) using his von Neumann–Morgenstern utility function over extended lotteries. In other words, all interpersonal utility comparisons are transformed into intrapersonal utility comparisons for the observer.

The idea of deriving substantive principles of morality based on rational individual decision making in a hypothetical situation in which the decision maker is deprived of morally irrelevant information is one of Harsanyi's greatest achievements.[8] In Rawlsian terminology, Harsanyi's impartial oberver can be described as operating from behind a veil of ignorance, and the decision problem that he is facing can be called an original position. As we have seen, Rawls (1971) used his own version of the original position to derive his principles of justice. For reasons presented at length in Rawls (1971, sections 27 and 28), he rejected both Harsanyi's version of utilitarianism and his formulation of the original position with its reliance on expected utility theory, which Rawls argued leads the impartial observer to gamble on the principles that govern the structure of the most basic social institutions of society. In Rawls's original position, less information is permitted behind the veil, with the consequence, or so Rawls argued, that social institutions should be designed to maximize the prospects of the worst-off individuals as measured by an index of primary goods, once priority has been given to ensuring that everyone enjoys equal liberties and fair equality

[8] Unbeknowst to Harsanyi, a similar idea had been suggested by Vickrey (1945), but Vickrey did not develop this idea in any detail.

of opportunity.[9] As we have noted earlier, in a response to Rawls, Harsanyi (1975) defended his use of expected utility theory and argued that Rawls's maximin reasoning leads to unsatisfactory outcomes.

In Harsanyi (1955), Harsanyi introduced a conceptually distinct argument in support of a weighted form of utilitarianism, also based on von Neumann–Morgenstern expected utility theory, what Weymark (1991) has called Harsanyi's social aggregation theorem. As in Harsanyi's impartial observer theorem, the objective is to provide a social ranking of the lotteries $\mathcal{L}$ on a finite set of sure outcomes X. But now, in addition to there being n individual preferences on these lotteries that satisfy the von Neumann–Morgenstern expected utility axioms, there is also assumed to be a social preference on this set that also satisfies these axioms. Harsanyi interpreted this social preference as being the moral or social preference of an actual individual. The individual and social preferences are related to each other by some form of the Pareto principle. In its Pareto Indifference form, this principle requires that if everyone is indifferent between two alternatives, then the social preference should be as well. With these assumptions, Harsanyi showed that if von Neumann–Morgenstern utility functions are used to represent the individual and social preferences over lotteries, then the lotteries in $\mathcal{L}$ are socially ranked according to a weighted sum of the individual utilities associated with them. That is, there is a weight a_i for each individual i such that lottery p is socially preferred to lottery q if and only if $\sum_{i=1}^{n} a_i U_i(p) > \sum_{i=1}^{n} a_i U_i(q)$, where U_i is the von Neumann–Morgenstern utility function used to represent person i's preferences over lotteries. Note that scaling all of the weights by multiplying them by a common positive number does not affect this ranking.

In this version of Harsanyi's social aggregation theorem, there is no guarantee that the individual weights are unique up to a factor of proportionality, nor that they are all positive. Uniqueness of the weights in this sense follows if an assumption called Independent Prospects, which is implicitly used by Harsanyi (1955) in his proof, is adopted. This assumption requires that, for each individual, there exists a pair of lotteries for which this person is

[9] Rawls does not view his original position in terms of a single individual decision making problem. Rather, he regards his original position as a hypothetical situation in which a number of parties – representatives of family or genetic lines (Rawls, 1971, section 25) or representative citizens (Rawls, 2001, section 6) – seek to reach agreement about the basic structure of society. But, as Barry (1989, p. 196) and other commentators have noted, the parties in Rawls's original position all have the same information and objectives, so there is no substantive difference between his multiparty bargain and a description of the original position that only has one individual behind the veil.

not indifferent, but for which everyone else is. Positivity of the weights is then obtained by strengthening Pareto Indifference to Strong Pareto, which requires that the social preference weakly prefers lottery p to lottery q if every individual weakly prefers p to q, with strict social preference between these lotteries if at least one of these individual preferences is strict. Harsanyi interpreted his social aggregation theorem as providing a defense of using (weighted) utilitarianism to rank alternatives socially.

1.2.3 The Sen–Weymark Critique

Beginning with the critique of Sen (1976), there has been considerable controversy about whether Harsanyi's impartial observer and social aggregation theorems should be interpreted as theorems about utilitarianism. Sen has argued that each of the axioms of von Neumann–Morgenstern expected utility theory only place restrictions on the rankings of lotteries, and, hence, their theory is ordinal. As a consequence, any increasing transform of a von Neumann–Morgenstern utility function defined on a set of lotteries, a set of extended lotteries, or a set of sure outcomes is a satisfactory representation of the corresponding preference relation. But if this is the case, then von Neumann–Morgenstern utility functions are not cardinal, as required if they are to be used in utilitarian calculations. Furthermore, there is no reason why the utility function that measures an individual's well-being for the purpose of such computations should be a von Neumann–Morgenstern utility function, rather than some increasing, but nonaffine, transform of such a function, even if this individual's preferences conform to the axioms of expected utility theory. John Broome, in his contribution to this volume, calls this argument the "standard objection" to Harsanyi's theorems.

The force of Sen's objection can be most easily seen by considering its implications for Harsanyi's social aggregation theorem. Suppose that when the von Neumann–Morgenstern utility functions U_i are used, lottery p is socially preferred to lottery q if and only if $\sum_{i=1}^{n} a_i U_i(p) > \sum_{i=1}^{n} a_i U_i(q)$. But suppose that for each person i, the welfare-relevant utility function is $\bar{U}_i = (U_i)^3$, the cube of U_i. To obtain the same social ranking of the lotteries using the welfare-relevant utility functions, it must now be the case that p is socially preferred to q if and only if $\sum_{i=1}^{n} a_i \sqrt[3]{\bar{U}_i(p)} > \sum_{i=1}^{n} a_i \sqrt[3]{\bar{U}_i(q)}$.[10] Thus, when the von Neumann–Morgenstern utility functions U_i are used, it appears that lotteries are socially ranked using a weighted sum of utilities,

[10] Note that $\sqrt[3]{\bar{U}_i(p)} = U_i(p)$.

whereas when the utility functions $\bar{U}_i$ are used, it appears that lotteries are socially ranked using a weighted sum of the cube roots of utilities, even though in both cases the social ranking is the same.

The "standard objection" applies equally well to both the impartial observer and social aggregation theorems. Sen (1986) has raised a further issue that only pertains to the latter theorem. His argument can be illustrated with a simple two-person example. As we have seen, given the assumptions of this theorem, for each person i, $i = 1, 2$, if the von Neumann–Morgenstern utility function V_i is used to represent i's preferences over the lotteries in $\mathcal{L}$, then there is a weight a_i for him such that lottery p is socially preferred to lottery q if and only if $a_1 V_1(p) + a_2 V_2(p) > a_1 V_1(q) + a_2 V_2(q)$. Instead of using V_2 to represent person 2's preferences, we could equally well use $\bar{V}_2 = \frac{1}{2} V_2$ because $\bar{V}_2$ is an increasing affine transform of V_2. Because person 2's utility for each of the lotteries is now half of what it was before taking the transform and because the social preferences have not changed, to obtain the same social ranking of the lotteries using a weighted utilitarian rule, we must now multiply person 2's weight by 2. In other words, lottery p is socially preferred to lottery q if and only if $a_1 V_1(p) + 2a_2 \bar{V}_2(p) > a_1 V_1(q) + 2a_2 \bar{V}_2(q)$. However, weighted utilitarianism requires that the same weights be used to aggregate the individual utilities, regardless of what the individual utility functions turn out to be. In constrast, in Harsanyi's theorem, the weights depend on the choice of utility functions used to represent the individual preferences.

Sen's arguments are quite informal and easily misunderstood.[11] They were subsequently formalized and extended by Weymark (1991). A good introduction to Sen's critique and its formalization by Weymark from someone who endorses their conclusions may be found in Roemer (1996, chapter 4).

1.2.4 Roemer on the Sen–Weymark Critique

Weymark's formalization of Sen's argument that von Neumann–Morgenstern utility theory is ordinal and, therefore, that Harsanyi was not justified in interpreting his theorems as providing choice-theoretic foundations for utilitarian principles is fairly abstract. In Chapter 5, John Roemer's main purpose is to make this argument more widely accessible by presenting a simple example that captures the essential features of this critique. He does so using Harsanyi's impartial observer theorem as the basis for his discussion.

[11] Harsanyi (1977b), in his response to Sen, appears to have misinterpreted what Sen was saying about these issues.

What emerges clearly from Roemer's chapter is that the impartial observer needs to make the units in which individual gains and losses are measured comparable, and this is done by taking increasing affine transforms of each person's von Neumann–Morgenstern utility function so that the utility it assigns to any sure outcome is the same as the utility assigned by the impartial observer's utility function to that outcome when he is that person for certain. However, as his example shows, the resulting interpersonal comparisons need not be the ones that would be made if they were instead based on the utility functions that measure individual well-being.

Roemer goes on to show that an analogous problem arises with individual choice using the state-contingent alternatives model of uncertainty. In this model of uncertainty, there are a number of states of the world, only one of which will be realized once the uncertainty is resolved. Suppose that there are M possible states and that in each state m, the set of possible outcomes is X. These are the outcomes that are possible ex post once the uncertainty is resolved. An ex ante alternative is a list $x = (x^1, \ldots, x^M)$ that specifies the outcome x^m in X that will occur should state m eventuate.[12] If the same outcome is achieved in every state, then there is no uncertainty, and the alternative is said to be certain. Before the realization of the state, the decision maker chooses an ex ante alternative from among those that are feasible.

Versions of expected utility theory have been developed for this model of uncertainty by Arrow (1964) and Savage (1954), among others. In these theories, the axioms on preferences imply that there is a state-independent utility function V on X (representing this person's tastes or values for ex post outcomes) such that ex ante alternative x is preferred to ex ante alternative y if and only if $EV(x) > EV(y)$, where for all x in X^M (the set of all possible ex ante alternatives), $EV(x) = \sum_{m=1}^{M} p_m V(x^m)$ and p_m is the probability of state m occuring.[13] That is, x is preferred to y if and only if the expected utility from x exceeds the expected utility from y. Unlike in the von Neumann–Morgenstern framework, the probabilities that are used to take these expectations are the same for all alternatives. Depending on the particular version of this model that is employed, these probabilities may be objective or subjective. The function V is called a Bernoulli utility function.

[12] In some versions of this model, an ex ante alternative is called an act. In their contribution to this volume, Blackorby, Donaldson, and Weymark call an ex ante alternative a state-contingent alternative, and they call an ex post alternative either a prospect or a social alternative.

[13] In Savage's version of this model, the number of states is infinite and sums are replaced by integrals when computing an expected utility.

It evaluates outcomes x in X ex post once the uncertainty has been resolved. In contrast, the expected Bernoulli function EV evaluates alternatives x in X^M ex ante, that is, before the resolution of the uncertainty.

What Roemer shows is that the utilities obtained in different states play a role similar to that played by the utilities of different individuals in Harsanyi's impartial observer theorem. As a consequence, the implicit interstate comparisons of utility gains and losses provided by the expected utility representation of preferences need not coincide with the inter-state comparisons that are obtained using a utility function that measures well-being in each state.

Versions of Harsanyi's social aggregation theorem for the state-contingent model of uncertainty have been developed by Hammond (1981) and Blackorby, Donaldson, and Weymark (1999) for the case in which everyone has the same probabilities.[14] Blackorby, Donaldson, and Weymark have argued that, as in von Neumann–Morgenstern expected utility theory, the state-contingent alternatives model of uncertainty is also ordinal and, therefore, is subject to the criticisms raised by Sen.

1.2.5 Social Welfare Functionals and Welfarism

Arrow (1951) modeled the social choice problem as one of aggregating profiles of individual preference orderings over a set of alternatives, one for each person, into a social ordering of the same alternatives. The mapping that assigns a social ordering to each profile of preferences in its domain is a called an Arrovian social welfare function. In Arrow's problem, this aggregation procedure is designed before the individual preferences are known and, hence, is meant to apply to more than one profile. In contrast to this multiprofile approach, Harsanyi assumed that the individual preferences are known, so in Arrovian terminology, he can be viewed as engaging in a single-profile aggregation exercise. The Arrovian social choice framework is not rich enough to take account of any information that might be available concerning interpersonal comparisons of utility. For that reason, Sen (1970) proposed an alternative framework in which profiles of utility functions on the set of alternatives are aggregated into a social ordering of these alternatives. The mapping that associates social orderings with profiles of utility

[14] Later in this section, we shall discuss social aggregation for the state-contingent alternatives model for the case in which probabilities are subjective, as in Savage (1954), and, hence, can differ from person to person. See Mongin and d'Aspremont (1998) for a discussion of variants and extensions of Harsanyi's social aggregation theorem that have been established using different models of choice under uncertainty.

functions is a called a social welfare functional. As with Arrow's aggregation procedure, different profiles, now profiles of utility functions, can, in principle, be assigned different social orderings of the alternatives.

Note that utility functions, not preferences, are primitive concepts for a social welfare functional. However, any profile of utility functions $\mathbf{U} = (U_1, \ldots, U_n)$ on a set X of alternatives implicitly defines individual preferences on X. For example, individual i prefers x to y if and only if $U_i(x) > U_i(y)$.

The ability to make intrapersonal and interpersonal utility comparisons may be limited. This limited information is formalized by grouping profiles of utility functions that cannot be distinguished from one another because they contain the same usable information into sets and then requiring the social welfare functional to assign the same social ordering to each profile in the same set. This grouping results in a partition of the domain of the social welfare functional into sets of informationally equivalent profiles. Because different social orderings can be assigned to profiles in different cells of this partition, if it becomes possible to distinguish some profiles of utility functions that were previously indistinguishable, say $\mathbf{U}$ and $\bar{\mathbf{U}}$, then the partition becomes finer and $\mathbf{U}$ and $\bar{\mathbf{U}}$ no longer need to be assigned the same social preference because they are now in different cells of the finer partition.

For example, suppose that, as in Arrow (1951), only preference information is available. If, for each person i, $\bar{U}_i$ is an increasing transform of U_i, with possibly different transforms used for different individuals, then the corresponding profiles of utility functions $\bar{\mathbf{U}}$ and $\mathbf{U}$ represent, person by person, the same individual preferences and, hence, are grouped together and assigned the same social ordering. In this case, the utility functions are said to be ordinally measurable and (interpersonally) noncomparable. Suppose, instead, that the profile $\bar{\mathbf{U}}$ is informationally equivalent to the profile $\mathbf{U}$ if and only if there exist numbers $a_1, \ldots, a_n$ and a positive number b such that $\bar{U}_i = a_i + bU_i$. It is now possible to make interpersonal comparisons of utility differences, as required by utilitarianism, because for any pair of individuals i and j and any pair of alternatives x and y, $\bar{U}_i(x) - \bar{U}_i(y) > \bar{U}_j(x) - \bar{U}_j(y)$ if and only if $U_i(x) - U_i(y) > U_j(x) - U_j(y)$. In this case, the utility functions are said to be cardinally measurable and unit comparable. Using these kinds of transforms results in a finer partition of the possible profiles of utility functions than in the Arrovian case.

Utilitarianism, in either its classical or weighted formulations, is an example of a welfarist principle. Such principles are consequentialist in the sense that only the consequences of decisions matter for the purposes of social evaluation. Furthermore, it is only the utility consequences that matter, not the physical outcomes that are achieved or the individual utility functions used to generate these utilities. Utility consequences are summarized by a vector of numbers, the utilities of the individuals being considered.[15] If a social welfare functional is welfarist, then the social ranking of alternatives is completely determined by a social ordering of utility vectors, a property called *welfarism*. This social ordering is called a social welfare ordering. Specifically, there is a social welfare ordering of the attainable vectors of individual utilities with the property that if x and y are any two alternatives in the set of possible alternatives, $\mathbf{U} = (U_1, \ldots, U_n)$ is the profile of individual utility functions, and $u = (U_1(x), \ldots, U_n(x))$ and $v = (U_1(y), \ldots, U_n(y))$ are the vectors of individual utilities obtained from x and y, respectively, with the utility functions in $\mathbf{U}$, then u is socially preferred to v according to the social welfare ordering for utility vectors if and only if x is socially preferred to y for the social preference over alternatives that is generated by the social welfare functional when the profile of utility functions is $\mathbf{U}$. With classical utilitarianism, u and v are compared by seeing which of these two utility vectors has the highest utility sum.

A social welfare ordering can be interpreted as being a social preference over vectors of individual utilities. As is the case with utilitarianism, it may be possible to represent this preference by a utility function. Such a function is called a social welfare function.[16]

If there is only one profile of utility functions in the domain of a social welfare functional, as in Harsanyi's social aggregation theorem, then welfarism is equivalent to requiring that the social welfare functional satisfies Pareto Indifference. On the other hand, if the domain of a social welfare functional includes all possible profiles of utility functions on the set of alternatives, then welfarism is satisfied if and only if it satisfies Pareto Indifference and Independence of Irrelevant Alternatives. The latter condition requires that the social ranking of any two alternatives x and y should be the same for the profiles of utility functions $\mathbf{U}$ and $\bar{\mathbf{U}}$ if the individual utilities obtained from x and from y are the same in both of these profiles.[17] On this unrestricted

[15] A vector of numbers (here, individual utilities) is a list in which the order the numbers appear matters.

[16] Unfortunately, the phrase "social welfare function" is used both for this kind of function and for an Arrovian social welfare function. In the sebsequent discussion, when the phrase "social welfare function" is used by itself, we are referring to a real-valued function whose arguments are vectors of utility numbers.

[17] When utilities are ordinal and interpersonally noncomparable, this condition is equivalent to requiring the social ranking of two alternatives to depend only on the individual orderings of them, which is the original Arrovian Independence of Irrelevant Alternatives condition.

domain, welfarism is also equivalent to Strong Neutrality, which requires the social ranking of x and y when the profile of utility functions is $\mathbf{U}$ to be the same as the social ranking of w and z when the profile of utility functions is $\bar{\mathbf{U}}$ if the individual utilities for x (resp. y) in the profile $\mathbf{U}$ are the same as the individual utilities for w (resp. z) in the profile $\bar{\mathbf{U}}$. This neutrality condition is simply a formal definition of what it means for a social welfare functional to be welfarist.[18]

1.2.6 Blackorby, Donaldson, and Weymark on Social Aggregation under Uncertainty

In Chapter 6, Charles Blackorby, David Donaldson, and John Weymark reformulate Harsanyi's social aggregation problem for the state-contingent alternatives model of uncertainty in which everyone agrees on the probabilities using Sen's social welfare functional framework.[19] By using a framework that can accommodate interpersonal comparisons of utility, they are able to avoid the first of Sen's criticisms, namely, that the interpersonal utility comparisons required for utilitarianism to be meaningful are not available if only preference information is considered. However, because they start with individual utility functions that are meant to serve as the appropriate measures of well-being for social welfare purposes, it is then necessary to say when a utility function satisfies the expected utility hypothesis. They consider two possibilities. In the first case, a utility function U on the set of ex ante alternatives X^M is said to satisfy the expected utility hypothesis if U is the expected value of some Bernoulli utility function V on the set of ex post outcomes X. Broome (1991) calls this the Bernoulli hypothesis. In the second case, U is only required to be an increasing transform of a utility function that satisfies this hypothesis. In both cases, the preferences over ex ante alternatives underlying these functions are expected utility preferences in the sense described above when discussing John Roemer's contribution. For this reason, Blackorby, Donaldson, and Weymark argue that expected utility theory, by itself, provides no reason for restricting attention to the first of these possibilities.

In their chapter, Blackorby, Donaldson, and Weymark suppose that the profiles of individual utility functions that they consider satisfy the expected utility hypothesis in one of the two variants just discussed. They consider

[18] See Mongin and d'Aspremont (1998) and Bossert and Weymark (2004) for detailed discussions of welfarism and the literature on social choice with interpersonal utility comparisons.

[19] It is a straightforward exercise to restate their results using the lottery model of uncertainty. See also Mongin (1994) and the chapter by d'Aspremont and Mongin in this volume for related multiprofile extensions of Harsanyi's social aggregation theorem.

a number of different domains for a social welfare functional, including (i) a domain in which there is only a single profile of utility functions (as in Harsanyi's social aggregation theorem), (ii) a domain in which every profile of utility functions satisfying the expected utility hypothesis is possible, and (iii) a domain in which every profile corresponds to the same profile of expected utility preferences. For each of these cases and for each of the two versions of the expected utility hypothesis, they identify the kinds of social welfare functionals that satisfy the Strong Pareto principle and, in the multiprofile cases, some further assumptions. Either explicitly or implicitly, for each of their multiprofile domains, they assume that every possible vector of individual utilities can be generated by some alternative and some profile of utility functions, a property that d'Aspremont and Mongin (see below) call Complete Utility Attainability. As they show, their assumptions in the multiprofile cases imply that welfarism is satisfied, a property that is implied by just Strong Pareto when there is only one profile of utility functions (because Strong Pareto implies Pareto Indifference).

For each of their domains, Blackorby, Donaldson, and Weymark show that when only utility functions that satisfy the Bernoulli hypothesis are considered, then any social welfare functional satisfying their assumptions is weighted utilitarian, with the same weights being used for every profile in the domain (thereby, avoiding the last of Sen's criticisms). However, when utility functions that satisfy their less restrictive definition of the expected utility hypothesis are also considered, they show that for the first of their multiprofile domains [case (ii)], no social welfare functional satisfies their assumptions, whereas for the other two domains, transforms must be applied to the welfare-relevant utilities before computing a weighted sum, as was done in Section 1.2.3 when taking the cube roots of utilities in the example used to illustrate one of Sen's criticisms of Harsanyi's theorems. Therefore, the social welfare functional is not weighted utilitarian.

In effect, this latter result is another manifestation of Sen's point that Harsanyi's utilitarian conclusions depend on the use of expected utility representations of the individual preferences. As Blackorby, Donaldson, and Weymark see no good reason for supposing that the welfare-relevant utility functions satisfy the Bernoulli hypothesis, they conclude that their extensions of Harsanyi's analysis do not provide support for utilitarianism.

1.2.7 D'Aspremont and Mongin on Welfarism and Social Aggregation

In Chapter 7, Claude d'Aspremont and Philippe Mongin employ a generalization of the von Neumann–Morgenstern framework in which the set of alternatives, which to avoid possible misunderstanding is now denoted Y, is a convex set. A set Y is convex if for any number λ with $0 \leq \lambda \leq 1$ and any two alternatives x and y in Y, the weighted combination $\lambda x + (1 - \lambda)y$ is also in Y.[20] The set of lotteries $\mathcal{L}$ considered earlier is a convex set. In d'Aspremont and Mongin's model, a von Neumann–Morgenstern utility function is a function U on Y with the property that utility assigned to $\lambda x + (1 - \lambda)y$ is $\lambda U(x) + (1 - \lambda)U(y)$. If λ and $1 - \lambda$ are probabilities, this expression is simply the formula for computing the expected utility of the lottery in which x and y are obtained with the probabilities λ and $1 - \lambda$, respectively. When this is not the case, this model shares the formal properties, but not the expected utility interpretation, of the von Neumann–Morgenstern model.

D'Aspremont and Mongin consider a social welfare functional whose domain consists of all of the profiles of von Neumann–Morgenstern utility functions that represent the same profile of preferences on Y. This domain is the analogue in their model of the third of the domains analyzed by Blackorby, Donaldson, and Weymark for the case in which the utility functions are assumed to satisfy the Bernoulli hypothesis. On this domain, they show that the social welfare functional is weighted utilitarian if the social orderings of the alternatives in Y assigned to each profile of utility functions satisfy the von Neumann–Morgenstern axioms and both Strong Neutrality (which is a strengthening of Pareto Indifference) and Complete Utility Attainability are satisfied.[21] As discussed above, the weights used to aggregate the individual utilities are unique up to a factor of proportionality and positive if Independent Prospects and Strong Pareto are assumed.

To establish their theorem, d'Aspremont and Mongin first show that welfarism holds on their domain when Strong Neutrality and Complete Utility Attainability are assumed. With this attainability assumption, the set of possible vectors of individual utility vectors is the n-dimensional Euclidean space $\mathbb{R}^n$, which is a convex set. Provided that the social orderings of the alternatives in Y satisfy the von Neumann–Morgenstern axioms, d'Aspremont and Mongin show that the social ordering of $\mathbb{R}^n$, interpreted as an ordering of vectors of utilities, whose existence is guaranteed by the welfarism

[20] A necessary condition for a set to be convex is that it exhibits sufficient mathematical structure so that it makes sense to multiply an alternative by a real number and to add alternatives together. This is the case when the alternatives are vectors of numbers, for example, vectors of probabilities or utilities. For the vectors x and y, the weighted combination $\lambda x + (1 - \lambda)y$ is computed component by component.

[21] It should be pointed out that the chapters by Blackorby, Donaldson, and Weymark and by d'Aspremont and Mongin were prepared independently.

theorem, in fact satisfies the von Neumann–Morgenstern axioms adapted to apply to the model in which the set of alternatives can be any convex set. Furthermore, one of these axioms (von Neumann and Morgenstern's independence axiom) is equivalent to the restriction on the social ordering of utility vectors that is implied when utilities are cardinally measurable and unit comparable. In effect, the social decision maker is implicitly making interpersonal comparisons of utility differences using the individual von Neumann–Morgenstern utility functions. Whether these are the appropriate utility functions to compare when making utilitarian calculations is, as we have explained, one of the main issues considered in Sen's critique of Harsanyi. D'Aspremont and Mongin believe that they are and, therefore, that their theorem provides an axiomatization of utilitarianism that is in the spirit of Harsanyi's social aggregation theorem.

Harsanyi was interested in providing a choice-theoretic foundation for the classical form of utilitarianism in which the individual welfare weights are all equal. As d'Aspremont and Mongin note, in order for a social aggregation procedure to both determine the welfare weights independently of the utility functions used to represent the individual preferences and to have these weights all be equal, it is necessary to employ a multiprofile approach, as they have done. They obtain the classical form of utilitarianism by requiring their social ordering of utility vectors to treat individuals symmetrically in the sense that permuting utilities among the individuals is a matter of social indifference.

1.2.8 Hild, Jeffrey, and Risse on Ex Ante versus Ex Post Social Aggregation

When probabilities are subjective, new issues arise for social aggregation, and they are the subject of Chapter 8 by Matthias Hild, Richard Jeffrey, and Mathias Risse. Now, not only can the tastes or values of individuals differ, so can their probability assessments. In the ex ante approach to social aggregation, individuals' ex ante expected utility functions are aggregated into a social expected utility function. This is the approach previously discussed for the state-contingent alternatives model of uncertainty with common probability assessments. Alternatively, the individual probabilities and ex post utilities can be separately aggregated into social or group probabilities and utilities, which are then combined to form a social expected utility function – the ex post approach to social aggregation. Hild, Jeffrey, and Risse first provide an introduction to the problems that can arise with the ex ante approach and relate these results to Harsanyi's social aggregation theorem.

They then show that the ex post approach can exhibit a rather undesirable feature that they call flipping. We begin with the ex ante approach.

One of the most striking results in the ex ante approach to social aggregation is what Broome (1991) calls the probability agreement theorem. Suppose that uncertainty is modeled using state-contingent alternatives with a finite number of states and that the utility functions used to represent the individual and social orderings of the ex ante alternatives satisfy the Bernoulli hypothesis, which requires that the utility of an alternative is the expected value of the utilities of the correponding ex post outcomes. However, now suppose that the probabilities can differ between individuals. Further suppose that for each individual there exists a pair of certain (or, equivalently, ex post) alternatives for which this person is not indifferent, but for which everyone else is. This assumption is simply the Independent Prospects assumption from the lottery model of uncertainty reformulated for state-contingent alternatives. Given these assumptions, in order for the Strong Pareto principle to be satisfied, it turns out that everybody must use the same probabilities, a result that is generally interpreted as being an impossibility theorem.[22]

If weaker versions of the Pareto principle are used, further possibilities emerge. For example, if Pareto Indifference is used instead of Strong Pareto in the preceding discussion and everyone has different probabilities (more precisely, the individual probabilities are affinely independent), not just different preferences over the certain alternatives, then someone is both a probability and utility dictator in the sense that the social probabilities are his probabilities and the social preferences for ex post outcomes either coincide with or completely reverse the preferences of this individual.

Mongin (1995) has established the most general versions of these ex ante impossibility theorems. Somewhat imprecisely, he has shown that when there is taste heterogeneity, then there must be probability agreement or a probability dictator and if there is probability heterogeneity, then there must either be a utility dictator or there must be a great deal of taste homogeneity, depending on which version of the Pareto principle is used. Hild, Jeffrey, and Risse discuss two of these impossibility theorems. For further discussion, see Mongin and d'Aspremont (1998, section 5.4).

Hild, Jeffrey, and Risse argue that the source of the ex ante impossibilities is its conflation of views about facts and values. For example, suppose there are only two individuals who have the same expected utilities, but both

[22] Broome (1991, section 7.1) presents a simple proof of this theorem for the two-person case.

differ in their probabilities and in their preferences over the ex post outcomes. Pareto Indifference would then imply that the social preference over ex ante alternatives should agree with this common individual preference even though this common preference arose for very different reasons. By adopting an ex post perspective in which probabilities and tastes are aggregated independently, matters of fact and of values can be kept separate.[23] However, this way of circumventing the ex ante impossibilities comes at the cost of Hild, Jeffrey, and Risse's flipping theorem.

For this ex post impossibility theorem, Hild, Jeffrey, and Risse use yet another version of expected utility theory due to Bolker (1967) and Jeffrey (1965), which, as in Savage (1954), allows for subjective probabilities. Recall that in the Savage model, there is a set X of possible ex post outcomes and a set of states of the world, which we shall denote by S. We now need to introduce the concept of an event, which is simply a subset of states. To say that event E has occurred means that we know that the true state is one of the states in E. If state s in S occurs for certain, this is also an event, denoted as $\{s\}$.

In Bolker–Jeffrey expected utility theory, there is no separate set of ex post outcomes X. Rather, outcomes are identified with events in the state space S. An alternative specifies which event occurs. As a consequence, probabilities and utilities are both assigned to the same kind of objects – events. Preferences are also over events. In the Bolker–Jeffrey theory with a finite number of states, there is a probability $p(E)$ and a utility $U(E)$ associated with each event E such that (i) the probability of an event is the sum of the probabilities of the states that comprise this event and (ii) the utility of an event is the expected utility of this event conditional on this event occuring for certain. Formally, this expected utility property says that $U(E) = \sum_{s \in E} p(\{s\}|E)U(\{s\})$, where $p(\{s\}|E)$ is the conditional probability of state s given the occurrence of event E. Broome (1990) and Bradley (2005) have provided particularly lucid introductions to Bolker–Jeffrey expected utility theory. For this model of uncertainty, Broome (1990) has established a version of Harsanyi's social aggregation theorem by

[23] Hild, Jeffrey, and Risse attribute this view to Raiffa (1968, section 12). As they note, Raiffa's point was made in the context of a discussion of a theorem due to Richard Zeckhauser that was published many years later in Hylland and Zeckhauser (1979). It might seem that the Hylland–Zeckhauser theorem is an ex post aggregation impossibility theorem because probabilities and utilities for ex post outcomes are aggregated separately. However, the conflict between their axioms arises because the Pareto principle is applied to ex ante expected utilities.

making assumptions that ensure that everyone agrees on the probabilities and Bradley (2005) has established a version of the probability agreement theorem.

In the Hild–Jeffrey–Risse flipping theorem, social utility is the sum of the individual utilities and social probabilities are the average of the individual probabilities. In order for this utilitarian aggregation procedure to be well defined, it is assumed that the individual utilities are calibrated in such a way that summing utilities is meaningful. In describing the states of the world, we must specify how finely states are to be discriminated from one another. For example, should the description of a state be "dessert is chosen" or should the description be more specific, say, by distinguishing between "ice cream is chosen" and "pie is chosen," if ice cream and pie are the two possible kinds of desserts. What the flipping theorem shows is that when individual utilities and probabilities are aggregated in the manner previously discussed, for some specifications of these utilities and probabilities, the social ranking of an event E relative to its complement can reverse as we move to a finer description of states in E. However, the individual utilities of these events have not changed and, hence, their utilitarian sum is unchanged as well. Therefore, the social ranking of E and its complement have not changed. We, thus, have a contradiction.

For example, initially suppose that there are two states, s and t, where s is the state "dessert is chosen." Consider the event $E = \{s\}$; that is, E is the event consisting of the single state s. At this level of refinement, the social utility of E is simply the sum of the individual utilities for E. Now divide s into two states s_1 and s_2, say "ice cream is chosen" and "pie is chosen." Provided that the individuals assign utilities and probabilities to the events $\{s_1\}$ and $\{s_2\}$ in a way that is consistent with what was done at the coarser level of description, then their utilities for the event $E = \{s\} = \{s_1, s_2\}$ will not have changed. Hence, social utility for E is unchanged when social utilities are computed using the original description of the set of states of the world. However, if social utilities are instead computed using the new level of refinement, then the social utility for the event E is obtained by first determining social utilities and probabilities for the events $\{s_1\}$ and $\{s_2\}$ and then calculating the expected value of the social utilities of $\{s_1\}$ and $\{s_2\}$ conditional on the event E occuring for certain. In general, the social utility of E computed in this way will differ from the first way of computing this utility, and, for some values of the individual utilities and probabilities, they may differ by so much that the ranking of E and its complement can flip. In other words, the possibility that the analysis can be carried out with different levels of refinement in the description of states can result in a discrepancy between the social ranking of two events obtained at one level with the social ranking of the same two events when the analysis is carried out at a different level of refinement.

1.3 Goodness and Well-Being

As Rawls (1971, p. 24) has argued, the two central concepts of any moral theory are the good and the right. A teleological theory has an independent concept of the good and what is right is to maximize the good. Utilitarianism is a teleological theory. With utilitarianism, goodness is equated with the sum of some measure of individual goodness or well-being. A variety of different concepts of well-being, including preferences, happiness, pleasure, and want or desire satisfaction, have been employed in utilitarian and other teleological theories. The contributions to Part 3 of this volume take up a number of issues related to the concepts of goodness (both individual and social) and well-being.

1.3.1 Broome on the Coherence of Preference-Based Utilitarianism

The Sen–Weymark critique discussed in the preceding section casts doubt on the meaningfulness of any utilitarian theory based solely on ordinal preferences, whether they are preferences for actual social alternatives (e.g., lotteries) or hypothetical alternatives that also specify what position one is to occupy in society (e.g., extended lotteries). However, there may be more to preferences than simply ranking alternatives. An individual may, for example, have a well-defined concept of the strength of preference that can be used to compare differences in utility. In Chapter 9, John Broome considers whether there can be any preference-based version of utilitarianism that is coherent. He concludes that any utilitarian theory must contain nonpreferencist features.

For Broome, utility is a measure of the goodness of alternatives. There are two challenges that a preference-based utilitarian theory must face. First, it must establish that there is a relevant concept of goodness for which interpersonal comparisons of differences in individual good are meaningful. Second, it must justify ranking alternatives using the sum of the good that the individuals in society obtain with each alternative.

Broome frames his arguments using the lottery formulation of uncertainty. Because the second of these issues is considered at some length in his monograph, *Weighing Goods* (Broome, 1991), he deals with it rather briefly here. Essentially, Broome argues that Harsanyi's social aggregation theorem does not provide a satisfactory justification for the additivity of the utilitarian social objective function because it presupposes that everyone uses the same probabilities. However, as a factual matter, this is clearly false. Broome's alternative argument in support of the additivity of social preferences in his monograph employs both preferencist and nonpreferencist elements.

The main focus of Broome's chapter is on the measurement of goodness. For Broome, individual goodness is a quantitative measure of a betterness relation that Broome identifies with the preferences that this individual would have in certain ideal circumstances (he is well informed, not subject to the heat of the moment, etc.) when he is making judgments on behalf of himself in his actual nonideal circumstances. Broome (1991) has given qualified support for supposing that this betterness relation should satisfy the axioms of expected utility theory. Assuming that these axioms are satisfied, a von Neumann–Morgenstern utility function represents this betterness relation, but, as we have seen, so does any increasing transform of such a function. If there is some reason for singling out von Neumann–Morgenstern representations of the betterness relation from among all of its possible utility representations, then we have identified a measure of individual goodness for which differences in goodness are meaningful intrapersonally.

Broome argues that there is a good reason for using a von Neumann–Morgenstern measure of individual goodness, which he develops using the following example. Preferences (as embodied in the betterness relation) are such that (i) sure outcome A is preferred to sure outcome B, which in turn is preferred to sure outcome C and (ii) the lottery in which A is received with probability 1/3 and C is received with probability 2/3 is indifferent to B for certain. Using the analogy of weighing objects in a pan balance, Broome regards the two (out of three) chances of the loss associated with moving from B to C as exactly balancing the one chance of the gain associated with moving from B to A. It is therefore "natural" to say that the strength of preference for A over B is twice that of B over C and, hence, that the goodness attached to the lottery in (ii) is a probability-weighted sum of the goodnesses attached to its sure outcomes.[24] Thus, it is this appeal to naturalness that justifies the use of a von Neumann–Morgenstern representation to measure individual goodness. Similarly, while we could weigh objects using the cube of their weight in, say, kilograms, such a measure is less natural than using

[24] Harsanyi has also argued that the willingess to take risks, as in Broome's example, provides a way of determining an individual's strength of preference. See, for example, Harsanyi (1979, pp. 296–297). However, as explained in Weymark (1991, pp. 305–307), Harsanyi's argument, unlike that of Broome, presupposes that preferences must be represented by a von Neumann–Morgenstern utility function, which we have seen need not be the case.

kilograms because the latter measure simply combines weights by taking sums.[25]

Assuming that an appeal to naturalness justifies the use of a von Neumann–Morgenstern utility function as a cardinal measure of individual goodness, Broome then asks if these measures can be used to compare differences in good interpersonally. Recall that Harsanyi's impartial observer makes interpersonal comparisons using preferences over extended lotteries in which an outcome specifies a sure alternative together with the name of the individual the observer imagines being. Applying Broome's naturalness requirement to the choice of the observer's goodness measure for extended lotteries, it follows that differences in goodness are interpersonally comparable. For Harsanyi, the observer's preferences are the impartial preferences of an actual member of society. However, if that is the case, then, as Broome argues, there is no reason to believe that different individuals will make the same interpersonal comparisons when they adopt this impartial perspective; hence, one cannot rely exclusively on individual preferences to deliver the interpersonal comparability of well-being that utilitarianism requires.

Harsanyi would dispute this conclusion. Harsanyi regarded everyone as being fundamentally the same (what he called the similarity postulate), so that an individual's well-being can be thought of as being a function, common to everyone, of the alternative obtained and the objective causal variables determining this person's characteristics (including his preferences). As a consequence, everyone would reach the same conclusion about how well off anyone is with a particular alternative by imaging how well off he would be with this person's alternative and causal variables. Furthermore, these estimates of individual well-being can be used to compare the well-beings of individuals with different values for the causal variables.[26]

[25] This justification for how to measure individual goodness is similar to the the one offered by von Neumann and Morgenstern (1944) for their use of a von Neumann-Morgenstern utility function to represent preferences. However, von Neumann and Morgenstern were interested in explaining choices in uncertain circumstances, not in using their measure of utility in a normative theory. As Arrow (1951, p. 10) has pointed out, a von Neumann–Morgenstern utility representation does not have "any particular ethical significance." Weymark (2005) has argued that "naturalness" is an appropriate criterion to use when choosing a utility representation for descriptive purposes, but not for use in a normative theory. It is noteworthy that Broome (1991) offers a different preference-based justification for the use of von Neumann-Morgenstern representations than he does here. In his monograph, Broome supplements the preference (betterness) relation on lotteries with a second preference relation that directly compares strength of preference. See also Risse (2002).

[26] This description of the logical foundations of interpersonal utility comparisons was first proposed in Harsanyi (1955) and was further developed in Harsanyi (1977c, pp. 57–61) and Harsanyi (1992, section 5).

Broome, briefly in this volume and in more detail in Broome (1993), considers Harsanyi's causal argument and finds it wanting. Broome believes that Harsanyi's argument mistakenly treats causes of preference as objects of preference. According to Broome, each specification of the causal variables determines a different preference over the objects of preference, and nothing in Harsanyi's argument provides a basis on which to compare preferences that correspond to different values for the causal variables.[27]

Broome also argues that when Harsanyi applies his ideas to a concrete example in which, through a process of empathetic identification, an individual compares the situation of someone else (with that person's tastes and values) with that of his own, this individual's extended preferences are based, in part, on an estimate of the benefits of being in the other person's situation. But, if this is the case, nonpreferencist elements (the benefit estimates) have been used to help determine extended preferences and, therefore, Harsanyi has a view about what constitutes individual good that is nonpreferencist. Broome thinks that this is inevitable and, hence, there cannot be a completely preference-based utilitarianism. This view is supported by Harsanyi's own account of the nature of interpersonal utility comparisons in his later writings, notably in Harsanyi (1992, section 5), where utilities are explicitly interpreted as measuring amounts of satisfaction.

1.3.2 Sugden on a Common Currency of Advantage

The goodness of a situation for an individual may be broadly construed to include all factors that make it worthwhile for this person, including both the outcome achieved and the opportunities available to him. In Chapter 10, Robert Sugden calls such a conception of goodness "advantage." He asks if there is some quantitative measure of advantage – a currency of advantage – that can be used to measure an individual's well-being in the same sense that a utility function represents preferences, and, if so, whether there is

[27] Broome (1993) also argues that the concept of a fundamental preference proposed by Kolm (1972) is subject to the same shortcomings that he attributes to Harsanyi's causal argument. Kolm (1972, pp. 79–80) says that everyone has the same fundamental preferences if the variables that distinguish individuals are treated as objects of preference, that is, if these variables are added to the list of arguments of the utility function used to represent the preferences. Rawls (1982, p. 179), after discussing at length this concept of fundamental preference, what he calls "a shared highest-order preference," rejects it on the grounds that "the notion of a shared highest-order preference function is plainly incompatible with the conception of a well-ordered society in justice as fairness. For in the circumstances of justice citizens' conceptions of the good are not only said to be opposed but to be incommensurable."

a common currency of advantage that can be used to make interpersonal comparisons of well-being. In particular, he asks a variant of the question posed by Broome in this volume: Can preference satisfaction provide such an intrapersonal currency, and, if so, is there some interpersonal common currency of preference satisfaction that can be used to provide a normative evaluation of different social situations?

Sugden considers whether Harsanyi's concept of extended preferences provides a common currency in which to compare different individuals' situations. For reasons similar to those advanced by Broome, Sugden thinks not. If they are to serve this purpose, then everyone must have the same extended preferences, which, as we have seen, Harsanyi believes follows from his similarity principle and his causal argument. However, even assuming that Harsanyi is correct that two individuals will make the same estimates of the benefits of (or the psychological reaction to) being in the situations of, say, Bob and John complete with their tastes and values, Sugden argues that they may nevertheless make different choices about which of these two situations they would choose. Therefore, their extended preferences need not coincide.[28]

As Sugden notes, Rawls (1971, p. 174) does not believe that it is possible for everyone to have the same extended preferences because we cannot "evaluate another person's total circumstances, his objective position plus his character and system of ends, without any reference to the details of our own conception of the good." There is no disembodied perspective from which to make such evaluations. Rawls believes that it is a basic fact about society that individuals will not agree on a common conception of the good. Therefore, in Sugden's terminology, Rawls uses his index of primary goods as a common currency of advantage because these are goods that rational individuals would want, regardless of their conceptions of a good life. It is this index that Rawls uses to determine whether everyone has a fair share of resources and opportunities. As we have seen, for Rawls, how individuals make use of these primary goods is their own responsibility and not a matter of justice.

However, to construct an index of primary goods, it is necessary to find some way of aggregating the quantities of each of the primary goods (assuming that the holdings of each primary good is quantifiable) into a

[28] Sugden also considers a second way in which common extended preferences might arise based on what he calls the "common preference principle." However, he does not see how this principle can be justified except by positing the preexistence of a common currency of advantage, in which case extended preferences cannot be this currency. See also Broome (1993, section 6).

number. There is widespread skepticism about the possibility of constructing a nonperfectionist index of primary goods that is not dependent on individual utility functions. Arrow (1973), for example, expressed skepticism about the possibility of determining who the worst off are in terms of primary goods, when, among the disadvantaged, some individuals have more of some primary goods and less of others. See also Gibbard (1979). More recently, Arneson (1990) has argued that Rawls's theory faces a dilemma because either the same index must be used for everyone, thereby imposing a perfectionist weighting of the various primary goods regardless of individuals' preferences, or an individualized index must be adopted. However, according to Arneson, if the latter option is endorsed, then Rawls's insistence on the incommensurability of individual conceptions of the good life becomes untenable if his maximin principle is adopted.

Sugden proposes an alternative way of identifying a common currency of economic opportunities that is preference dependent in a way that treats individuals with the same objective circumstances in an impersonal way.[29] An opportunity set for an individual specifies the range of opportunities for consuming private and public goods available to him. Beginning with the seminal work of Jones and Sugden (1982) and Pattanaik and Xu (1990), a variety of indices have been proposed for measuring the extent to which an opportunity set provides freedom of choice. Using this literature as his inspiration, Sugden's objective is to construct a measure of the value of an opportunity set that can serve as a common currency of advantage. Once such a metric has been obtained, it can then be used to determine a fair distribution of opportunities.

The basic features of Sugden's construction are most easily seen in the special case in which there are only private goods. One way to construct a utility function (provided that all goods are desirable) for an individual is to use what is known as a money metric representation of preferences. A money metric utility function is defined by first specifying a set of reference prices for the goods and then determining, for each commodity bundle x, the least-cost way of obtaining a commodity bundle indifferent to x. This cost can serve as the utility assigned to x. The money metric of an opportunity set for this individual is then the utility assigned to his most-preferred commodity bundle in his opportunity set. In effect, this procedure identifies an opportunity set that everyone who has the same value m of the

[29] As in Arneson's chapter, Sugden wants to allow for individuals with different objective circumstances (the factors outside their control) to be treated differently, while simultaneously making individuals responsible for their own choices.

money metric regards as being equally desirable as his actual opportunity set, namely, the set of all commodity bundles that cost no more than *m* using the reference prices.

This particular procedure for valuing opportunity sets is not impersonal as individuals with different preferences will, in general, have different money metrics for the same opportunity set. To arrive at a measure that treats individuals with the same objective circumstances impersonally, Sugden advocates averaging the values of these money metrics over the the actual distribution of preferences of those individuals who share the same objective characteristics. The justification for doing so uses a veil of ignorance argument in which the actual distribution of preferences in society is known. To complete his construction of a common currency of advantage, Sugden needs to determine which reference prices should be used to value opportunity sets. He suggests using current market prices. In a way somewhat reminiscent of Broome's rationale for using Von Neumann–Morgenstern representations of preferences, Sugden argues that this choice is appealing because of its naturalness.

1.3.3 Fleurbaey and Maniquet on Fair Social Orderings

One way to determine the relative goodness of social alternatives is to use a social welfare function. Recall that such a function ranks different distributions of utilities independently of how these utilities are generated. Given a profile of utility functions, for each pair of alternatives, a social ordering of the alternatives is obtained by first determining what distributions of utilities they generate and then seeing how the social welfare function ranks these distributions. In this way, a social welfare function is used to construct a social welfare functional, that is, a function that assigns a social ordering of the alternatives to each profile of utility functions of interest. Ultimately, however, a social decision must be made from whatever subset of the alternatives that turns out to be feasible. This can be done by identifying which of these alternatives are best according to the social ordering of all of the alternatives (feasible or not), that is, by maximizing this social ordering on the feasible set. Note that best alternatives identified in this way depend both on what profile of utility functions is considered and on what set of alternatives is feasible.

If utility is ordinal and interpersonally noncomparable, as in Arrow (1951), this procedure in effect first determines a social ordering of the alternatives as a function of the individual preferences – an Arrovian social welfare function – and then chooses best alternatives from feasible sets as previously described. Thus, the social welfare function approach to ranking social alternatives includes the Arrovian approach as a special case. A somewhat less ambitious way of making social decisions is to directly identify socially best alternatives as a function of the individual preferences and feasible set without going through the intermediary of an Arrovian social welfare function. Such a function is called an allocation rule (or, alternatively, a social choice function). With this approach, no attempt is made to rank the non-best alternatives in terms of their relative social goodness.

In the preceding discussion, the alternatives can be anything – politicians running for office, job candidates, or allocations of resources, for example. Naturally, economists have focused much of their attention on problems in which the social alternatives are distributions of private and public goods. The social welfare function approach in general, and the utilitarian social welfare function approach in particular, is widely used in welfare economics and the normative part of public economics. The allocation rule approach has a number of applications, the most relevant here being in the theory of fair allocation.[30] In this theory, the socially best alternatives in a feasible set are those that satisfy some criterion of fairness as well as some other desirable properties. For a comprehensive introduction to the literature on fair allocation rules, see Thomson (2005).

The constrast between the social welfare function approach of traditional welfare economics and the theory of fair allocation echoes the opposition between Harsanyi and Rawls with their focus on utility consequences and fair social institutions, respectively. In Chapter 11, Marc Fleurbaey and François Maniquet discuss the pros and cons of these two ways of making social decisions and offer an alternative of their own for making decisions about the distribution of economic resources. To illustrate their proposal, they consider the problem of allocating fixed quantities of a number of private goods (the social endowment) among a group of individuals who only care about their own personal consumptions, but their ideas have much wider applicability. In such division problems, the combination of a profile of preferences and a set of feasible alternatives is called an economy.

In their view, the social welfare function approach has the advantage that it provides a way to compare any pair of social alternatives on the basis of

[30] Allocation rules are also used in implementation theory. In order to use an allocation rule, it is necessary to learn what the individual peferences are. However, if individuals know how an allocation rule works, they may have an incentive to misrepresent their preferences so as to achieve an outcome that is better for them. In implementation theory, the objective is to design incentive mechanisms for which such manipulation is not possible.

individual utilities. Unfortunately, in the Arrovian version of this approach, we run up against Arrow's impossibility theorem – only dictatorial social welfare functions are compatible with Arrow's desiderata for an acceptable Arrovian social welfare function (see Arrow, 1951). To avoid this concentration of decision-making power, as described earlier, the social rankings are required to take into account information about the interpersonal comparability of utility. However, little is said about how one could actually perform these comparisons. In economics, the ordinalist approach is usually justified in pragmatic terms – only preferences are revealed by choices and nonchoice information about utilities is unreliable because individuals may not have an incentive to convey this information truthfully. Fleurbaey and Maniquet note that Rawls's view that individual conceptions of the good life are incommensurable provides an additional argument in favor of ordinalism. The theory of fair allocation has the advantage that it only identifies fair allocations on the basis of information about resource holdings and individual ordinal preferences about personal consumption. However, it is unable to compare the relative desirability of unfair allocations.

Fleurbaey and Maniquet introduce the concept of a social ordering function, which is a function that determines a social ordering of the feasible alternatives for each economy. That is, for each possible social endowment of goods, a social ranking of the possible distributions of this endowment is determined as a function of the individual preferences and of what is feasible. In contrast, with an Arrovian social welfare function, all alternatives (feasible or not) are socially ranked and these rankings are independent of what is feasible. With a social ordering function, unlike an allocation rule, all feasible alternatives (i.e., distributions of the social endowment) are ranked. The ability to rank all of these alternatives may be useful if only a subset of the distributions of the social endowment are attainable because of incentive or political constraints. Because a social ordering function does not employ information about utilities other than what is contained in individual preferences over personal consumption, it shares the ordinalism and interpersonal noncomparability of Arrovian social choice theory.

In view of the nihilism of Arrow's impossibility theorem, one of Fleurbaey and Maniquet's main objectives is to show that social orderings that incorporate fairness norms and other appealing properties exist in spite of the ordinalism of their approach to social decision making.[31] They do this in two different ways: (i) by directly constructing social welfare orderings whose

[31] There has been considerable skepticism in welfare economics and social choice theory about the possibility of achieving this objective. See Fleurbaey and Mongin (2005).

socially best alternatives in each economy yield well-known allocation rules found in the fair allocation literature and (ii) axiomatically.

To illustrate the first approach, we use Fleurbaey and Maniquet's example of the Pazner–Schmeidler egalitarian-equivalent allocation rule (see Pazner and Schmeidler, 1978). For each economy, this rule selects from among the Pareto optimal distributions those for which there is an individual commodity bundle that is proportional to the social endowment and that everyone regards as indifferent in preference to their assignment by the allocation rule.[32] This universal indifference to a common consumption bundle is the fairness norm embodied in this rule. The social ordering function that Fleurbaey and Maniquet use to generate this allocation rule using the maximization procedure previously described has the property that the economy-dependent social orderings of alternatives take the form of maximin criteria applied to particular measures of individual bundles of resources. As a consequence, Rawls's intuition that a distributive criterion such as the maximin principle can be applied to some index of resource holdings without resorting to interpersonally comparable information about individual utility appears vindicated. At the same time, these social orderings satisfy the Pareto criterion and respect individual preferences, so that, *pace* Arneson (1990), perfectionism is avoided as well. Fleurbaey and Maniquet's other examples of fair social ordering functions employ different fairness norms, but they all require that there be some commodity bundle or individual opportunity set that everyone regards as being indifferent to what they are allocated. In this respect, these fairness norms have much in common with the way that Sugden constructs a money metric measure of an opportunity set.

As in Arrow (1951), the axiomatic approach lists a number of a priori properties that one would like the social decision-making procedure to satisfy (e.g., it should respect the Pareto principle). These properties are the axioms and, in this context, they are typically normative criteria that formalize various ethical principles. Fleurbaey and Maniquet argue that the appeal of these axioms is context dependent, which the abstractness of Arrow's framework fails to consider. In particular, they argue that Arrow's Independence of Irrelevant Alternatives axiom loses much of its appeal when economic problems are considered. They suggest alternative independence axioms that, like Arrow's axiom, limit the information about preferences that can be used when ranking a pair of alternatives, but are motivated by the structure of the economic problems they consider. Furthermore,

[32] An allocation is Pareto optimal if there is no other feasible allocation that everyone prefers to it.

using the social ordering functions that they construct to generate the fair allocation rules they consider, Fleurbaey and Maniquet show that each of their independence axioms is consistent with a number of basic normative criteria, including the requirement that individuals be treated evenhandedly.

1.3.4 Barry on Want Satisfaction

Utilitarianism and many other teleological theories agree that, in general, it is a good thing to satisfy the wants of individuals (even if want satisfaction is not synonymous with what is good). However, if wants are subject to change, the question becomes, What wants? In Chapter 12, Brian Barry defends the thesis that it is individuals' actual wants. He argues that the goodness of want satisfaction is not subject to objections raised by Elster (1982) and Rawls (1982). According to Elster, individuals tend to limit their wants to what is achievable, and such wants are an inappropriate basis for a utilitarian to evaluate different social states. Rawls, however, has argued that want satisfaction implies that everyone should modify their wants so that they are easily satisfied, which denies that individuals are autonomous beings with their own determinate conceptions of the good.[33] Both options facilitate want satisfaction, but if one does not regard them as being equally satisfactory ways of promoting the good, then doubt is cast on the validity of the thesis that unqualified want satisfaction is a good thing.

Barry denies that Rawls's conclusions follow from his premise. Specifically, he denies that a utilitarian is committed to regarding it to be an improvement if an individual deliberately changes his tastes and aspirations so that his wants are more easily satisfied. To accept that such a change is desirable undermines the concept of what it means to be a person. Thus, Barry agrees with Rawls that a fundamental characteristic of individuals is that their preferences cannot be completely malleable, but he sees no conflict between this observation and a utilitarian concern with want satisfaction. This is not to say that a utilitarian (or a nonutilitarian for that matter) cannot say that it is better if individuals had different wants because what constitutes the good may be more extensive than actual want satisfaction.

An implication of the limited form of the want-satisfaction thesis that Barry defends is that it can only provide a partial ranking of states of affairs. It can be used to justify a want-satisfaction version of the Pareto principle – given the wants people actually have, it is better if some individuals' wants

[33] Recall from our discussion of Sugden's chapter that Rawls believes it is incoherent to think of a disembodied person who has no conception of the good that helps to define him.

are more fully satisfied provided that nobody else has his wants satisfied less. Even if it is possible to compare quantities of want satisfaction before and after a change in tastes, it is illegitimate to extend the Pareto criterion to such situations because an increase in someone's satisfaction might have occurred simply because his tastes are now more easily satisfied.

Elster frames his discussion of adaptive preference formation in terms of the parable of the fox who declares some grapes to be sour and, thus, undesirable when he realizes that they are unattainable. Elster argues that a utilitarian should not assess the relative goodness of social states based on preferences formed in this way. Barry considers many specifications of Elster's description of the fox and the grapes parable, as well as some alternative ways in which tastes might change. In each case, he concludes that the wants that should be promoted are the wants based on actual preferences, even if these preferences were shaped by limiting what is wanted to what is feasible.

In his discussion of the fox and the grapes parable, Elster focuses on whether there would be a welfare loss if the fox did not receive the grapes. He does not explicitly consider the implications of his argument for questions of distributive justice. Barry argues that an implication of the concerns raised by Elster is that if an individual's low expectations result in him not wanting much, then any theory of distributive justice that takes account of want satisfaction will sacrifice his interests at the expense of those whose wants are not so easily satisfied. The objection that individuals who are efficient at producing good from resources will be favored by a good-maximizing theory such as utilitarianism is a serious one. However, according to Barry, this objection provides a reason to adopt a theory of justice that focuses on the design of institutions that generate the allocation of resources and not on the use that is made of them.

1.4 Sharing the Gains from Social Cooperation

A view widely held by social contractarians is that society is a cooperative enterprise and that the objective of a social contract is to provide a framework that facilitates the realization of the potential benefits from social cooperation. In liberal versions of social contract theory, individuals are free to pursue their own goals as they wish, within certain bounds. Gains from cooperation can be shared in many ways, so a liberal social contract theory needs to address the issue of how individuals are to coordinate their actions on one of the many possible ways of reaping the benefits of cooperation. Furthermore, to the extent that the coercive power of government is used to help shape the outcome of these individual decisions, for example, by instantiating principles of justice into the design of the basic institutions of society, then it is also necessary to investigate why individuals should voluntarily agree to be governed by these principles. The contributions to Part 4 address these issues of coordination and compliance.

1.4.1 Naturalistic versus Normative Theories

The standard descriptions of Rawls's liberal egalitarianism regard his principles as constituting normative ideals that can serve as guiding principles for the design of basic social institutions or, at least, as principles that can inform public debates relating to such institutions. Similarly, Harsanyi's utilitarian principles are meant to serve as normative guidelines for individual or collective action. The hope of the normative theorist is that by contributing to the debate about fundamental issues of society, his ideas will help convince citizens and decision makers of the rightness of some principles of morality or justice and thereby influence the actual decisions that are made by individuals and society.

Rawls explicitly assumes that the principles identified from behind a veil of ignorance will be complied with once the veil is lifted. For example, Rawls (1971, p. 245) says that "strict compliance is one of the stipulations of the original position; the principles of justice are chosen on the supposition that they will generally be complied with." Rawls justifies this assumption by describing his theory of justice as a contribution to ideal theory, that is, a theory for a well-ordered society in which everyone is assumed to comply with and support the principles of justice. This is not to say that Rawls ignores problems of noncompliance. On the contrary, parties behind the veil are to use their general knowledge of human psychology to ensure that the principles that are adopted are ones that are self-supporting in a well-ordered society. Nevertheless, while Rawls (1971, p. 303) argues that suitably modified versions of his principles of justice have relevance for some situations in which strict compliance does not hold, he acknowledges that if we depart sufficiently from his ideal case, it may be necessary to abandon his principles altogether.

Other liberal egalitarian theories in the contractarian tradition can also be described as contributions to ideal theory. For example, Barry (1995) offers an account of principles of justice based on reasonable agreement – what he calls justice as impartiality – that is based on an assumed desire on the part of all individuals "to live in a society whose members all freely accept its rules of justice and its major institutions" (Barry, 1995, p. 164).

We live in a nonideal world, and for this reason, many scholars regard normative theories developed for ideal circumstances as fundamentally flawed or at least lacking compelling foundations. For such skeptics, normative social contract theories designed for ideal worlds leave many questions unanswered: What social arrangements are viable and compatible with the motivations actually held by individuals? How do their motivations evolve? How are moral and social norms actually generated? These questions invite a naturalistic response. Suppose that norms of behavior evolve over time through ordinary experience and owe little to reasoning and theoretical debates, so that the set of feasible social arrangements is narrowly determined by human evolution, leaving little scope for conscious action.[34] From this perspective, the normative theorist has little influence over the course of history. For someone who subscribes to this view, theoretical justifications of ethical or political principles are unpersuasive. Rather, to the extent that individuals are observed to behave morally or justly, the explanation lies in the evolutionary processes that shaped who we are.

1.4.2 Game Theory

The contributions to Part 4 make extensive use of game theory. Before turning to these chapters, it is useful to summarize the main features of the relevant game theory. A game can be thought of as a stategic situation in which the outcome depends on the joint choice of actions by a set of individuals called players. In traditional game theory, these actions are chosen by rational individuals who pursue their own goals. A basic distinction needs to be made between cooperative and noncooperative games. In a cooperative game, any agreements reached by the players are binding and hence must be complied with if sanctions are to be avoided. In contrast, in a noncooperative game, no binding agreements are possible.

In a noncooperative game, the players are viewed as independently choosing strategies. If there is only one time period, a stategy for a player consists of choosing a single action, and this is done simultaneously by all of the players. If the game involves decision making over time, a strategy specifies what action a player chooses in each of the situations in which this person might be called on to act. These stategies can be thought of as being conditional plans announced simultaneously at the beginning of the game by each player. With a pure strategy, all decisions are deterministic. With a mixed

[34] Although Hayek (1960, p. 24) did not adopt such an extreme view, he cautions that reason operates within bounds and that “it is the state of civilization at any given moment that determines the scope and possibilities of human ends and values.”

strategy, a player plays according to a probability distribution over his pure stategies. When there are a finite number of pure strategies, a mixed strategy can be written as a vector x whose kth component is the probability that the kth pure strategy is chosen. A pure strategy is a mixed strategy in which all of the probability is put on it. When thinking of a pure strategy as a degenerate mixed strategy, it is customary to write it as the vector e^k that has a 1 in the kth component and a 0 elsewhere. A strategy profile is a list describing what strategy each player chooses. The payoff function of a player specifies the expected value of his von Neumann–Morgenstern utility as a function of the strategy profile. Thus, payoffs are measured ex ante before the random devices that players use to implement their mixed strategies identify which pure strategies are played ex post.

In noncooperative game theory, it is assumed that players rationally pursue their own ends. Both this motivational assumption and the structure of the game are common knowlege. As a consequence, simply reasoning by himself, a player can determine what strategy or strategies maximize his expected utility for any possible combination of the strategies for the other players. Such a maximizing strategy is called a best reply. A strategy profile is a Nash (1951) equilibrium if each player's choice of strategy is a best reply to the strategies of the other players. If the game involves decision-making over time, playing according to a Nash equilibrium may not be credible, as a player may find it profitable to deviate from his announced strategy as time progresses. A Nash equilibrium is subgame perfect if nobody has an incentive to deviate from his announced plan whenever it is his turn to make a decision.

The only cooperative theory that we consider is the Nash (1950) theory of two-person cooperative bargaining. The Nash theory is welfarist, so all agreements can be described in terms of their utility consequences. A bargaining problem, then, can be characterized by (i) the set of vectors of individual utilities that can be achieved by some feasible agreement and (ii) a disagreement point, which is the utility vector obtained if the bargaining is not resolved. A bargaining solution specifies the utility vector that is agreed to in each bargaining problem. Any such choice belongs to the bargaining set, which consists of the utility vectors that are both Pareto optimal and that make no one worse off than with the disagreement point.

The most prominent solution is the (symmetric) Nash (1950) bargaining solution, which selects the utility vector in the bargaining set that maximizes the product of the individual utility gains with respect to the disagreement point. The Nash solution can be supported both by normative arguments, as in Nash (1950), and by the fact that it is the subgame perfect Nash equilibrium outcome in a natural model of noncooperative bargaining in which individuals with equal bargaining power make successive offers.[35] If the bargaining set is asymmetric in favor of some individual, say, because this person is better able to convert physical outcomes into utility, then the Nash bargaining solution will typically allocate a greater utility gain to him.

An alternative bargaining solution has been proposed by Kalai and Smorodinsky (1975). In the two-player case, this solution is determined as follows. First, for each player i, the difference Δ_i between the utility achieved in his best outcome in the bargaining set and the utility at the disagreement point is calculated. Then, the utility vector (u_1, u_2) in the bargaining set is chosen for which the ratio of the utility gain of individual 1 to the utility gain of individual 2 is equal to the ratio of the maximum possible gains Δ_1/Δ_2. The minimax relative concession solution used by Gauthier (1986) in his bargaining approach to social contract theory is closely related to, and inspired by, the Kalai–Smorodinsky solution.

Game theory is often criticized for assuming too much rationality and computing ability on the part of the players. An extreme way of modeling bounded rationality is to follow the lead of evolutionary game theory by assuming that a player is a pure automaton that employs a fixed strategy.[36] Evolutionary game theory is concerned with identifying successful strategies. Success can be measured in two ways. First, successful strategies are ones that can withstand the introduction of "mutant" strategies. Second, they are strategies that are favored by a selection mechanism that determines which strategies reproduce over time in successive generations of players.

The concept of an evolutionary stable strategy, introduced by Maynard Smith and Price (1973), focuses on the first of these forms of success. There are two equivalent ways of defining an evolutionary stable strategy. Suppose that two players are repeatedly drawn at random and with equal probability from a large population to play a symmetric noncooperative game in which the set of possible mixed strategies is S. The expected utility of a player who uses strategy x when faced with a player who plays strategy y is $u(x, y)$.[37] Consider a situation in which the players have been led by evolutionary forces to use the same incumbent strategy x. Now imagine that there is a mutation and the mutants play the strategy y. If the fraction of the population that are

[35] For a discussion of the noncooperative foundations of the Nash bargaining solution, see Binmore (1998, section 1.7).

[36] Evolutionary game theory has its origins in the work of Maynard Smith and Price (1973). For a good introduction to evolutionary game theory, see Weibull (1995).

[37] In a nonsymmetric game, this utility could depend on the identity of who plays x and who plays y.

mutants is ε, then any player will meet a mutant with this probability. The incumbent strategy x is evolutionary stable if for any mutant strategy $y \in S$, there is a value of ε such that playing x yields a higher expected payoff than playing y when the other player is chosen randomly from the postmutant population, as previously described, and the fraction of mutants does not exceed ε. In other words, x cannot be upset if a small proportion of the population mutates.

Equivalently, x is an evolutionary stable strategy if (i) (x, x) is a Nash equilibrium in the symmetric two-person game with strategy set S and (ii) if a strategy $y \neq x$ gives the same payoff as x when played against x (i.e., $u(x, x) = u(y, x)$), then x gives a higher payoff when played against y than playing y (i.e., $u(x, y) > u(y, y)$). This latter characterization of evolutionary stable strategies regards individuals as actively choosing their strategies and makes no explicit reference to evolutionary phenomenon. The establishment of a connection between stable states of evolution with players modeled as automata and standard concepts of strategic equilibrium for rational players (such as the Nash equilibrium) has been one of the most significant achievements of evolutionary game theory (see, e.g., Young, 1998).

The replicator dynamics provides one way of modeling the process by which strategies are selected over time. Suppose that, at any given time, two players are chosen with equal probability from a large population to play a symmetric game. However, now they are programmed to play one of a finite number of pure strategies in the set X. The vector x whose kth component x_k is the proportion of the population that plays the kth of these pure strategies can also be interpreted as a mixed strategy drawn from the set of mixed strategies S on X. Because of the linearity of von Neumann–Morgenstern utility functions in the probabilities, the expected utility $u(e^k, x)$ of a player who plays the kth pure strategy is the same whether he meets someone at random from a population whose distribution over pure strategies is given by x or whether he meets a single individual for sure who has the mixed strategy x. The average payoff in the population described by x is simply $u(x, x)$, that is, the expected payoff of someone playing mixed strategy x when the other player uses the same strategy.

In the discrete time version of the replicator dynamics proposed by Taylor and Jonker (1978), a player programmed to play pure strategy k reproduces itself in proportion to the ratio of his expected payoff to the average payoff of the whole population, that is, in proportion to the ratio $u(e^k, x)/u(x, x)$. In the next generation, the distribution of strategies is the vector y for which $y_k = x_k u(e^k, x)/u(x, x)$. Thus, strategies with better than average payoffs proliferate and strategies with less than the average payoff decline. Note that strategies reproduce themselves without error, so there are no mutations in this process. A distribution of pure strategies is stationary if it simply replicates itself. With the replicator dynamics, stationarity is achieved if all pure strategies that are played with positive probability yield the same expected payoff.

One can also consider the stability of the selection process with respect to mutations. A selection process is asymptotically stable at the distribution of pure strategies x if after a small perturbation in this distribution due to mutations, the dynamic process moves the distribution back toward x. Every evolutionary stable strategy is stable in this sense for the replicator dynamics. A distribution is an attractor of a selection process if the dynamic process converges to this distribution from some initial conditions. A stationary distribution is an attractor. The set of distributions that converge to an attractor is called its basin of attraction.

1.4.3 Binmore on Natural Justice

In Chapter 13, Ken Binmore first considers Harsanyi's social aggregation theorem, what he calls Harsanyi's teleological defense of utilitarianism. He justifies the use of von Neumann–Morgenstern representations of preferences in this theorem, and hence its utilitarian interpretation, by supposing that individuals can make strength of preference comparisons, not just rank lotteries, as in Broome (1991). Harsanyi claims that individuals have a moral obligation to pursue the common good as expressed by the social preference, which Binmore regards as simply begging the question about why they have this obligation. It would be better, Binmore argues, to regard Harsanyi's theorem as providing guidance about what actions a benevolent government should require of its citizens.

The main part of Binmore's chapter offers a synopsis of some of the central features of his social contract theory, as expounded at greater length in Binmore (1994, 1998, 2005). Binmore, drawing inspiration from David Hume's writings on the origins of social conventions, offers a naturalistic account of the evolution of fairness norms, as embodied in an original position, that can serve as a coordinating device for determining how to share the gains from social cooperation.[38] In Binmore's view, theories that

[38] Binmore's discussion of the use of fairness norms as coordinating devices bears some resemblance to the role Appiah (2006, p. 28) attributes to evaluative language when he says: "Our language of values is one of the central ways we coordinate our lives with one another. We appeal to values when we are trying to get things done *together*."

make unrealistic motivational demands are simply utopian. Not wishing to engage in utopian thinking, Binmore requires any social contract agreed to behind a veil of ignorance to be self-enforcing when the veil is lifted. As we shall see, Binmore reaches very Rawlsian conclusions using a description of the original position that has much more in common with Harsanyi than with Rawls.

In Harsanyi's version of the veil of ignorance, a single person, the impartial observer, is choosing behind the veil. In contrast, with Binmore, the decision making behind the veil is modeled as a cooperative bargaining problem. Specifically, Binmore extends the Nash (1950) theory of two-person cooperative bargaining so that it can be applied behind a veil of ignorance. Outside the veil, the feasible set of alternatives and the disagreement alternative are also modeled as vectors of utilities, as in a cooperative bargaining problem. Binmore refers to the agreements that generate these utilities as social contracts. Even though each of the individuals knows exactly how different agreements benefit him, he employs the device of an original position, thereby agreeing to negotiate as if ignorant of his true identity because it is a useful device for working out the implications of a shared norm to treat one another fairly. In other words, hypothetical bargaining behind the veil of ignorance commands our attention not because of abstract metaphysical arguments, but simply because it expresses very well our ingrained do-as-you-would-be-done-by principles of fairness.

For simplicity, as in much of bargaining theory, Binmore supposes that there are only two individuals, who he calls Adam and Eve. Behind the veil, these two decision makers are players I and II, respectively. In the hypothetical identity lottery that these players imagine themselves facing, there are two equally likely outcomes once the veil is lifted, player I is Adam and player II is Eve, or vice versa, denoted AE and EA, respectively. Although the true state is AE, behind the veil, the players do not know that this is the case. Unlike Harsanyi, what the players choose to do once the veil is removed is allowed to be contingent on whether the true state is AE or EA.[39] As with Harsanyi's impartial observer, the players behind the veil need to imagine what it is like to be either Adam or Eve with each of the possible social contracts that they could agree to. For the reasons given in his discussion of Harsanyi's social aggregation theorem, Binmore assumes that both the empathetic preferences (the analogues of Harsanyi's extended preferences) behind the veil and the actual preferences over social contracts outside the

[39] In Harsanyi's impartial observer theorem, the same lottery over the set of social alternatives is chosen regardless of who the observer turns out to be.

veil are represented by von Neumann–Morgenstern utility functions. Further, these empathetic preferences satisfy Harsanyi's Principle of Acceptance. The veil is assumed to be thin enough that Adam and Eve in their roles as players I and II behind the veil use their actual empathetic preferences, although they do not know who they in fact are. In principle, the two individuals need not agree on how to make interpersonal utility comparisons, and so each may use a different scaling factor to convert a unit of utility for Adam into an equivalent amount of utility for Eve.

To facilitate the comparison of his analysis with that of Harsanyi, Binmore initially assumes that there is a government that can enforce agreements reached behind the veil. Using their own ways of commensurating utilities, the players behind the veil can determine the feasible set of utility vectors and the disagreement utilities (all of which are expected values of the utilities obtained in each of the two states AE and EA outside the veil) for the bargaining problem that they are faced with. Applying the Nash bargaining solution to this problem not only determines the expected utilities agreed to behind the veil, but also the actual utilities in each of the two states that the players agree to implement outside the veil.

If it happens that Adam and Eve agree on how to make interpersonal utility comparisons, then it turns out that the same social contract is adopted in each state, and this contract is what would be obtained by maximizing a weighted utilitarian objective function on the bargaining set in either state AE and EA. Furthermore, the relative weights in the objective function are given by the scaling factor used to convert Adam's utility into utility for Eve. In effect, in this special case, Binmore's bargaining approach to the veil of ignorance provides support for Harsanyi's impartial observer defense of utilitarianism.

At this point, Binmore introduces the additional assumption that there is no external enforcement mechanism that compels the individuals to implement a particular agreement once the veil is lifted. More precisely, they retain the right to call up a new round of bargaining *behind the veil of ignorance.* This option is attractive to the individual with the smaller utility in the realized state because, by invoking the veil, he places himself once more in a situation in which he has an equal chance of obtaining the higher realized utility. The only situation in which the worst off has no incentive to appeal to the veil of ignorance is when the two utilities are equal. The requirement that agreements be self-enforcing outside the veil constrains what agreements are possible behind the veil in such a way that the Nash bargaining solution results in the same agreement that would have been chosen had the players used the maximin principle to make their choice instead. Thus, by explicitly taking the compliance issue into account, rather than assuming that individuals have a duty to comply with the results of their hypothetical bargaining, Binmore has provided some justification for Rawls's use of the maximin principle (applied to utilities, not an index of primary goods) without appealing to Rawls's assumption that players behind the veil of ignorance employ maximin reasoning, which both Binmore and Harsanyi regard as a rather dubious motivational assumption.

In the special case in which Adam and Eve agree on how to make interpersonal utility comparisons, the agreement reached when the compliance constraint is taken into account is also the agreement that a utilitarian would choose, not just what is chosen by Nash bargainers or Rawlsian maximiners. We therefore have the unexpected conclusion that by requiring agreements to be self-enforcing, a Rawlsian maximinner and a utilitarian agree on what social contract to adopt, even though they do so for very different reasons.

Binmore argues that biological and cultural evolutionary forces will, over time, lead individuals to make the same interpersonal utility comparisons. In his view, the origin of empathetic preferences lies in the need primitive hunter-gather societies had to empathize with one another in order to find a way to share food between those who were lucky enough to find food and those who were not. Just as successful behavioral patterns are imitated and propagated over time, empathethic preferences that benefit those who hold them are also imitated and propagated. Binmore argues that this dynamic process will eventually converge to an evolutionary stable situation in which everybody shares the same empathetic preferences.

Although it is evolutionary forces that lead to everybody sharing the same empathetic preferences, Binmore uses the second characterization of an evolutionary stable strategy to check whether the conditions for evolutionary stability hold. He considers a thought experiment in which each player behind the veil, in addition to bargaining, can announce what empathetic preferences he is employing and does so by choosing an announcement that benefits himself the most in the subsequent bargaining, given what the other player announces. As we have seen, a necessary condition for evolutionary stability of the empathetic preference formation process is that there is a Nash equilibrium in this announcement game in which both players make the same announcement.

Thus, if sufficient time is allowed for this evolutionary process to stabilize, then there is a commonly agreed to standard for converting one person's utility into the utility of the other. Furthermore, using this common conversion factor, the outcome of the bargaining behind the veil is the same as if Adam and Eve had simply used either the maximin utility or utilitarian decision-making criterion instead or had bargained directly without appealing to the original position.[40]

By providing naturalistic foundations for both the use of the original position and the way that empathetic preferences are formed, Binmore has thereby provided an account of moral behavior that eschews any normative justification in evolutionary stable situations. But this does not imply that morality has no role to play in the short run when evolutionary forces have not had time to do their work. In the short run, in response to changing circumstances, the fairness norms and standards for making interpersonal comparisons inherited from the past allow individuals to reach a new social contract reasonably smoothly by their bargaining. It is in these short-run coordination problems that individual fairness norms have normative significance. However, a short-run social contract may not be evolutionarily stable, and so a new round of adjustments to the empathetic preferences is initiated until once again the moral content of the agreements reached is eroded. We thus see that there is some role for normative considerations in Binmore's social contract theory, but it is considerably reduced in comparison with theories that Binmore rejects as being utopian.

1.4.4 Skyrms on the Evolutionary Viability of Fairness Norms

In Chapter 14, Brian Skyrms is in a sense even more radical than Binmore because he abandons all normative considerations of fairness and impartiality in order to examine the evolutionary viability of fairness norms in a purely naturalistic way. Like Binmore, Skyrms draws his inspiration from Hume. He argues that one should study how the social contract actually evolves: "Throw away the veil.... People just bargain, over and over" (p. 337). Skyrms focuses on the simplest possible problem of distributive justice: the game in which two individuals must decide how to share a windfall amount of money. Each gets his proposed share if the sum of these shares does not exceed one and nothing otherwise. Assuming that individuals are self-centered and prefer more money to less, this game has an infinite number of Nash equilibria of the form $(x, 1 - x)$, where $0 \leq x \leq 1$, that is, in which the sum of the shares demanded exactly equals one. Fairness suggests that equal sharing (i.e., $x = 1/2$) is an attractive solution in this context, and this appears to be a salient option in experiments.

[40] Recall that there is no incentive to appeal to the veil of ignorance if Adam and Eve obtain the same utilities, which would be the case with the maximin principle.

Skyrms analyzes this game from an evolutionary perspective. He first considers the case in which everyone's utility is measured by the amount of money obtained. In this symmetric game, equal sharing is the unique evolutionary stable strategy. However, with the replicator dynamics, for any x with $0 < x < 1$, there is an asymptotically stable distribution in which one subgroup of the population demands x and the rest of the population demands $1 - x$.[41] Evolutionary stability, by itself, does not help much in narrowing down what fairness norms can evolve.

Suppose that the population is split between greedy individuals who demand $x > 1/2$ and modest individuals who demand $1 - x$. As these individuals are randomly matched, a modest person always gets what he asks for regardless of whether he meets another modest type or whether he meets a greedy type. However, a greedy individual gets x if he meets a modest type and he receives nothing otherwise. As a consequence, it is not certain that all of the prize will be allocated. Only when $x = 1/2$ is the outcome always Pareto optimal ex post.

If equal sharing has a relatively large basin of attraction, then this possible inefficiency may not be very important. To investigate this issue, Skyrms supposes that the prize can only be divided into a finite number of equal-sized amounts. Using computer simulations, he then determines the attractor that the replicator dynamics converges to starting from a distribution over the possible claims that is chosen randomly, assuming that all possible initial distributions are equally likely. By repeating this procedure a large number of times, Skyrms is able to estimate how large the basin of attractions are for different stable distributions.

Note that if two individuals both claim more than an equal share, the one with the larger claim gets a higher utility if matched with a sufficiently modest player, but he also has a higher probability of getting nothing because the sum of the claimed shares exceeds one. The greedier the individual, the more the second effect tends to dominate. Also note that among individuals who claim less than an equal share, the ones who demand more tend to have higher utilities in the random matching process. This suggests that equal division has a much larger basin of attraction than the other options, and indeed Skyrms's simulations confirm this intuition. Furthermore, they show that the $(x, 1 - x)$ attractors with large basins of attraction have values of x close to $1/2$. One is then left to wonder whether the appeal of equal division comes

[41] In this distribution, the fraction of the population requesting the smallest amount is equal to the ratio of their demand to that of the other demand.

from normative fairness considerations or from an evolutionary process that has wired this solution into our brains.

In the rest of his chapter, Skyrms introduces an asymmetry by supposing that the utility functions can be one of two types and that an individual has one of these functions when he plays the role of the first player and the other function when he plays the role of the second. As in the symmetric case, demands of the form $(x, 1 - x)$ for $0 < x < 1$ are evolutionarily stable. However, his computer simulations using the replicator dynamics show that the Nash bargaining solution has the largest basin of attraction and that almost all initial distributions converge to outcomes close to it. However, Skyrms's simulations also show that if some positive correlation between the players' strategies is introduced into the matching process, then the largest basin of attraction is shifted away from the Nash solution toward the utilitarian solution. Furthermore, when types are uncorrelated in the matching process, the probability distribution over the possible $(x, 1 - x)$ attractors need not be symmetric, but can be skewed toward either the utilitarian or Kalai–Smorodinsky solutions to this division problem, depending on the specification chosen for the utility functions.

Skyrms concludes from his simulations that philosophers should devote more attention to the Nash bargaining solution than they have done. However, he is careful not to claim too much, as the variations in the basic model that he has considered do not unequivocally identify one solution as having the largest basin of attraction. Rather, more modestly, he views evolutionary analyses, such as his, as helping to identify the realistic possibilities for solving problems of distributive justice.

1.4.5 McClennen on the Use of Cooperative Dispositions as a Coordinating Device

In Chapter 15, Edward McClennen pursues and extends ideas developed in McClennen (1990) about rationality and cooperation, and, as an alternative to the narrow self-interested behavior underlying Nash equilibria, he proposes a theory of rational social interactions that includes some role for Rawlsian maximin reasoning. In this alternative account of rationality, individuals are assumed to display certain cooperative dispositions, thereby providing some role for naturalistic considerations to play in his theory. Many noncooperative games have multiple Nash equilibria and, even though there have been a number of criteria proposed in the literature for selecting among them, there is no agreed on theory that both selects a unique equilibrium in every game and explains how the players coordinate their strategy choices so that this equilibrium is achieved. Furthermore, even if there is a unique Nash equilibrium, the outcome may not be Pareto optimal. In contrast, individuals with the dispositions posited for them by McClennen coordinate on a Pareto optimal outcome, even if it is not a Nash equilibrium.

The difference between McClennen's theory of rational behavior and Nash's theory can be illustrated using a game of pure coordination of the kind analyzed by Schelling (1960). In such a game, the players have common interests (i.e., they rank the outcomes associated with each strategy profile in the same way), but there are multiple Nash equilibria and they are not able to communicate with one another to coordinate their strategy choices. Consider, for example, a situation in which Adam and Eve can meet at either location x or y. They both prefer to meet at location x rather than at location y, but above all they prefer to meet rather than to miss each other. This scenario is a coordination game that has two pure-strategy Nash equilibria (both choose x and both choose y) and one mixed-strategy equilibrium that involves randomizing over which of the two locations to go to. If Adam thinks that Eve will go to x, he will go there as well, but he may be mistaken, in which case they will fail to meet. In this game, it is hard to see how the Nash theory can help us predict what Adam and Eve would do. However, it is natural to suppose that Adam and Eve will coordinate on x even though they are unable to communicate with each other because meeting at x is the unique Pareto optimal outcome. In the terminology of Schelling (1960), both choosing x is a focal point.[42]

In the preceding discussion, we viewed the Pareto criterion as a focusing device for selecting among Nash equilibria in a coordination game. McClennen argues that one should instead use the Pareto criterion to provide an alternative account of rational behavior in noncooperative games. Specifically, he suggests that individuals have a disposition to choose a strategy that can be combined with strategies of the other individuals to produce an outcome that is both Pareto optimal and Pareto superior to what they

[42] If Adam and Eve think that meeting at either location is equally good, then this Pareto-based selection criteria does not help predict their behavior. Nevertheless, casual observation and experiments suggest that coordination occurs quite frequently in such games. Schelling has offered a reason for this. He has argued that one should look for clues in the real-world game being played, not the abstract formulation considered here, to see how this coordination problem is solved. For example, conventions that exist in one's society might single out one of the equilibria as a focal point, such as meeting an arriving passenger at the baggage claim in an airport, rather than at the check-in counter.

expect would be the outcome had these individuals simply pursued their own interests, as in the Nash theory.[43] In McClennen's view, the problem with the Nash theory is that it is based on behavior that may be appropriate for a single individual deciding what to do in the face of uncertainty generated by nature, but not for individuals who want to coordinate their actions to exploit fully the gains from cooperation. As he notes, in computing a best reply, it makes no difference whether an individual thinks that what he can obtain is being constrained by what others choose or whether his choice has been constrained by nature. In effect, in the Nash theory, an individual views himself as someone pursuing his own myopic self-interest, not as someone engaged in a cooperative enterprise with the rest of society.

McClennen's Pareto-based coordination principle yields a unique prediction in some games with conflicts of interests, such as the prisoner's dilemma. The prisoner's dilemma is a symmetric two-person noncooperative game with the following features: (i) each player has two strategies, say, cooperate and not cooperate, (ii) no matter what strategy one player chooses, the other player's best response is to not cooperate, and (iii) both players have higher payoffs if they both cooperate than if they both do not. This game has a unique Nash equilibrium in which nobody cooperates. The only strategy combination that Pareto dominates this outcome has both players cooperating, which is what McClennen's account of rationality identifies the players as doing.

Matters become more complicated when there is more than one way in which the benefits from cooperation (relative to a benchmark like that provided by a non-Pareto optimal Nash equlibrium) can be distributed. If individuals are not satisfied with the way in which the benefits of cooperation are shared, then their willingness to cooperate may dwindle, and they may engage in rent-seeking or other noncooperative strategies aimed at promoting their narrow self-interest.

To avoid, or at least minimize, this kind of destructive behavior, McClennen assumes that individuals also have a disposition to accept an egalitarian principle reminiscent of Rawls's maximin principle. This principle requires inequalities in the sharing of the gains from cooperation to be only permitted if they are mutually advantageous. McClennen argues that this principle offers the worst-off individuals the assurance that they could not be made better off. It is this assurance that helps secure their cooperation. Rawls (1971, p. 176) uses a similar reasoning when arguing that his

[43] Alternative x is Pareto superior to y (or, equivalently, x Pareto dominates y) if everyone prefers x to y. McClennen uses the phrase "Pareto-efficient" for this domination relation.

principles of justice can be adhered to without great difficulty. However, this argument, while important, is given less prominence by Rawls than the argument that it is the special features of the veil of ignorance that lead parties behind the veil to adopt his maximin principle (see Rawls, 1971, section 26). McClennen believes that arguments based on the need for securing cooperation, of which he offers various forms related to the stability and adaptability of social arrangements, help give legitimacy to his maximin principle.

1.5 Rights and Liberties

Rights and liberties are essential to the fabric of modern societies, but the proper way of conceiving their foundations, their value, and how best to promote and protect them remains a matter of debate. Karl Marx, for example, argued that real freedom, as opposed to formal freedom, is what matters to individuals. The poor are formally free to buy whatever they want, but their real freedom is very limited. Marx's conception of freedom has been influential ever since he articulated it, but it is just one of many conceptions of freedom.

Modern discussions of liberties owe much to Isaiah Berlin's (1958) distinction between negative and positive freedom. An individual possesses negative freedom to the extent that his actions are not subject to the deliberate interference of others. Interference here is broadly construed to include threats, not just physical coercion. In contrast, an individual possesses positive freedom to the extent that he is his own master in pursuing his rational ends. Although Berlin acknowledged the importance of both types of freedom, he was wary of the ways in which totalitarian regimes have justified coercive policies by claiming that individuals who do not accede to their wishes are mistakenly not pursuing their "real"interests. For Berlin, the plurality of rational ends is a fundamental feature of human nature. To respect this pluralism, he argues that individuals must be free to choose their own ends without interference from others.

Libertarians, such as Friedrich Hayek, take negative freedom as the primary value to be protected and define a state of liberty as being one in which "coercion of some by others is reduced as much as is possible in society" (Hayek, 1960, p. 11). This goal is achieved by awarding the state a monopoly on coercion, so as to limit the coercion of individuals, while at the same time requiring the state to operate under known general rules so that its coercive powers are constrained as much as possible. Advocates of this view, such as Milton Friedman (1962), argue that, with limited exceptions, competitive capitalism – private enterprises competing in a free market – is the only economic institution compatible with these libertarian principles. Property rights need to be defined and enforced by the state, but activities such as progressive taxation for redistributive purposes are, in this view, inimical to liberty.

Rawls (1971, section 32) regards the dispute between proponents of negative and positive freedom about the definition of freedom as being somewhat misguided. He believes that the real dispute is about the relative weight to be placed on different kinds of liberties when they conflict. Nevertheless, it is instructive to consider how Rawls's theory of justice employs conceptions of freedom similar to those previously discussed. By granting absolute priority to the protection of basic liberties, his principles of justice provides some solace to libertarians and others who value negative freedom by guaranteeing that egalitarian policies do not encroach on individuals' rights to act as they wish within an extensive protected sphere of activity. Rawls's maximin principle, with its focus on and concern for those who are disadvantaged in terms of primary goods, embodies an endorsement of Marx's plea for real freedom, as well as a belief that the protection of basic liberties does not preclude substantial redistribution. Furthermore, his conception of citizens as individuals who are free to form and revise their own conceptions of the good life and the duty he attributes to society to provide individuals with the means to behave as autonomous moral agents in this sense can be viewed as a particular instantiation of Berlin's concept of positive freedom. At the risk of some oversimplification, we can describe Rawls's principles of justice as promoting equal real and positive freedom within the bounds delineated by the protection of negative freedom.

1.5.1 Pettit on Republicanism

In Chapter 16, Philip Pettit argues that a more satisfactory alternative to freedom as noninterference (i.e., Berlin's concept of negative freedom) is provided by the republican ideal of freedom as nondomination. The origins of republicanism can be traced to ancient Roman times and it was influential during the Renaissance and seventeenth and eighteenth centuries. Pettit has been instrumental in helping to revive this tradition, most notably in Pettit (1997). His chapter provides a good introduction to the basic tenets of republicanism (see also Skinner, 1978).

With Berlin's concept of negative freedom, an individual is not free to the extent that he is subject to actual coercive interference by some other party. This interference may or may not be arbitrary (i.e., subject to the discretion of the intervenor). In contrast, the republican concept of freedom regards someone as being free if he is not subject to the arbitrary interference, actual or potential, of some other party (i.e., if he is free from domination). In effect, freedom as nondomination regards an individual as being free to the extent that he is not subject to the mastery of others, thereby sharing the idea of a protected sphere with the concept of negative liberty and sharing the idea of self-mastery with the concept of positive liberty (see Pettit, 1997, pp. 21–22). Republicanism articulates a conception of society in which domination is limited as much as possible.

The republican critique of freedom as noninterference should not be confused with the Marxian critique that the absence of interference is compatible with a lack of real freedom because of naturally or socially generated impediments to the exercise of autonomy. For republicans, it is not a question of whether an individual's freedom has been constrained because of some actual interference in his affairs. Rather, the issue is whether he is subject to the *possible* arbitrary interference of others, whether this potential interference is exercised or not. For example, in terms of actual interference in one's activities, the condition of a slave with a good master was, in ancient times, not qualitatively much different from that of a modern worker. But there is a significant difference in their situations because the slave master had the ability to arbitrarily interfere with the slave's life in ways that are now prohibited.

Republicanism is compatible with state-imposed constraints on one's actions provided that they are not arbitrary. For example, the citizen of a democratic state who submits to a tax adopted by an assembly of elected representatives suffers from interference in Berlin's sense, but this is not an instance of domination because this interference is not arbitrary, having been established by what Pettit calls "a fair rule of law." Thus, laws are necessarily coercive from the perspective of freedom as noninterference, whereas being subject to the rule of law does not limit one's freedom in the republican sense if the law is not arbitrary. In this regard, republicans share with libertarians such as Hayek the view that "when we obey laws, in the sense of general abstract rules laid down irrespective of their application to us, we are not subject to another man's will and are therefore free" (Hayek, 1960, p. 153). To the extent that laws are freely complied with and reflect the will of the people as expressed through democratic institutions, they are legitimate. By legitimizing this kind of coercive activity, republican freedom incorporates the defining feature of Berlin's concept of positive freedom – self-mastery.

Viewing liberty as freedom from noninterference severely constrains the role of the state in promoting effective freedom, as any redistributive policy aimed at improving the lot of the disadvantaged is viewed as an intrusive interference in the economic activities of individuals. Pettit argues that for such a policy to be justified according to this view of freedom, it must be shown that this restriction on liberty results in a lowering of the probability of interference by other agents of sufficient magnitude to offset the direct reduction in freedom, which he regards as being a rather dubious proposition.

What separates liberals such as Rawls from libertarians such as Hayek is the rejection by the former of the view that the state should limit its activities to promoting equal freedom from noninterference for all. Such liberals also want to promote effective freedom, as exemplified by Rawls's difference principle. Pettit argues that the republican ideal of maximizing equal freedom from nondomination is not only compatible with a concern for effective freedom, but that it actually *requires* the state to engage in activities, both redistributive and regulatory, that limit the inequalities in access to resources and positions of authority that are conducive to the domination of some individuals by others. For example, introducing unemployment insurance reduces the dependence of an employee on the good will of his employer.

More generally, Pettit argues that there is no presumption that a redistributive policy restricts freedom provided that it is enacted under a fair rule of law even though any redistribution involves depriving the wealthy of resources that they could otherwise enjoy. Provided that a redistributive policy serves the goal of expanding equal freedom from nondomination, such a policy can be justified even to the rich on the basis of a shared commitment to promoting republican freedom. Furthermore, according to Pettit, it is often much easier to ascertain that a policy will reduce the capacity of some agents to interfere arbitrarily in the affairs of others (a qualitative assessment) than to estimate the reduction of actual interference that it will entail (a quantitative assessment), as required to determine whether a policy reduces overall freedom from noninterference. Pettit is well aware that by providing an extensive role for the state, there is a danger that it might use its powers arbitrarily. He therefore cautions that sufficient safeguards must be put in place to prevent this from happening.

In summary, republicanism, with its focus on freedom as nondomination, is more open to egalitarian policies than the standard liberal view of freedom as noninterference. Moreover, republicanism actually demands that the state engages in activities that can be shown to reduce the domination experienced by any segment of society.

1.5.2 Riley on Rule Utilitarianism and Liberal Priorities

Harsanyi was a strong proponent of rule utilitarianism. Rule utilitarianism, first proposed in Harrod (1936), is the doctrine in which the utilitarian criterion is applied to moral rules, not to individual acts, as is the case with act utilitarianism. In Chapter 17, Jonathan Riley describes and critically examines Harsanyi's game-theoretic version of rule utilitarianism, as exposited in Harsanyi (1992).[44] Riley argues that rule utilitarians necessarily give moral priority to equal rights and liberties over other values. He also suggests a way in which Harsanyi's theory can be modified so that it gives priority to certain kinds of rights over others.

We have already seen that giving absolute priority to rights and liberties in Rawls's theory over other socioeconomic advantages has been criticized by Harsanyi in his contribution to this volume. However, utilitarians like Harsanyi are themselves criticized for providing no serious protection to individuals from intrusions on their basic rights because any such intrusion is acceptable if it increases the sum of utilities. Security and integrity of the person are very important values for most, if not all, individuals and the benefits from protecting property rights can be substantial; therefore, one might expect that basic liberties will be widespread if the utilitarian criterion is adopted. Indeed, Harsanyi (1992, section 11) believes that in a rule utilitarian society certain rights and obligations would be protected in the sense that they should not be overridden, even if the immediate direct effects of doing so increases aggregate utility, except in catastrophic situations. However, this outcome is contingent on what preferences individuals actually have. Utilitarians could very well accept violating the rights of a small minority of people if doing so greatly benefits the rest of the population. Harsanyi (1992, p. 696) acknowledges as much when he says that it is acceptable to reduce the scope of our liberties somewhat if the resulting benefits in terms of other social values is sufficiently great. As Riley notes, Rawls considers it an essential goal of his theory of justice to provide a more secure foundation for basic liberties than is provided by utilitarianism.

In Harsanyi's version of rule utilitarianism, an action is morally right if it conforms with an optimal moral code, which in turn consists of the set of moral rules whose acceptance would maximize the sum of utilities if everyone abided by them. Harsanyi models a rule utilitarian society

[44] Early statements of Harsanyi's views on rule utilitarianism may be found in Harsanyi (1977a, 1977d). His formalization of rule utilitarianism evolved over time and we, like Riley, take the presentation in Harsanyi (1992) as being definitive.

as a two-stage game. In the first stage, an optimal moral code is chosen cooperatively from among all feasible moral codes. This moral code identifies a set of permissible actions (i.e., strategies, to use game-theoretic terminology) that are the same for everyone.[45] The second stage is a noncooperative game in which individuals choose strategies from among those permitted by the moral code to advance their own personal interests. To identify an optimal moral code, it is necessary to predict what the equilibrium behavior in the second-stage game will be. This behavior depends on what moral code is chosen in the first stage.[46]

Thus, for Harsanyi, rule utilitarianism consists in selecting the moral code that governs social interactions from the set of all possible moral codes in order to achieve the highest possible aggregate utility when individuals with their ordinary, more or less selfish, motives make decisions in conformity with this code. Harsanyi's model of a rule utilitarian society is similar to how economists generally model social decision problems when there are incentive constraints (e.g., in the theory of optimal taxation). In such problems, there is a set of feasible rules of the game and the objective is to choose those rules that yield the best consequences (as measured by some social objective function) given (i) the distribution of the characteristics (preferences, labor productivities, etc.) in the society under consideration and (ii) a prediction of how these individuals will behave once the rules of the game are enforced.

Because the moral code implied by act utilitarianism, which requires every decision to be made so as to maximize the sum of utilities, is one of the codes that could be adopted by a rule utilitarian, Harsanyi (1992, p. 686) argues that an optimal rule utilitarian moral code can be no worse than act utilitarianism. Furthermore, it is strictly better for at least three reasons. First, act utilitarianism is too burdensome for ordinary individuals, who would have to suppress their natural tendencies to give priority to the interests of themselves and their loved ones in order to maximize aggregate utility in all of their daily decisions. In contrast, rule utilitarianism permits more scope for pursuing one's own interests. Second, rule utilitarianism provides individuals with more freedom of choice than act utilitarianism. With rule

[45] If the decision makers are all rule utilitarians, then it is possible to think of the optimal moral code as being chosen by a single impartial observer operating from behind a veil of ignorance, as in Harsanyi's impartial observer theorem. Harsanyi is not particularly clear if he adopts this interpretation. Riley thinks that he does and frames his discussion of Harsanyi's model in terms of an impartial observer, but this is not essential for the main points that he makes.

[46] If one is not a rule utilitarian, there may be no good reason to abide by the utilitarian moral code in the second-stage game. Harsanyi also considers a version of his model in which some individuals are not committed rule utilitarians.

utilitarianism, it is only necesary to determine whether one's actions are in conformity with the moral code, not if one has done the best one can possibly do to promote aggregate utility. Third, an optimal rule utilitarian code provides desirable incentives to take actions (to work hard, to save and invest, etc.) that benefit society as a whole, and it provides assurances that individual interests will be protected (e.g., the interest of a lender in having a loan repaid). These effects are typically fairly small when considering individual acts in isolation, but can be considerable when the complete set of rights and obligations that constitute a rule utilitarian's optimal moral code is considered.

We thus see that Harsanyi's version of rule utilitarianism privileges, but does not give absolute priority to, certain rights and liberties (and their correlative duties) because individuals with their particularistic concerns value the freedom to make their own choices and because significant benefits result from the incentives and assurances provided by an extensive system of rights and liberties. These rights and liberties are embodied in the set of permissible strategies made available to individuals by an optimal moral code. Everyone has the same set of permissible strategies and, hence, there are equal rights and liberties. Harsanyi believes that moral duties should have absolute priority over nonmoral considerations (see Chapter 2), from which it follows that rule utilitarians have a duty to restrict their strategy choices to those that are permitted by an optimal moral code. According to Riley, it therefore follows that rule utilitarians must give absolute priority to the equal rights and liberties embodied in an optimal moral code over all other social values. However, he argues that this system of rights and liberties need not resemble a political liberal's system of rights and liberties. Indeed, the exact content of the rights and liberties in a rule utilitarian optimal moral code are contingent on the characteristics of the individuals who make up society, and so could vary from one society to another. Furthermore, for the reasons previously discussed, these moral codes would not embody absolute priority rules between different kinds of rights and duties, as is the case with the priority given to political liberties in Rawls's theory of justice.

Riley challenges Harsanyi's claim that a rule utilitarian would not give basic rights absolute priority within an optimal moral code. He attributes this view to Harsanyi's belief that requiring moral choices to be rational implies that preferences must admit finite trade-offs between rights and liberties on the one hand and other social advantages on the other.[47] Put

[47] This implication would not follow if the preference continuity assumption that is included in Harsanyi's list of Bayesian rationality postulates were dropped.

another way, there is a single good, measured by utility, that can be used to adjudicate all conflicts in values, a view that Riley calls strong monism. Riley notes that strong monism is not defended by Harsanyi.

Riley proposes an alternative to strong monism that is inspired by the work of Sen (1980–81), who suggests that utility is best thought of as having several distinct components. In Riley's proposal, different types of actions are distinguished. Within a type, actions are ranked according to expected utility, but between types, there is a fixed lexicographic priority ranking. Thus, when comparing actions, one first checks whether they are of different types, in which case they are ranked using the type priority ordering. If, and only if, they are of the same type, then the expected utility criterion is used to rank them. Riley illustrates this proposal by considering the case in which there are only two types of actions. The first type of action is sufficiently important that individuals attach claim rights to its performance, whereas the second type of action, while permissible, is not sufficiently valuable to require moral protection. We therefore have a set of liberal priorities, with claim rights given absolute priority over "mere" liberties, built into the basic structure of moral preferences.

Riley's chapter demonstrates that utilitarianism is compatible with giving absolute priority to certain kinds of rights. The contrary view has been perpetuated by the strong monism underlying the concept of utility employed by utilitarians in the past, whether this monism is hedonistic, as in the writings of the early utilitarians, or is based on more modern conceptions of utility, as in the writings of Harsanyi. Respecting liberal priorities, Riley argues, does not necessarily require abandoning the utilitarian focus on the sum of utilities or on expected utility calculations from behind a veil of ignorance. They may be accommodated simply by enlarging the domain of admissible kinds of preferences that a moral agent can employ when choosing an optimal moral code. Note, however, that in Riley's proposal, the protection of basic rights is contingent on preferences taking a certain form; therefore, Riley's revised version of utilitarianism is also subject to Rawls's criticism that utilitarianism does not provide a secure foundation for these rights.

1.6 Concluding Remarks

As the preceding discussion makes clear, the debate between utilitarians and liberal egalitarians is far from being resolved. Nevertheless, we believe that the chapters that follow have moved this debate forward by raising new issues to be considered and by shedding further light on issues that have previously been discussed in the literature. Despite their differences, our contributors agree that John Harsanyi and John Rawls are owed our heartfelt thanks for their role in helping to make social justice a subject of academic inquiry once again and for helping to promote an interdisciplinary dialogue on this subject. We hope that this volume will serve as a fitting memorial to them both.

References

Appiah, K. A. 2006. *Cosmopolitanism: Ethics in a World of Strangers.* W. W. Norton, New York.

Arneson, R. J. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56, 77–93.

Arneson, R. J. 1990. Primary goods reconsidered. *Noûs* 24, 429–454.

Arneson, R. J. 2000. Luck egalitarianism and prioritarianism. *Ethics* 110, 339–349.

Arrow, K. J. 1951. *Social Choice and Individual Values.* Wiley, New York.

Arrow, K. J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, 91–96; translated from the original 1953 article published in French.

Arrow, K. J. 1973. Some ordinalist-utilitarian notes on Rawls's theory of justice. *Journal of Philosophy* 70, 245–263.

Atkinson, A. B. 1973. How progressive should income tax be? In *Essays in Modern Economics,* ed. M. Parkin and A. R. Nobay. Longmans, London, pp. 90–109.

Atkinson, A. B. 1995. *Public Economics in Action: The Basic Income/Flat Tax Proposal.* Clarendon Press, Oxford.

Barry, B. 1989. *A Treatise on Social Justice*, Vol. 1: *Theories of Justice.* University of California Press, Berkeley.

Barry, B. 1995. *A Treatise on Social Justice*, Vol. 2: *Justice as Impartiality.* Clarendon Press, Oxford.

Berlin, I. 1958. *Two Concepts of Liberty.* Clarendon Press, Oxford.

Binmore, K. 1994. *Game Theory and the Social Contract*, Vol. 1: *Playing Fair.* MIT Press, Cambridge, MA.

Binmore, K. 1998. *Game Theory and the Social Contract*, Vol. 2: *Just Playing.* MIT Press, Cambridge, MA.

Binmore, K. 2005. *Natural Justice.* Oxford University Press, New York.

Blackorby, C., Donaldson, D., and Weymark, J. A. 1999. Harsanyi's social aggregation theorem for state-contingent alternatives. *Journal of Mathematical Economics* 32, 365–387.

Bolker, E. D. 1967. A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science* 34, 333–340.

Bossert, W., and Weymark, J. A. 2004. Utility in social choice. In *Handbook of Utility Theory*, Vol. 2: *Extensions*, ed. S. Barberà, P. J. Hammond, and C. Seidl. Kluwer Academic, Boston, pp. 1099–1177.

Bradley, R. 2005. Bayesian utilitarianism and probability heterogeneity. *Social Choice and Welfare* 24, 221–251.

Broome, J. 1990. Bolker-Jeffrey expected utility theory and axiomatic utilitarianism. *Review of Economic Studies* 57, 477–502.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty and Time.* Basil Blackwell, Oxford.

Broome, J. 1993. A cause of preference is not an object of preference. *Social Choice and Welfare* 10, 57–68.

Choné, P., Laroque, G. 2005. Optimal incentives for labor force participation. *Journal of Public Economics* 89, 395–425.

Cohen, G. A. 1989. On the currency of egalitarian justice. *Ethics* 99, 906–944.

Dershowitz, A. M. 2006. *Preemption: A Knife That Cuts Both Ways.* W. W. Norton, New York.

Dworkin, R. 1981. What is equality? Part 2: Equality of resources. *Philosophy & Public Affairs* 10, 283–345.

Dworkin, R. 2000. *Sovereign Virtue: The Theory and Practice of Equality*. Harvard University Press, Cambridge, MA.

Elster, J. 1982. Sour grapes – utilitarianism and the genesis of wants. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge University Press, Cambridge, pp. 219–238.

Fleurbaey, M. 1998. Equality among responsible individuals. In *Freedom in Economics: New Perspectives in Normative Analysis*, ed. J.-F. Laslier, M. Fleurbaey, N. Gravel, and A. Trannoy. Routledge, London, pp. 206–234.

Fleurbaey, M., and Mongin, P. 2005. The news of the death of welfare economics is greatly exaggerated. *Social Choice and Welfare* 25, 381–418.

Friedman, M. 1962. *Capitalism and Freedom.* University of Chicago Press, Chicago.

Gauthier, D. 1986. *Morals by Agreement.* Clarendon Press, Oxford.

Gibbard, A. 1979. Disparate goods and Rawls's difference principle. *Theory and Decision* 11, 267–288.

Hammond, P. J. 1981. *Ex-ante* and *ex-post* welfare optimality under uncertainty. *Economica* 48, 235–250.

Harrod, R. F. 1936. Utilitarianism revised. *Mind* 45, 137–156.

Harsanyi, J. C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61, 434–435.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321.

Harsanyi, J. C. 1975. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review* 69, 594–606.

Harsanyi, J. C. 1977a. Morality and the theory of rational behavior. *Social Research* 44, 623–656.

Harsanyi, J. C. 1977b. Non-linear social welfare functions: A rejoinder to Professor Sen. In *Foundational Problems in the Special Sciences*, ed. R. E. Butts and J. Hintikka. D. Reidel, Dordrecht, pp. 293–296.

Harsanyi, J. C. 1977c. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, Cambridge.

Harsanyi, J. C. 1977d. Rule utilitarianism and decision theory. *Erkenntnis* 11, 25–53.

Harsanyi, J. C. 1979. Bayesian decision theory, rule utilitarianism, and Arrow's impossibility theorem. *Theory and Decision* 11, 289–317.

Harsanyi, J. C. 1992. Game and decision theoretic models in ethics. In *Handbook of Game Theory with Economic Applications*, Vol. 2, ed. R. J. Aumann and S. Hart. North-Holland, Amsterdam, pp. 669–707.

Hayek, F. A. 1960. *The Constitution of Liberty*. University of Chicago Press, Chicago.

Hylland, A., and Zeckhauser, R. 1979. The impossibility of Bayesian group decision making with separate aggregation of beliefs and values. *Econometrica* 47, 1321–1336.

Jeffrey, R. C. 1965. *The Logic of Decision*. McGraw-Hill, New York.

Jones, P., and Sugden, R. 1982. Evaluating choice. *International Review of Law and Economics* 2, 47–65.

Kalai, E., and Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43, 513–518.

Kolm, S.-C. 1972. *Justice et Équité*. Editions du Centre National de la Recherche Scientifique, Paris, translated with a new foreword as: Kolm, S.-C. 1999. *Justice and Equity*. MIT Press, Cambridge, MA.

Laslier, J.-F., Fleurbaey, M., Gravel, N., and Trannoy, A., eds., 1998. *Freedom in Economics: New Perspectives in Normative Analysis*. Routledge, London.

Mandler, M. 1999. *Dilemmas in Economic Theory: Persisting Foundational Problems of Microeconomics*. Oxford University Press, New York.

Maniquet, F., and Sprumont, Y. 2004. Fair production and allocation of an excludable nonrival good. *Econometrica* 72, 627–640.

Maynard Smith, J., and Price, G. R. 1973. The logic of animal conflict. *Nature* 246, 15–18.

McClennen, E. F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press, Cambridge.

Mongin, P. 1994. Harsanyi's aggregation theorem: Multi-profile version and unsettled questions. *Social Choice and Welfare* 11, 331–354.

Mongin, P. 1995. Consistent Bayesian aggregation. *Journal of Economic Theory* 66, 313–351.

Mongin, P., and d'Aspremont, C. 1998. Utility theory and ethics. In *Handbook of Utility Theory*, Vol. 1: *Principles*, ed. S. Barberà, P. J. Hammond, and C. Seidl. Kluwer Academic, Boston, pp. 371–481.

Nash, J. F., Jr., 1950. The bargaining problem. *Econometrica* 18, 155–162.

Nash, J. F., Jr., 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–295.

Pattanaik, P. K., and Xu, Y. 1990. On ranking opportunity sets in terms of freedom of choice. *Recherches Economiques de Louvain* 56, 383–390.

Pazner, E. A., and Schmeidler, D., 1978. Egalitarian equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92, 671–687.

Pettit, P. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford.

Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Addison-Wesley, Reading, MA.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA.

Rawls, J. 1974. Some reasons for the maximin criterion. *American Economic Review, Papers and Proceedings* 64, 141–146.

Rawls, J. 1982. Social unity and primary goods. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge University Press, Cambridge, pp. 159–185.

Rawls, J. 1993. *Political Liberalism*. Columbia University Press, New York.

Rawls, J. 2001. *Justice as Fairness: A Restatement*. Harvard University Press, Cambridge, MA.

Risse, M. 2002. Harsanyi's "utilitarian theorem" and utilitarianism. *Noûs* 36, 550–577.

Roemer, J. E. 1996. *Theories of Distributive Justice*. Harvard University Press, Cambridge, MA.

Roemer, J. E. 1998. *Equality of Opportunity*. Harvard University Press, Cambridge, MA.

Savage, L. J. 1954. *The Foundations of Statistics*. Wiley, New York.

Schelling, T. C. 1960. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

Sen, A. 1970. *Collective Choice and Social Welfare*. Holden-Day, San Francisco.

Sen, A. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7, 243–262.

Sen, A. 1980–81. Plural utility. *Proceedings of the Aristotelian Society* 81, 193–215.

Sen, A. 1985. *Commodities and Capabilities*. North-Holland, Amsterdam.

Sen, A. 1986. Social choice theory. In *Handbook of Mathematical Economics*, Vol. 3. K. J. Arrow, and M. D. Intriligator. North-Holland, Amsterdam, pp. 1073–1181.

Sen, A. 1992. *Inequality Reexamined*. Harvard University Press, Cambridge, MA.

Skinner, Q. 1978. *The Foundations of Modern Political Thought*, Vol. 1, *The Renaissance*. Cambridge University Press, Cambridge.

Taylor, P. D., and Jonker, L. B. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156.

Thomson, W. 2005. Fair allocation rules. In *Handbook of Social Choice and Welfare*, Vol. 2, ed. K. J. Arrow, A. K. Sen, and K. Suzumura. North-Holland, Amsterdam, forthcoming.

Tungodden, B. 2000. Egalitarianism: Is leximin the only option? *Economics and Philosophy* 16, 229–245.

Vickrey, W. 1945. Measuring marginal utility by reactions to risk. *Econometrica* 13, 319–333.

von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Weibull, J. W. 1995. *Evolutionary Game Theory*. MIT Press, Cambridge, MA.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, pp. 255–320.

Weymark, J. A. 2005. Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare* 25, 527–555.

Young, H. P. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press, Princeton, NJ.

PART ONE

THEMES FROM RAWLS

2

John Rawls's Theory of Justice

Some Critical Comments

John C. Harsanyi

2.1 What Choices People Would Make in Ignorance of Their Own Personal Interests

Both Rawls's *A Theory of Justice*, 1971, and my own theory of moral value judgments (see, e.g., Harsanyi, 1953, 1977, chapter 4) can be interpreted as theories that try to answer the question of what social institutions people would choose if their choices were wholly unaffected by their own personal interests.

In Rawls's theory, this question takes the form of asking what social institutions people would choose in the *original position* where a "veil of ignorance" would prevent them from knowing what their own social positions and even what their own personal characteristics were and therefore from knowing their own personal interests.

In my own theory, this question takes the form of asking what social institutions people would choose for their society if they had to make their choices on the assumption that each of them would have the same probability $1/n$ of ending up in any one of the n possible social positions.

Yet, even though the basic questions Rawls and I ask are rather similar, our theories by which we try to answer them are *very different.* One important reason for this is that Rawls assumes that people in the original position would use the *maximin principle* as their decision rule, whereas I assume that people making moral value judgments would base their choices on *expected-utility maximization* in accordance with the Bayesian concept of rationality.

2.2 The Maximin Principle

Rawls's use of the maximin principle as a decision rule is rather surprising because it has been known since the early 1950s that it is an *irrational* decision rule, with very paradoxical implications (see Radner and Marschak, 1954; see also Harsanyi, 1974).

It is an irrational decision rule because it asks us to act on the assumption that any policy we may follow will always produce the *worst possible outcome*, even if that outcome has *near-zero* probability.

As a result, the maximin principle will be a *very poor* decision rule both in *everyday life* and in *ethics*. If we followed it in everyday life, then we could not eat any food because there is always a small chance that it will contain harmful bacteria. We could not cross even the quietest country road because we might be hit by a car. Nor could we marry because there is always some risk, perhaps a very small risk, that our marriage might end in a disaster.

The maximin principle is an equally poor decision rule in *ethics*. It leads Rawls to ask us to give *absolute priority* to the interests of the "least advantaged" social group over the interests of everybody else. In my view, this is an *unacceptably extreme* position. It would require us to give *absolute priority* to the interests of this social group even if they were a small minority, while the rest of society contained many millions of people. It would require us to do so even if this meant sacrificing some *very important* interests of many other people in order to protect some *very unimportant* interests of the people in the least advantaged social group.

For example, suppose the government of an advanced country decides to use a small percentage of its tax revenue to support some highbrow cultural activities, such as classical music, sophisticated theater performances, or scientific activities of great intellectual interest, yet without any clear applications. Suppose also that these cultural activities will be greatly enjoyed by a group of highly educated and relatively well-to-do people but would be of little interest to the less well educated and economically disadvantaged members of the community.

This government policy would no doubt violate Rawls's requirement of giving *absolute priority* to the interests of the least advantaged social group. But this only confirms my view that this requirement would entail prohibition of many morally unobjectionable social policies.

Many people do not seem to realize that Rawls's theory would require *very extensive* redistribution of income and wealth in our country, and that any attempt to implement this policy would give rise to serious economic and political problems. It would cause major dislocations in our economy. It would also create heated confrontations between the opponents and the supporters of this policy, possibly leading to widespread civic unrest and perhaps even to a civil war. In evaluating Rawls's theory, these problems must also be taken into account.

It is a remarkable fact that the two leading opponents of utilitarianism, John Rawls and Robert Nozick, both moved away from the middle-of-the-road liberalism of most philosophers of the previous generation to rather radical *extreme positions* in opposite directions of the political spectrum. Nozick moved very much to the *right*, opposing all forms of *income redistribution*, and therefore advocating such inhumane policies as abolition of all government programs using tax revenues to relieve poverty. In contrast, Rawls moved to the left, advocating quite *radical* policies of income redistribution.

2.3 Other Absolute-Priority Principles in Rawls's Theory

Other principles of absolute priority also play an important role in Rawls's theory. (For a rather complicated hierarchy of such principles, see Rawls, 1971, pp. 302–303.) But his arguments in support of these principles are rather vague and unconvincing.

I shall restrict my discussion to one such principle. Rawls argues that, when society reaches a high level of economic well-being, people will become less willing to accept restrictions in their personal liberty for the sake of economic benefits, and will in fact assign *absolute priority* to their basic liberties over their economic interests (Rawls, 1971, p. 542).

In my own view, this claim of Rawls is contrary to the facts. Even in our own society, people often make voluntary agreements restricting their freedom of action in certain ways in exchange for some economic compensation. They also support legislation and government regulations restricting all citizens' freedom of action in the hope that these restrictions will increase the economic well-being of our society as a whole.

More generally, to assume that one social value had absolute priority over another would imply that the former social value was *infinitely more important* than the latter. Yet, in situations where we have to choose between two different social values, we usually find that there is a *finite trade-off* between these two values, and that we have to decide what trade-off we are willing to accept between these values. For instance, we have to decide how much *individual freedom* or how much *economic efficiency* we are willing to give up for a given *decrease in economic inequalities.* Or we have to decide how to balance society's interest in *deterring crime* against its interest in ensuring the *fairness of criminal trials* or how to balance the interests of *gifted children* against those of *slow learners* in various schools, and so on. In none of these cases will we assign absolute priority to one social value over the other.

To be sure, there is a clear case of *absolute priority* of one social value over all others. It is the absolute priority we must assign to our *moral duties* over personal interests and over all other nonmoral considerations. For instance, we must not engage in immoral behavior for any amount of money, however large, or even as a favor to a friend.

2.4 Rawls's Attempt to Deny Moral Credit to Many People Performing Valuable Services to Society

According to the traditional commonsense conception of justice, which Rawls calls the *system of liberal equality* (Rawls, 1971, pp. 65, 73), when people perform valuable services to society then they *deserve* moral credit and suitable rewards for doing so as a matter of *justice* as such.

Rawls rejects this view in favor of a new conception of justice, which he calls the *system of democratic equality* (1971, pp. 65, 101–103). He admits that it is desirable for society to establish suitable rewards for people making valuable contributions as *incentives* for further similar contributions. But he denies that such people actually *deserve* any moral credit or any special rewards as a matter of *justice*.

He argues that when people make such contributions, the latter are based on their special *talents* or on their *good character* or both, thereby enabling them to make the required efforts. Yet, they do not owe either their talents or their good character to their own *personal merits*. Rather, they owe their talents to their *good luck* of being born with a favorable *genetic endowment*, and owe their good character to their *good luck* of having been born into a *good family* and into a *favorable social environment*. Accordingly, they *do not deserve* any moral credit or any special rewards for making these contributions.

After arguing that people do not deserve the talents they have been born with and the social position they have been born into, Rawls writes: "The assertion that a man deserves the superior character that enables him to make the effort to cultivate his abilities is equally problematic; for his character depends in large part upon fortunate family and social circumstances for which he can claim no credit. No notion of desert seems to apply to these cases" (1971, p. 104).

2.5 Some Comments on Rawls's Argument

This is a very strange and one-sided argument. No doubt talented people do not deserve any moral credit for their native talents. But they do deserve moral credit for *developing* their talents and for *using* them for our common benefit.

Even stranger is Rawls's attempt to deny that people with a good character deserve moral credit for their *good character* and for their *effort* to achieve socially desirable objectives because of their good character. For if we deserve moral credit for anything at all, then we surely deserve it for our good character and for our morally good behavior based on our good character.

Of course, Rawls is right in arguing that it is *much easier* to develop a good character under favorable than under unfavorable social conditions. But as a matter of common sense, to develop a good character even under very favorable social conditions will always be an important *personal achievement* and will be unquestionably to one's *moral credit*.

When Rawls tries to deny it, he is implicitly denying that people have *free will* and are under normal conditions *morally responsible* for their actions and for their moral character. In fact, his views amount to adopting some form of "*hard*" determinism[1] without actually saying so and without offering any argument in support of this rather uncommon and implausible philosophical position.[2]

2.6 Rawls's Conception of Justice and Social Policy

Some people might feel that the difference between the traditional and the Rawlsian conceptions of justice is purely academic and has no practical implications. They might argue that it should be immaterial whether society was to reward superior performance merely as a requirement of expediency and of accepted social conventions, or was to do so strictly as a requirement of *justice* – as long as performance was rewarded in a suitable manner.

Yet, this would be in my opinion a rather shortsighted point of view. Human excellence simply *cannot* be fostered by external rewards alone. It can flourish only in a society that truly *recognizes* its intrinsic value and its social importance, and that has genuine *respect* for individuals whose performance reaches the standards of excellence. It cannot flourish in a society that *denies* that superior performance merits and deserves social recognition and other rewards strictly as a matter of *justice*.

[1] Determinism is the view that people's behavior is ultimately dependent on their social and natural environment and on their own genetic makeup, in accordance with some strictly deterministic laws or, what is more likely, in accordance with some merely probabilistic laws. Determinism is called *soft* or *hard* depending on whether it is or is not assumed to be compatible with *free will* and with *moral responsibility*.

[2] In order to place Rawls's views in proper perspective, I shall briefly discuss the problems of free will and of moral responsibilty in Sections 2.7 and 2.8.

First, as I have tried to show, when Rawls claims that people with special abilities and with special moral qualities *do not deserve* any reward or even any moral credit for any superior performance, this claim lacks any rational basis and is, therefore, *very unfair* to these people. What is even more important from a practical standpoint, if our society actually accepted this highly unfair claim, then this could be only interpreted as sending these people the socially very counterproductive message that society *did not really appreciate* their contributions and, indeed, would actually prefer that the comfortable quiet of a Philistine acceptance of general mediocrity were never disturbed by upsetting instances of individual excellence and of superior performance. Yet, this would have been the wrong message to send at any time, but would be particularly inappropriate today.

For, under present conditions, no country can stand up to international competition, can enjoy a high – let alone a continually increasing – standard of living, or can maintain a high level of cultural creativity, unless it shows *genuine appreciation* for individual excellence and for superior performance, and unless it *encourages* all its citizens, both those with outstanding ability and those with lesser ability, to develop their talents to the very limit of their capacity.

To be sure, by encouraging all people to do their best in developing their abilities, we may in fact significantly *increase the difference* in performance between people of great ability and those of lesser ability. But we simply cannot afford educational policies and other social policies that would artificially *discourage* the ablest people, and often also those with lesser ability, from reaching their full potential, which is now done by many schools in the United States and in many other Western countries. The simple fact is that we absolutely need as many well-trained and creative individuals as possible.

Let me add that even though, unlike Rawls, I take the view that justice itself requires us to reward superior performance in a suitable manner, I agree with him that, in doing so, we must not create needless economic and social inequalities. (In fact, it seems to me that such a policy would be fully compatible with significantly smaller economic and social inequalities than we have today.)

2.7 Free Will and Moral Responsibility

As is well known, the question of moral responsibilty gives rise to a curious dilemma. On the one hand, our personal experience suggests that we are rational agents possessing *free will* because under normal conditions we seem to be in control of our own actions and, therefore, feel *morally responsible* for them.

On the other hand, both philosophical reflection and the results of modern science suggest a *deterministic* view of human behavior, according to which, even though our actions are normally under our immediate control, in ultimate analysis they must depend on some causal antecedents *beyond our own control*, such as our genetic endowment and many aspects of our natural and social environment.

Human organisms are parts of nature and, for all we know, are subject to the basic laws, strictly deterministic or perhaps merely probabilistic, of physics and chemistry. This means that, like other natural objects, they can be the causes of all kinds of effects, but cannot be their causally independent *ultimate causes*, not even the ultimate causes of their own actions.

Some philosophers, the *incompatibilists*, take the view that determinism and free will are *incompatible*. Accordingly, they feel compelled *either* to hold on to determinism and to reject free will or to hold on to free will and to reject determinism.

Other philosophers, very likely the overwhelming majority, are *compatibilists*, embracing *both* determinism and free will. Nonphilosophers are similarly divided. I am a *compatibilist*. I feel that if free will is *rightly interpreted* then there is no inconsistency between determinism and free will.

2.8 Another Interpretation of Free Will: The Bearer's Responsibility View

To be sure, there is a widespread uncritical intuitive interpretation of free will – I shall call it the *ultimate causal responsibility* view – that would make it impossible for human beings to have free will. It would interpret free will as being the causally independent *ultimate cause* of one's actions and, therefore, as having ultimate causal responsibility for them. As I have already argued, we *cannot* be the ultimate causes of our own actions, which means that we *cannot* have free will in the sense required by the ultimate causal responsibility view. This is so because, even though we are normally in *immediate* control of our own actions, we are not causally *independent* agents, in that our behavior is causally dependent on our inherited psychological characteristics and on many environmental conditions that are *not* under our own control.

Yet, it seems to me that we do have free will and are normally morally responsible for our actions in the following sense. Under normal conditions, our actions are not forced on us from the outside but rather spring from our own – self-chosen but ultimately heredity- and environment-dependent – *deepest moral attitudes*, which means that they spring from the very *inner core* of our own moral personality. Thus, our actions indicate how strong a commitment we have to respect other people's rights and legitimate interests and, more generally, to uphold the basic moral values. Indeed, it is our actions and our moral attitudes underlying our actions that indicate *what kind of persons* we really are.

Admittedly, in some cases, closer observation will show that some people's actions do not fully reflect their true moral attitudes but rather result from some momentary impulse, or even from a persistent habitual inclination that they may have been trying to overcome without full success. Or, these actions might result from mental illness. In such cases, as we know that these actions do not fully reflect their real deeper moral attitudes, we do not hold them fully responsible for them.

Yet, apart from such cases of reduced responsibility, we do hold people morally responsible for their actions and for their moral attitudes underlying their actions, even though we know perfectly well that human behavior and human attitudes are strongly influenced by the social and natural environment and by some inherited psychological attributes. We do so precisely on the assumption that their actions and their moral attitudes show *what kind of persons* they really are. More specifically, we do assign *moral credit* or *discredit* (i.e., positive or negative moral responsibility) to people when their actions express moral attitudes consistent or inconsistent with the morally required minimum standards.[3]

In contrast to the *ultimate causal responsibility* interpretation of free will and of moral responsibility, I shall describe the view just outlined as the *bearer's responsibility* view because it interprets people's moral responsibility as their eligibility to moral credit and discredit for being the *bearers* of the moral attitudes expressed by their actions and because it interprets free will as an ability to choose one's own actions in accordance with one's own self-chosen, even if heredity- and environment-dependent, moral attitudes.

It is important to understand, it seems to me, that human beings *do have* free will in the relevant sense and *do deserve* moral credit and discredit for their actions and for their moral attitudes underlying their actions. For it is an essential part of the notions of *morally good* and of *morally bad* actions and attitudes that they are to our *moral credit* or to our *moral discredit*. If this were not the case, that is, if we had no moral responsibility for them,

[3] These definitions of moral credit and discredit are based on criteria suggested by Brandt's (1985) theory of criminal responsibility.

then the very concepts of *moral value* and of *moral disvalue* would lose their meaning, and moral philosophy itself would become an empty intellectual enterprise without any real subject matter at all.

References

Brandt, R. B. 1985. A motivational theory of excuses in criminal law. In *Criminal Justice: Nomos* 27, ed. J. R. Pennock and J. W. Chapman. New York University Press, New York, pp. 165–198.

Harsanyi, J. C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61, 434–435. Reprinted in Harsanyi, 1976, as chapter 1.

Harsanyi, J. C. 1974. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review* 69, 594–606. Reprinted in Harsanyi, 1976, as chapter 4.

Harsanyi, J. C. 1976. *Essays on Ethics, Social Behavior, and Scientific Explanation.* D. Reidel, Dordrecht.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, Cambridge.

Radner, R., and Marschak, J. 1954. Note on some proposed decision criteria. In *Decision Processes*, ed. R. M. Thrall, C. H. Coombs, and R. L. Davis. Wiley, New York, pp. 61–68.

Rawls, J. 1971. *A Theory of Justice.* Harvard University Press, Cambridge, MA.

3

Rawls, Responsibility, and Distributive Justice

Richard Arneson

The theory of justice pioneered by John Rawls explores a simple idea – that the concern of distributive justice is to compensate individuals for misfortune. Some people are blessed with good luck; some are cursed with bad luck, and it is the responsibility of society – all of us regarded collectively – to alter the distribution of goods and evils that arises from the jumble of lotteries that constitutes human life as we know it. Some are lucky to be born wealthy, or into a favorable socializing environment, or with a tendency to be charming, intelligent, persevering, and the like. These people are likely to be successful in the economic marketplace and to achieve success in other important ways over the course of their lives. However, some people are, as we say, born to lose. Distributive justice stipulates that the lucky should transfer some or all of their gains due to luck to the unlucky.

In *A Theory of Justice,* Rawls suggests how to draw a line between the misfortune that is society's responsibility and the misfortune that is not by distinguishing between deep and shallow inequalities. The former are associated with inequalities in the "basic structure" of society in this passage:

> For us the primary subject of justice is the basic structure of society, or more exactly, the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation. By major institutions I understand the political constitution and the principal economic and social arrangements. . . . 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. Not only are they pervasive, but they affect men's initial chances in life; yet they cannot possibly be justified by an appeal to the notions of merit and

desert. It is these inequalities, presumably inevitable in the basic structure of any society, to which the principles of social justice apply.[1]

Rawls's idea is appealing. Think of two persons: one born on the "right," the other on the "wrong" side of the tracks; one blessed with capable and nurturing parents, the other cursed with parents from the bottom of the barrel; one born with a genetic endowment that predisposes her to talent and fortune, the other plagued by an unfortunate genetic inheritance; one wealthy from birth, the other poor. From the start, before either child has taken a step out of the cradle, they have unequal life expectations given their initial circumstances. The contrast between basic structural inequalities and nonbasic ones does not seem exactly to coincide with the distinction between deep and shallow inequalities: Inequalities in genetic inheritance do not arise from the way that the core institutions of society are set. The important contrast here seems to be between deep inequalities among persons, those that are present from birth, in given social circumstances, and shallow inequalities that arise later as a result of processes that are influenced by voluntary choice.

As is well known, Rawls's master proposal concerning justice is that these inequalities are justifiable just in case they are set so that over time the least advantaged individuals are rendered as well off as possible. Advantage is measured in terms of an index of what Rawls calls *primary social goods*, general-purpose resources of which any rational person would prefer to have more rather than fewer. In this chapter, I assume with Rawls that the morally appropriate response to misfortune specifies distributions that tilt in favor of worst-off individuals, give priority to the worst off; the exact degree of tilt that is appropriate is an important issue, but not one this chapter considers.

A complication enters when Rawls separates the primary social goods into basic liberties and the rest. The basic liberties are associated with the status of citizens in a democracy and required to be equal for all citizens. The idea of maximizing from the standpoint of the worst off is applied to the holdings of the rest of the primary social goods, and holdings of income and wealth are taken to be a rough proxy for these. Rawls then supposes that in applying his principles of justice there are two relevant

[1] John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), p. 8. The objection might be raised that I am making too much of this one passage in Rawls and ignoring his more central lines of thought on responsibility. I focus on the contrast between deep and shallow inequalities because I believe it to be plausible and worth considering quite independently of its degree of centrality in Rawls's own thinking.

social positions, that of equal citizen and that determined by one's place in the distribution of wealth. Rawls proceeds to reiterate the idea that our concern should be unchosen basic structural inequalities: "Since I assume that other positions are entered into voluntarily, we need not consider the point of view of men in these positions in judging the basic structure."[2] What is puzzling is that the distribution of income and wealth is as much the outcome of voluntary choice as of unchosen starting points. Rawls makes two suggestions for defining the worst-off class of individuals: Either take all those with the income and wealth of the typical unskilled worker or less or take all persons with less than half of the median income and wealth. This group then constitutes the worst-off group whose long-run expectation of primary social goods is the job of social justice to maximize.

When I first read these passages, I was reminded of Alfred Doolittle, the sagacious worker in George Bernard Shaw's *Pygmalion*.[3] Doolittle, seeking a handout, proclaims himself to be one of the undeserving poor, whose needs are just as great as the needs of the most deserving. The least advantaged class, as defined by Rawls, is a heterogeneous group, whose members differ in characteristics that should render them differentially entitled to assistance from the better-off members of society. The point here is not, or anyway need not be, that the Alfred Doolittles of the world are morally disreputable persons who should be penalized. The point is that they are by any reasonable standard among the better-off members of society, not the worst off. A person who is very talented and possesses desirable traits such as charm and gregariousness may have a decided and steady preference for leisure over moneymaking activity and may adopt a plan of life that involves voluntary avoidance of such activity. Even though his bank-account wealth and income are low, he is living well, but Rawlsian justice lumps him together with the desperately poor who are barely able to find marginal employment. One might also suppose that some individuals with income and wealth above the average were not blessed with good fortune in the natural lotteries of talent, inherited wealth, and early socialization. These individuals simply work with above-average zeal to make the most of their opportunities, and they may also have special unchosen needs that require them to have a high income to have a decent life. It might seem that this point concerns the degree to which it is reasonable to take income and wealth as a proxy for one's index of primary social goods. Presented with this difficulty, this is

[2] Rawls, *A Theory of Justice*, p. 96.

[3] George Bernard Shaw, *Pygmalion, A Romance in Five Acts* (Baltimore: Penguin, 1951 [originally published 1916]).

the line that Rawls has taken in response.[4] He has proposed that we should count leisure among the primary social goods and should stipulate that anyone who enjoys voluntary unemployment be credited automatically with a larger share of primary social goods than anyone who works for a living. But the core difficulty is that, according to Rawls's own stated rationale for his principles of justice, they should compensate for otherwise unacceptable inequalities in people's unchosen circumstances, the luck of fortune that puts individuals on the right side or the wrong side of the tracks at birth. The difference principle mixes together deep and shallow inequalities promiscuously. And whatever Rawls's own views might be, surely justice requires society to distinguish the cases that Rawls lumps together and, if feasible, to treat in different ways inequalities that are beyond one's power to control and inequalities that arise from voluntary choices for which individuals can take responsibility.

From this point on I shall mostly ignore the distinction between Rawls's general conception of justice, which identifies it with the maximization of the primary social goods holdings of the group in society that has the least of these goods, and the special conception, which holds only under conditions of modern society, when it becomes rational to single out the basic liberties of constitutional democracy for special priority over all other primary social goods. This complication does not matter in what follows, so Rawls's theory can be represented by the general conception.

3.1 Rawls on Deservingness and Responsibility

In an interesting discussion in *A Theory of Justice,* Rawls attacks the idea that notions of merit or deservingness should be included among the values that the principles of justice should assert as fundamental. He urges that the principle of distribution according to merit must in the end reward individuals for inherited traits for which the bearers of these favored traits can claim no credit. This point holds even for conceptions of merit that, to the naive theorist, might seem attainable equally by anybody. Rawls writes, "Even the willingness to make an effort, to try, and so to be deserving in

[4] See John Rawls, "The Priority of the Right and Ideas of the Good," *Philosophy and Public Affairs* 17, no. 4 (Fall 1988): 251–276. Much of this discussion is incorporated in his *Political Liberalism* (New York: Columbia University Press, 1993), Lecture V. See also Philippe Van Parijs, "Why Surfers Should Be Fed: The Liberal Case for an Unconditional Basic Income," *Philosophy and Public Affairs* 20, no. 2 (Spring 1991): 101–131. Much of this discussion is incorporated in his *Real Freedom for All: What (If Anything) Can Justify Capitalism* (Oxford: Oxford University Press, 1995).

the ordinary sense is itself dependent upon happy family and social circumstances."[5] And again, "the effort a person is willing to make is influenced by his natural abilities and skills and the alternatives open to him. The better endowed are more likely, other things equal, to strive conscientiously, and there seems to be no way to discount for their greater good fortune."[6] Rawls adds that notions of merit and deservingness may emerge and play a role within associations and schemes of cooperation. Within these contexts, the notions may work to motivate participants to put forward their best efforts in ways that will further the goals of the association. But the viability of this instrumental use of merit and desert has no tendency to show that the notions are fit to function as fundamental justice values.

To these considerations, Rawls adds another argument. He supposes that the best interpretation of a desert-based theory of justice would say that distribution of benefits and burdens should be set so that reward is proportional to virtue or moral worth. He then adds that the notion of moral worth is best understood as the disposition to comply fully with norms of justice, so one cannot define the notion until the norms of justice are independently defined and on hand. The idea of moral worth thus strikes him as an inherently secondary matter, logically unsuited to figure in a fundamental norm of justice: "For a society to organize itself with the aim of rewarding moral virtue as a first principle would be like having the institution of property in order to punish thieves."[7]

However, the logical difficulty that Rawls notes does not decisively sweep deservingness values into secondary, instrumental status. The fundamental deservingness idea could be that fault forfeits first. That is, if lesser life prospects must be imposed on some, it is morally better that those whose conduct is by comparison more faulty should suffer the imposition, wherein

[5] Rawls, *A Theory of Justice*, p. 74.

[6] Rawls, *A Theory of Justice*, p. 312. Rawls's rejection of deservingness and merit as fundamental moral determinants of an individual's just share is not based on hard determinism, the claim that every event is determined by preexisting conditions according to causal laws, that human actions are events, and that being determined in this sense precludes moral responsibility. Rawls's claim is rather epistemic. Matters for which people cannot be held responsible mix with matters for which people might be held responsible to cause outcomes in such a way that we can never reliably tell to what extent an individual is genuinely morally responsible for the outcomes of her actions. Moreover, even if in private life one sometimes can know enough to make a reasonable attribution of responsibility, at the level of public institutions, we cannot gain the information that would be needed to make reliable global judgments of individuals' lifetime deservingness of the sort that would render the implementation of deservingness-based justice a feasible project.

[7] Rawls, *A Theory of Justice*, p. 313.

the relevant notion of fault depends on context. In some contexts, the deserving are those who strive conscientiously, and in some contexts, the relevant conscientious striving is trying to be prudent. Here there is no vicious circularity: The idea of desert can be specified independently of the content of the requirements of justice.

As far as deservingness and responsibility are concerned, Rawls wants to walk delicately on a tightrope. He wants to deny that we should set up institutions with the aim of rewarding the deserving, but neither does he wish to deny a role to individual agency and individual responsibility within his theory of justice. After all, the distinction between deep and shallow inequalities rests on the idea that individuals sometimes make voluntary choices for which they are responsible, such that it is morally appropriate that they bear the consequences for their lives that result from these choices.

In Rawls's scheme, justice is responsible for securing a fair share of resources to individuals. That is, justice stipulates that institutions be established and sustained that will operate in conjunction with individual choices to maximize the primary social goods holdings of those with least. For those whose primary social goods holdings place them above the worst-off class, what one gets by way of primary social goods depends on how one chooses to act within these institutions. The uses that people make of their resources in their private lives are not the concern of justice. Whether one organizes one's romantic life well or poorly, for example, is not a social justice issue. Rawls also wants to hold that individuals bear responsibility for their ends, in the sense that each individual is deemed capable of affirming and, if appropriate, of revising her own conception of the good, and is responsible for the consequences for her life that flow from her embrace of one rather than another conception of the good. A conception of the good may be regarded as a set of final ends plus an account that shows how the individual final ends are coherently connected to one another and together express an idea of what is worth striving for in life or what constitutes a meaningful life. To say that an individual is responsible for her conception of the good or for her individual choices is in this context to hold that society is not obligated to compensate her for bad consequences she suffers because of having that conception or making those choices. (Nor is society authorized to take away the good fortune the individual comes to have because of her conception of the good and because of the choices she makes and transfer some of this good fortune to others who are less fortunate.) In a nutshell, the Rawlsian idea of justice is that society is obligated to provide for individuals a fair share of opportunities and resources that correct to some extent for the natural lotteries of birth and upbringing so that the expectations of the worst off are as high as they can be made. What individuals make of their opportunities and resources, the goodness or badness of the lives they fashion for themselves guided by their own individual conceptions of the good, is their own business, not in any way the responsibility of society.

3.2 The Canonical Moment Version of Rawlsian Justice

This synthesis of mutual obligation and individual responsibility sounds attractively liberal, but collapses under examination. One cannot regard people's income as fixed beyond their power to control; the employment and self-employment decisions that individuals make reflect their values, aims, and choices as well as their initial unchosen assets and the ensemble of circumstances fixed by the actions of others. One possible way to reinstate the line between deep and shallow inequalities within Rawls's system would be to adopt the simplifying device of a canonical moment at which individuals enter adulthood and are deemed fully responsible for their choices and for the further socialization and values-altering regimes they undergo. The canonical moment version of the Rawlsian difference principle would then require that at the onset of adulthood each individual be provided a fair share of primary social goods (other than basic liberties, whose distribution is to be equal). This fair share will be such as to maximize the long-run sustainable potential expected level of primary social goods of the class of individuals whose potential for acquiring primary social goods is least. In other words, on this conception, justice requires not the maximization of the expected level of primary social goods over the course of one's life of those who are worst off in this respect, but rather the maximization of the expected level of primary social goods that the worst off could anticipate if each of them chose the plan of life of those available that would provide the highest expected level of primary social goods.[8]

A regime that satisfied the canonical moment version of the difference principle would be identifying the deep or basic structural inequalities with

[8] There is an unclarity in this formulation that emerges once one notes that each individual's choice of the plan of life that is most prudent depends on what other individuals are rationally anticipated to be choosing. In forming a prudent life plan, the individual needs to anticipate not what others if ideally rational and well informed would choose, but what they will actually choose. It is not prudent for me to seek to date Ted if I know in advance that he will not seek to date me, even if I also know that if he were ideally prudent he would be willing to date me. For decisions in the economic sphere, we can finesse this difficulty by supposing that the individual reasonably expects to be making choices in a competitive environment in which the choices she makes will not have a significant impact on what others are anticipated to choose.

the inequalities in the potential for primary social goods acquisition that individuals face at the onset of adulthood, these being identified with unchosen inequalities in circumstances the just regulation of which is the primary subject of justice. The justifying idea would be that when any inequalities in the expected lifetime level of primary social goods that individuals could reach if they tried worked to maximize the expected potential level of the worst off over the long run, then and only then are those inequalities morally justified. Since compensation for the disadvantages that one suffers is set by the expected level of primary goods one could acquire rather than by the level one actually reaches, society is not in the position of compensating individuals for the consequences that fall on themselves as a result of their free and voluntary choices. This version of a just political regime does not seek to eliminate the influence of luck on the quality of lives that individuals reach. The initial basic structural inequalities that aroused our concern were not inequalities in guaranteed lifetime wealth and income levels. Being born in fortunate or unfortunate circumstances does not guarantee one a fortunate or an unfortunate life, just a greater or lesser prospect of such.

This revised Rawlsian doctrine on social justice is close in spirit to the "equality of resources" proposal espoused by Ronald Dworkin.[9] Dworkin proposes equalizing shares of resources and Rawls proposes maximinning resource shares, but this difference may not signify any serious moral disagreement because Dworkin limits himself to interpreting the ideal of equality and does not address the issue of how much weight in policy making to assign equality when it conflicts with other moral values. Dworkin could then affirm a Rawlsian maximin principle without retracting any of his assertions about the ideal of equality of resources. Rawls's view that the primary subject of justice is the way the basic structure of society distributes initial inequalities, with its implicit contrast between deep and shallow inequalities, bears significant similarity to Dworkin's contrast between option and brute luck and his identification of justifiable inequalities with those that arise because of option luck rather than brute luck given fair initial conditions. His initial formulation of equality of resources stipulates that equality of resources obtains among persons when each starts with a share of resources dictated by a theoretical equal auction and any subsequent inequalities in their resource holdings arise through option luck.

Dworkinian *option luck* is chance that affects a person's life through gambles that the person either deliberately chooses or could have chosen. *Brute*

[9] Ronald Dworkin, "What Is Equality? Part 2: Equality of Resources," *Philosophy and Public Affairs* 10 (Fall, 1981): 283–345.

luck is chance that befalls a person without any mediation of choice. Being harmed by a chance event against which there was no possibility of purchasing insurance or taking protective measures would be an instance of brute luck. But where insurance is available, the decision to purchase it or not transforms the chance event into option luck, and even if one does not advert to the possibility of purchasing insurance and make a deliberate choice, still, one could have done so, and this circumstance suffices to change brute luck to option luck. Because option luck is present in virtually all choices made by adults, I described the canonical moment version of Rawls's general conception of justice as close in spirit to Dworkin's ideal of equality of resources, which incorporates the norm that the outcomes of brute, but not option, luck should be equalized.

3.3 Responsibility for Voluntary Choices is Problematic

The proposed fusion of Rawls and Dworkin on distributive justice is an unstable doctrine. It combines the ideas that distributive justice requires compensating individuals for their unchosen talent deficiencies and that distributive justice forbids compensating individuals for the outcomes of their free and voluntary choices provided that these choices proceed from a fair prior distribution of resources. In a slogan, the proposal is that individuals should be held responsible for their choices but not for their unchosen circumstances in which choices are made. The problem is that prominent among individuals' deficiencies in talents are deficiencies in their choice-making and choice-following abilities. Consider a decision problem in which complex reasoning is required to reach a prudent decision. Two individuals may strive equally conscientiously to arrive at a prudent choice, but one has been favored with better reasoning ability and succeeds while the other fails. Or suppose instead that the decision problem is easily solved by both individuals, but it requires heroic willpower to carry out the decision, and one individual is blessed with far greater executive abilities than the other and successfully implements the chosen decision, while the other succumbs to what is for him nearly irresistible temptation. In both cases, unchosen talent differences bring about an outcome in which the talented individual is well off and the untalented individual is badly off. In such cases, the norm that justice requires compensation for unchosen differences in talent and forbids compensation for differences in well-being that arise from the quality of individual choices yields contradictory recommendations for and against compensation.

This conclusion might seem too hasty. After all, a canonical moment distributive principle can specify that the fair initial shares of resources that individuals are given should be adjusted to reflect differences in their talents, including their choice-making and choice-following talents. The individuals then proceed to make their own choices and plan their lives as they please, but ex hypothesi compensation has already been made at the start that appropriately offsets their disabilities that affect their choices. However, in general, it will not be the case that the appropriate protection for individuals with choice-making deficiencies can be determined in advance of their actual choices. Consider that any adjustment made in the initial stake of resources that a canonical moment theory of justice assigns the individual might be swamped immediately by a bad decision of that individual: Following the initial "fair" distribution, the individual engages in high-stakes gambling with a poor betting strategy and predictably loses her entire resource stake. Or suppose that immediately following the initial "fair" distribution of resources, the unfortunately endowed individual makes a mistake in judgment for which she cannot reasonably be held blameworthy and proceeds to ride a motorcycle at excessive speed on a deserted road and suffers an expectable bad accident, which leaves her subsequent life prospects gravely diminished. Adequate compensation for choice-making and choice-following talent deficits will sometimes have to take the form either of paternalistic restriction of people's liberty in contexts where disastrous choices are predictable or ex post compensation to restore individuals' life prospects following choice-inflicted personal disaster (or some mix of restriction and amelioration).

Where do these criticisms of canonical moment views leave us? It may seem that we have come full circle. I began by invoking and endorsing Rawls's idea that the primary subject of justice is the basic structure of society regarded as engendering inequalities that do not arise from individual voluntary choice, for which individuals may be held responsible. I then criticized Rawls's difference principle for its failure to distinguish inequalities due to individual choices from inequalities due to circumstances and to treat these two kinds of inequality differently. This criticism prompted a reformulation of the Rawlsian difference principle, the canonical moment difference principle, which does respect a sharp distinction between what arises from individual voluntary choice and what arises from unchosen circumstances. The canonical moment difference principle has important affinities with Ronald Dworkin's approach to distributive justice. But the distinction between inequalities arising from choice and inequalities arising from unchosen circumstances turns out to be confused because unchosen circumstances include each individual's talent endowment, and among one's talents is the ability to make and implement good choices in formulating a conception of the good and in devising a plan of life. Is the original difference principle then vindicated after all? My answer is: No.

Consider a simple stylized example. Smith and Jones have identical native talents and equally favorable childhood socialization experiences. Over the course of their lives, Smith chooses a life plan that gives her an expectation of a high level of income and other resources over the course of her life, whereas Jones chooses a life plan that gives her an expectation of a much lower resource level, which happens to place her among the Rawlsian worst-off class. The Rawlsian difference principle will recommend institutions such as a tax and transfer policy that redistributes resources from a group that includes Smith to a group that includes Jones. But Jones has freely decided to pursue life goals that do not involve maximizing her resource holdings, either because given her values, prudence does not lead her to choose this form of maximization or because she chooses to pursue life goals other than those dictated by prudence (for example, she may choose to sacrifice her earnings prospects in favor of service to a worthy cause). In either case, the transfers recommended by the difference principle are unfair. The conclusion to be drawn from the discussion to this point is that neither the difference principle nor the canonical moment difference principle adequately incorporates responsiveness to individual responsibility in the theory of distributive justice.

3.4 Responsibility for Ends Reconsidered

Pressing on the thought that individuals are not reasonably held responsible for their talents, which are available to them owing to circumstances beyond their control, leads to the idea that in some cases it is wrong to hold individuals responsible for bad outcomes they suffer that are the consequence of their inept choices of fundamental life aims, for this incompetence may have arisen inexorably from circumstances beyond the individual's power to control. This latter thought sounds vaguely menacing. It is opposed by the plausible liberal idea that each individual is responsible for the quality of the fundamental aims that she affirms and for the consequences for the quality of her life that flow from her embrace of these aims and her pursuit of a plan of life based on them. The doctrine of responsibility for ends has the implication, which many find attractive, that even in principle, much less in practice, justice does not call for resource provision to individuals for the purpose of compensating them for their tastes, should their tastes happen to be expensive. If one individual is satisfied with popcorn and beer and another has a refined sensibility that is satisfied only with plover's eggs and prephylloxera claret, the doctrine of responsibility for ends holds that the person with expensive tastes has no claim for extra compensation by appeal to distributive justice principles.

The idea that each citizen must take responsibility for her final ends and for the plan of life she follows is closely linked to Rawls's decision to measure the condition of individuals for purposes of distributive justice by their resource holdings as measured by an index of primary social goods. Rawls articulates the rationale for primary social goods as follows:

> Justice as fairness [i.e., Rawls's doctrine] . . . does not look behind the use which persons make of the rights and opportunities available to them in order to measure, much less to maximize, the satisfactions they achieve. Nor does it try to evaluate the relative merits of different conceptions of the good. Instead, it is assumed that the members of society are rational persons able to adjust their conceptions of the good to their situation.[10]

The notion of rationality alluded to here is a range property: As long as one meets a minimal threshold of rationality, one is considered rational – period – and variations in rational capacity above the threshold do not dictate different treatment of different individuals in a Rawlsian scheme as far as the doctrine of responsibility for ends is concerned. The rough idea is that if one is nonfeebleminded and noncrazy, the soundness of one's conception of the good and the viability of one's plan of life are not questioned.

This may sound attractively liberal, but the consequence should be noted: If one assumes that at least to some extent and in some cases one can make objective determinations that some people's fundamental aims and life plans are defective and ruinous for their lives, the principled refusal to use this information as a basis for social policy will lead a Rawlsian just society to treat the predictably blighted lives of some of its least fortunate members as a matter beyond the scope of justice and not a legitimate social concern. This is individualism with a vengeance.

The claim then is that the principled refusal to look behind the uses that people make of their opportunities and liberties to see what quality of life they reach is unfair at least to those who predictably and through no fault of their own end up with avoidably unfortunate lives. I next consider several objections against this claim.

[10] Rawls, *A Theory of Justice*, p. 94.

Objection 1. One could avoid this individualism-with-a-vengeance result by setting the threshold of minimal rationality very high, but then much of social policy will be treated by principles of justice that are not Rawls's and one wants to know the content of these principles. At any rate, there is still the difficulty that by means of the threshold one is treating as an either/or a morally relevant factor that varies by degree.

Objection 2. Defending Rawls's doctrine of responsibility for ends, Norman Daniels writes that if individuals egregiously fail to be rational in their choice of fundamental life aims, the appropriate response by society is to provide medical care that will restore the individual's capacity for choice.[11] In this way, responsibility for ends is upheld: Individuals are responsible for their choices of final ends, provided they have a threshold capacity for choice, and if they lack the capacity, the just society owes them aid to restore the capacity, not compensation that restores them to some putatively fair level of satisfaction of their rational ends.

However, notice that there are moral costs to the resolve to stand by responsibility for ends come what may. First, providing aid that attempts to rehabilitate rational faculties may be in some cases an inefficient means of helping the individual attain a better quality of life understood as degree of fulfillment of choiceworthy ends. Insistence on responsibility for ends then means we help a badly off person less rather than more. Second, the Daniels version of responsibility for ends holds to be beyond the purview of social justice differences in the quality of the final ends that individuals affirm, no matter how large the differences, provided that the individuals are above the threshold standard of rationality. In some of these cases, the individuals with worse ends will have arrived at their ends by a process not reasonably deemed within their power to control. Inborn or socially acquired excessive susceptibility to cultural cues and insufficient reasoning power may be the factors that determine the differences in the quality of individual ends, not any blameworthy negligence or recklessness on the part of the choosing individuals. In this range of cases, the Rawls–Daniels position turns a blind eye to significant differences in life prospects among

[11] Norman Daniels, "Equality of What: Welfare, Resources, or Capabilities?" *Philosophy and Phenomenological Research* 50, Supplement (Fall, 1990): 273–296. A similar point is asserted by Christine Korsgaard in "Commentary on Amartya Sen's 'Capability and Well-Being' and Gerald Cohen's 'Equality of What? On Welfare, Goods, and Capabilities'," in *The Quality of Life*, ed. by Martha Nussbaum and Amartya Sen (Oxford: Oxford University Press, 1993), pp. 54–61.

individuals that cannot reasonably be deemed the responsibility of the individuals themselves. In practice, perhaps often little can be done to ameliorate these discrepancies, but in principle, the theory of justice (I claim) should register them.

Objection 3. An alternative response is to query an assumption that is implicit in the individualism-with-a-vengeance worry: That society as a whole can reach sufficient agreement on reasonable final ends to be able to base public policy on substantive claims about the good. One might hold that Rawls's primary reason for restricting the interpersonal comparisons for the theory of justice to differences in people's holdings of primary social goods is a sensible skepticism that society-wide reasonable agreement on worthwhile human ends and a correct conception of the good are possible. If the members of a diverse democracy cannot agree on the good, then something like the primary goods idea must be accepted.[12]

I doubt that the denial of the possibility of any reasonable agreement about what goals are worthy of pursuit, hence a blanket denial of interpersonal comparisons beyond comparisons of different person's holdings of income and other primary social goods, is consistent with any insistence that distributive justice requires compensation for disadvantage. If we really are faced with incommensurability of the good, such that we have no basis at all for asserting that a sick, destitute, and illiterate individual with few primary social goods is likely to be leading a worse life than a healthy, wealthy, and well-educated individual blessed with many primary social goods, then on what basis do we claim that redistribution between the worse-off and better-off person (as rated by the primary goods measure) is appropriate? After all, it would be fetishistic to care about lack of means unless lack of means can be known to bring about a lack of opportunity to achieve

[12] On the difficulty or perhaps impossibility of making interpersonal comparisons of well-being that can be employed in principles that determine the requirements of justice, Rawls's thinking appears to undergo evolution. In *A Theory of Justice*, he writes of the difficulties that afflict the making of interpersonal comparisons, "I do not assume, though, that a satisfactory solution to these problems is impossible." Rawls sees the issue of interpersonal comparison as bound up with the merits of utilitarianism as a theory of justice and observes that "the real difficulties with utilitarianism lie elsewhere." These quotations are from *A Theory of Justice*, p. 91. But in a later essay Rawls makes the basis of interpersonal comparisons central. See Rawls, "Social Unity and Primary Goods," in *Utilitarianism and Beyond*, ed. by Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982), pp. 159–185. A version of this same idea is crucial to the argument in Rawls, *Political Liberalism.*

worthwhile ends. Incommensurability implies agnosticism about what constitutes fair shares.

Even if two individuals happen to adopt exactly the same final ends, and they have unequal holdings of primary social goods, it is problematic to hold that it is morally important to get more primary social goods into the hands of those who have lesser shares. For one thing, for all that has been specified so far, it could be the case that the individual with a lesser amount of primary social goods might have a greater amount of nonprimary goods, so she can attain a higher level of satisfaction of her final ends then the person with the same final ends and more primary social goods. Why care about a subset of the means that people have to achieve their final goals? A second point to note is that if sets of final ends are incommensurable, then if the individual with fewer primary social goods had chosen a different conception of the good with a different and more easily satisfiable set of final ends, there would then be no basis for claiming that the one has a lesser prospect of attaining a satisfactory quality of life than the other – even if it were granted that if two persons have identical final ends and one has more primary social goods, the one with more primary social goods has a greater prospect of fulfilling these final ends. If sets of final ends are incommensurable, then the individual with lesser primary goods has it within her power by choosing a new set of final ends to bring it about that she does not have a lesser prospect of achieving a good life than the person with more such goods. Why care that someone has lesser means than another to achieve a shared set of goals if there is nothing especially normatively attractive about the pursuit of that set of goals as opposed to many others?

Objection 4. According to Rawls, the primary social goods idea relies on the assumption that individuals are "able to adjust their conceptions of the good to their situation."[13] If an individual is allotted a fair share of resources, it is up to her to adjust her life choices to achieve decent life prospects. But this adjustment process encompasses two different processes, only one of which is usefully described in the language of choice. Given a set of fundamental personal values, a person may choose a plan of life, a revisable set of goals that one then pursues in order to achieve one's values to the fullest possible extent. We expect that a person's life plan should adjust to the present and expectable circumstances of one's life. If I am a very poor peasant, my reasonable life plan may be

[13] Rawls, *A Theory of Justice*, p. 94.

limited to trying to keep my family alive so it can continue in the next generation. But it is not at all obvious that the individual's conception of the good, of what is valuable and choiceworthy in human life, should adjust to circumstances in this way. Why would my poverty affect the value of creating and appreciating great art and music? The idea of adjusting one's ends to one's circumstances makes sense only to the extent that what is at issue is the choice of a plan of life regarded as means to fulfill one's fundamental values. An individual's conception of values may be affected by his circumstances, but to think of one's values as formed by one's idiosyncratic circumstances is to think of them as partial and distorted: Insofar as being a peasant or a professor has given me access to some of the goods of life and not others, I should recognize my limited experience in forming my conception of the good life and try to offset it.

Of course idiosyncratic circumstances may favorably affect one's choice of values. But when this occurs, the circumstances are either working to improve the reasons and evidence on the basis of which one comes to affirm particular values or to improve one's abilities reasonably to incorporate evidence and reasons into one's reflections about values. What I am claiming does not make sense is the idea that having one rather than another set of limited experiences – eating fish but not fowl, reading books but not playing sports – can give one a better basis for making comparative assessments.

One chooses a plan of life, but not one's values, which are formed by belief and judgment. I can choose to engage in reflection, which may affect belief and judgment. I may engage in deliberation carefully or carelessly and be responsible for the degree of care taken as far as this lies within my power to control. But if I reflect, I cannot choose what conclusions I will be led to by reasons, and if I could so choose, the process of reflection would not be rational, controlled by the weight of reasons. There is a decisively passive aspect to the process of responding to reasons in forming beliefs. I cannot be responsible for my values in the way I am responsible for my choices.

3.5 Joint Responsibility on the Part of Individual and Society for Individual Ends

This discussion on responsibility for ends to this point might prompt the following response: We admit that it is not reasonable to hold individuals responsible for what does not lie within their control, and as an extension of this principle, it is not reasonable to hold two individuals equally responsible for what lies easily and costlessly within the control of the one and barely, at excruciating personal cost, within the control of the other. We then amend responsibility for ends as follows: Each individual should be held responsible for his choice of final ends insofar as this choice lies within his control. Moreover, the more difficult and costly it would be for a particular person to make a choice, the less one is fully responsible for that choice. But we affirm responsibility for ends subject to this proviso.

This last formulation also looks to be overly suspect and rigid. Imagine that Smith and Jones could have chosen their final ends differently and better, and it would not have been impossible or difficult for them to have done so. By the account just sketched, they are responsible for their choice of ends. Does this preclude the assumption of responsibility by society for the quality of the ends embraced by its members? Certainly it is possible that under the circumstances as sketched, the society could have altered the choice-forming environment in ways that would have increased the prospects of reasonable choice by Smith and Jones. In *On Liberty*, John Stuart Mill defends his proposed liberty principle, among other reasons, on the ground that a society that respects this principle will thereby provide an environment that is conducive to intelligent deliberation about goals and choice of life plans.[14] Here the metaphor of a division of responsibility between individual and society introduced by Rawls is potentially misleading.[15] That an individual in a particular context is responsible for her self-affecting choices in the sense that society will not compensate her for deficits in her well-being that result from those choices does not preclude the possibility that society is responsible for undertaking measures that will alter the environment in which choices are made that will predictably improve their quality. Moreover, if society fails to fulfill this obligation, it may incur an obligation to compensate those who suffer from this failure. This means that individuals might be responsible for their ends in the sense that the quality of the ends chosen lay within their power to control; yet, society might be responsible for compensating individuals for resultant low well-being because if society had done what it should, the deficient ends actually chosen would never have been selected.

[14] John Stuart Mill, *On Liberty*, ed. Elizabeth Rapaport (Indianapolis, IN: Hackett, 1978 [originally published 1859]), chapter 3.

[15] The notion of a division of responsibility between individual and society is advanced in Rawls, "Social Unity and Primary Goods," p. 170.

3.6 Effectively Equivalent Options

Suppose that Smith and Jones face crucial life decisions with large consequences for their expected well-being over the course of their lives. Each has available a prudent course of action that would guarantee a satisfactory outcome. To arrive at the prudent decision, one must solve a mathematical problem, which Jones can solve easily and which Smith can solve only by dint of great and costly effort that strains his faculties to the limit. Smith must reject many tempting options that would yield nice payoffs in the short run and disastrous payoffs in the long run to select the prudent option, whereas Jones faces no such tempting bad offers. Having made the prudent choice, Smith can carry it out only with great difficulty, and Jones can do it easily. To simplify, imagine that we can aggregate the factors that render prudent choice and action difficult or easy and painful or pleasant into a single scale of painful difficulty. We can then say that two agents facing different arrays of options have *equivalent* options if the well-being each would gain by acting perfectly prudently is the same and *effectively equivalent* options if making and implementing this perfectly prudent choice would be equally painful and difficult for each. One suggestion then is that individuals can reasonably be held responsible for their choices among options by comparison with the choices of other individuals who faced effectively equivalent options. Another suggestion is that to the extent that the difficulty and pain of making the prudent choice exceeds a level deemed tolerable, the individual's responsibility is mitigated in case she chooses and acts imprudently. In other words, we hold an individual responsible for doing as well as could reasonably be expected in her circumstances, given the value of the options available to her and the difficulty and pain of making and implementing the choice to do what she ought, given her circumstances.

I don't take this approach to responsibility to raise the free will issue. Even if one assumes that individuals have free will to make choices, the agent's native traits and talents influence the choices available to her in given circumstances and the difficulty and cost of determining and making the best choice. If, however, determinism holds, then either soft determinism obtains, in which case the suggested analysis still applies, or hard determinism obtains, in which case all questions of responsibility are moot.

However, another worry presses for attention.[16] It might be supposed that making the assumption that all members of society are fully rational agents

[16] I thank Wayne Martin and Philip Kitcher for pressing this concern.

expresses a normative commitment to treat all human beings as persons worthy of respect. This claim does not have the status of a weak empirical presumption to be adjusted continuously case by case in the light of the available evidence. We give up this claim only when forced to do so by confrontation with disabling mental illness or feeblemindedness. Short of that, we express respect for persons by treating every member of society as a fully rational agent, capable of appreciating and understanding the import of good reasons and capable of being moved to action by good reasons. Any other attitude denies respect for persons and licenses the treatment of individuals as objects to be manipulated in the service of ends that we suppose to be worthy but which the manipulated beings may not share.[17]

Various issues are surfacing here, most of which I must let sink back to the bottom of the pond. For present purposes, I simply want to register where I begin to disagree with the reflections of the previous paragraph. The problem starts with the slogan of "respect for persons." Whatever respect for persons entails, if the idea is to be acceptable it can require neither the denial of known empirical facts nor the treatment of people as though what's true were not true. People do differ in their capacities to appreciate reasons and in their susceptibility to be moved by them. These differences matter in everyday affairs, not just in the neighborhood of extremes of pathology. Often the pertinent facts are highly uncertain, and in virtue of the pervasive uncertainty, the choice of policy for coping with the variability in rationality across persons must be tentative and cautious. But if you know that I am incompetent in certain ways in some domain of policy making, it would not be disrespectful to take measures to cope with my incompetence, and perhaps to insulate me from decision-making responsibility in this domain, when policy choice is consequential for the well-being of other persons or myself. (I note that no elitist policy conclusions flow immediately from the remarks at this level of abstraction. Bentham's enthusiasm for Panopticon managerialism needs to be tempered by Mill's sober doubts concerning *quis custodiet custodies?* among other questions.)

3.7 Are We Responsible At Most for What Lies Within Our Control?

In this chapter, my starting point is the limiting principle that we should be held responsible at most for what lies within our power to control. I then

[17] This paragraph is an attempt to construe remarks in Rawls, *Political Liberalism*, pp. 178–187, esp. pp. 184–186. For an account of the moral import of Kant's analyses of rational agency and human freedom, see Henry E. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990); see esp. chapter 6.

amend this principle by noting that even if it lies within my power to secure an outcome, it may barely be within my power, so even if I could and should secure it, it may be unreasonable to hold me responsible for failing to secure it. In contrast, securing a similar outcome of similar importance may be easy for you, so if we both succeed in bringing the good outcome about, I should get more credit than you; if we both fail, I should be blamed less than you; and if one of us succeeds and the other fails, how much credit and blame should be assigned depends on which of us succeeded and which failed.

This account might be resisted at the outset by the denial that one should be held responsible at most only for what lies within one's power to control. In many situations, individuals assume responsibility for the quality of outcomes that may vary depending on factors beyond their power to control. In these scenarios, there is evidently voluntary control at one remove. But we might also envisage an assignment of responsibility for outcomes imposed on people without any mediation of voluntary choice. For example, a society might adopt a no-fault compensation scheme for automobile accidents. Under this scheme, everyone must purchase auto accident insurance, and when accidents occur, compensation is paid from the insurance fund to those who suffer losses, regardless of the faultiness of their conduct. Suppose the no-fault scheme is in place, and Smith and Jones, both dead drunk, recklessly cause an expensive accident. Responsibility for these losses is borne by all the members of society who are required to purchase the insurance, which pays for the costs that Smith and Jones incur. Here there is responsibility for outcomes beyond the responsible agent's power to control, and this responsibility is not incurred by voluntary choice. Whether this is fair depends on the system as a whole and its consequences for people's lives as it operates over time. One might insist that you cannot validly object to the scheme just by reciting the slogan, "No responsibility for outcomes that are beyond the individual's power to control."

To sort out these concerns, we must distinguish different senses in which an individual might be said to be responsible for the quality of some outcome. One might be responsible for an outcome in the sense of liable to praise or blame, reward or punishment, depending on the quality of the outcome. I take it that we should only be held responsible at most for what lies within our power to control. The no-fault insurance example does not challenge this claim.

An individual might be said to be responsible to some extent for an outcome just in case one will be required to pay some of its costs if the outcome falls below some threshold level of quality. One is fully responsible for negative outcomes if one is to bear all of the costs. (One might be responsible for positive outcomes as well, in which case one shares the gains.) Responsibility in this sense of liability to pay costs might sensibly be divorced from control.

Any theory of distributive justice which holds that society – all of us taken together – is obligated to compensate individuals for misfortune with a view to assuring everyone a fair share of opportunity for a good life necessarily assigns individuals responsibility in the cost-sharing sense for outcomes that are beyond their power to control. If a childhood disease epidemic places many individuals at a disadvantage unless they receive help that compensates for the disabling residue of the disease, then justice may dictate that the rest of us are obligated to provide this help, which means that we are responsible for sharing the costs of outcomes of disease that are clearly beyond our (i.e., the taxpayers') power to control.

The obligation of society to share the costs and benefits of good and bad luck by providing fair shares of opportunity to all corresponds to a right of each individual to receive a fair share of opportunities. My claim about personal responsibility as a determinant of fair shares to this point has been that one's fair share of opportunities is the share that would give one a fair share of human good or well-being if one used one's opportunities as prudently as could reasonably be expected, given one's unchosen circumstances and personal traits and talents. If one has received a fair share in this sense, deficits in well-being that arise from deficiencies in the way one has lived do not trigger further obligations on the part of society to compensate the individual so as to erase the deficits. Personal responsibility sets limits to morally desirable equalizing compensation done in the name of distributive justice.

The objection to this account is that we might conceive of ideas of personal responsibility merely as means to achieve other justice values. Viewed as a means in this way, a norm of responsibility might fail to respect the idea that one should be held responsible at most for what lies within one's power to control. To revert to the no-fault insurance scheme, one might justify an assignment of responsibilities to individuals beyond what lies within their power to control by the morally desirable consequences that the assignment brings about. To see matters in this way is to see the assignment of responsibility as political, not metaphysical.

The objection misfires. At least, the possibility of treating responsibility assignments as means to further goals does not at all preclude viewing aspects of responsibility assignments as intrinsically morally desirable. That responsibility assignments have instrumental value does not render them mere means. Once individuals have received a fair share of opportunities, it is morally better, other things being equal, that those who are truly responsible for faulty conduct that renders themselves or other persons (who have not consented to share these losses) worse off should pay for the consequences of such conduct. Of course, there may be costs to tailoring individual fortune to the quality of responsibility of their conduct, and sometimes these costs will outweigh the moral desirability of tailoring. This consideration has no power to undermine the claim that it is morally desirable for its own sake that fine-grained judgments of individual responsibility should affect what society owes the individual by way of opportunity provision over her life course.

The no-fault insurance scheme proposal illustrates the point. Perhaps the adoption of this scheme generates savings in administrative costs, which render everyone better off than they would be under alternative feasible schemes. This in no way denies that it is intrinsically more fair that if individuals harm themselves by faulty conduct and if a fine-grained theory of responsibility does not excuse their conduct but holds them fully accountable for it, the individuals themselves, and no one else, should absorb the costs of the harm. The assignment of responsibility in the sense of liability to bear costs is evidently both a means to other justice values and a way of apportioning responsibility fairly. A full theory of justice must give guidance on how we should balance these different fairness values when they conflict in particular cases.

3.8 The Hybrid Proposal

Suppose we resurrect the canonical moment idea and combine it with a standard of interpersonal comparison that looks beyond resource provision to the quality of life that individuals are enabled to achieve by given resources. The hybrid proposal is the opportunity for well-being conception.[18] According to it, two individuals enjoy the same opportunity for well-being just in case, at the onset of adulthood, resources have been allotted so that each faces an array of effectively equivalent life options in the sense that if each chooses as prudently as could reasonably be expected, each would have the same lifetime expectation of well-being.[19] (The notion of *well-being* here is a placeholder for whatever theory of human good is best.) This suggestion

[18] This is the view I advanced in "Equality and Equal Opportunity for Welfare," *Philosophical Studies* 56 (May 1989): 77–93.

[19] Equal opportunity for welfare so defined cannot always be implemented, as Marc Fleurbaey notes in "Equal Opportunity or Equal Social Outcome?" *Economics and Philosophy* 11 (1995): 25–55. When equal opportunity for welfare cannot be fully implemented, we need a measure that allows us to say, given two distributions of opportunities across a set of persons, which distribution comes closer to implementing this ideal.

is not subject to the two objections that plagued Rawls's view: That we are holding individuals responsible for matters beyond their power to control, and that we are misfocusing attention on resource holdings rather than on the extent to which an individual's resource holdings enable her to achieve a tolerable prospect of a good life.

The hybrid proposal resolves the problem of expensive tastes as follows: A distinction is made between an expensive taste that arises in a way that is reasonably held to be the responsibility of the individual who acquires the taste and expensive tastes for which it is not reasonable to hold the individual responsible. In principle, the latter are compensable. The expensive tastes problem also strongly suggests that mere satisfaction of an individual's basic preferences as such need not contribute much if anything to the choiceworthiness of her life. To some, popcorn and beer and plover's eggs and fine claret might appear equally to be frivolities. The individual's preferences, expensive or cheap, might not track what is reasonably deemed good for that person. The response to this aspect of the problem would explore the theory of the good. If the best account of human well-being does not identify it with satisfaction of actual preferences, then an oblique reply to the expensive tastes problem is available. The issue for distributive justice is not whether the person is enabled by his resource share to satisfy his tastes, be they expensive or cheap. The issue is whether the individual's resource share in the context of society's overall dealings with the individual provide her with a fair opportunity to achieve a good, valuable, choiceworthy life.

Does this hybrid position successfully integrate the considerations that unraveled the Rawls–Dworkin approach to individual responsibility within distributive justice?

3.9 Bert's Case

No. No doubt the hybrid proposal on responsibility contains multiple errors, but two are flagrant.[20] One error is that this approach to responsibility is too

[20] Good critical discussions of the hybrid proposal are in John E. Roemer, *Theories of Distributive Justice* (Cambridge: Cambridge University Press, 1996), chapter 8; Norman Daniels, "Equality of What? Welfare, Resources, or Capabilities?"; Thomas Christiano, "Difficulties with the Principle of Equal Opportunity for Welfare," *Philosophical Studies* 62, no. 2 (May, 1991): 179–185; Eric Rakowski, *Equal Justice* (Oxford: Oxford University Press, 1991), pp. 43–72. On the rationale of the family of views of which the hybrid proposal is a member, see G. A. Cohen "On the Currency of Egalitarian Justice," *Ethics* 99, no. 4 (July 1989): 906–944. See also Richard Arneson, "A Defense of Equal Opportunity for Welfare," *Philosophical Studies* 62, no. 2 (May 1991): 187–195; also Richard Arneson, "Property Rights in Persons," *Social Philosophy and Policy* 9, no. 2 (Winter 1992): 201–230.

unforgiving. A second error is that if we compensate for unchosen bad luck before the canonical moment, why ignore unchosen bad luck that occurs after it? Both errors are illustrated by Bert's case, posed by Marc Fleurbaey.[21] Starting with an allotment of opportunities at the canonical moment that is ex hypothesi fair, Bert squanders his resources by his own carelessly voluntary choice. He deliberately chooses to ride a motorcycle at high speed without protective headgear just for the thrill of the experience on a deserted road (so nobody is put at risk except himself), and without having purchased any accident insurance, even though he concedes the risk of accident is excessively high by comparison with the expected gains from speeding. In the event, he suffers an accident and is grievously injured. He is personally responsible for his plight, which has come about as a result of his heedlessly reckless choice. However, once he is injured he could be restored to normal health if society pays for brain surgery costing $10,000. Without this surgery he will swiftly degenerate into an irremediable vegetative state. Given that he is already the recipient of a fair share of opportunities, to provide him with the operation he needs would be to bestow on him an unfairly large set of opportunities – if he had a fair share, and he is now given extra resources, he gets more than what is fair. Nevertheless, it seems harsh to deny Bert his life-restoring operation. Bert behaves worse than could reasonably be expected of him. His behavior is faulty on a fine-grained conception of responsibility. Still, we should help him, I assume.

I assume, and do not here argue for, the "we should help him" response to Bert's case. Some might think that helping Bert at this point is required by charity, not justice. But there is a possibility of merely terminological disagreement here. I use *distributive justice* as a name for obligations to compensate fellow members of society for certain types of bad luck, these obligations being regarded as appropriately enforceable.

Some factors that may influence the response to Bert's case:

Opportunity Provision versus Maximal Utility. The description of Bert's case strongly suggests that offering Bert the resources he needs for the operation that would restore him to good health would be a very efficient use of resources to increase the sum total of human good. The strength of this consideration can be checked by varying the example in thought. We can imagine variants of Bert's case that are changed in only one respect: the cost-to-benefit ratio of giving Bert extra help becomes increasingly unfavorable.

[21] Fleurbaey, "Equal Opportunity or Equal Social Outcome?" pp. 25–55.

Initial Opportunities and Subsequent Bad Luck. After being allotted a set of resources that is supposed to give him a fair share of opportunities for well-being, Bert then chooses a course of life, experiences bad luck, and ends up with very low well-being despite initially bountiful resource provision. In Bert's case, he has bad luck in the course of following a poor plan of life, but bad luck could befall anyone who starts with a given set of opportunities, regardless of the quality of the life plan she chooses. Again, we can check the influence of the bad luck factor in our response to Bert's predicament by imagining otherwise similar variants of the example in which the bad luck lessens and then disappears.

Deservingness. In the example, Bert behaves imprudently and comes to harm through his own fault, but the "punishment" he receives is disproportionate to his "crime." Life is punishing Bert very severely for slight fault. We can bring this feature of the situation into relief by exaggerating it. Or we can imagine variants of the case in which Bert's negative deservingness increases and the ratio of his punishment to his crime diminishes as a result.

Priority to the Badly Off. Once Bert is injured, his life prospects absent any further aid are truly dismal. This factor may itself strengthen the case for helping him.

Efficiency. In the example as described, the resources that we could give to Bert have alternative uses. If we do not help Bert, we could help someone else. We might try to gauge the importance of this factor by imagining it altered. Suppose that the resources we could give to Bert have no alternative uses. We could help Bert or no one.

Does Bert's case indicate that distributive justice should be fundamentally concerned with the life outcomes that individuals actually reach rather than the opportunities they enjoy? Is provision of opportunities at most instrumentally morally valuable and not morally valuable for its own sake? Maybe one is just barking up the wrong tree when one tries to specify the content of distributive justice by articulating an ideal of fair provision of opportunities. However, the issue is still open.

The possibility of pointless opportunity provision might be thought to illustrate the futility of trying to devise principles of distributive justice according to which justice is some function of opportunity provision. Suppose that Smith and Jones live on separate islands and that Smith's resources are ample and Jones's resources are skimpy. Let's stipulate that Smith can improve Jones's opportunity to lead a good life in just one way, by constructing a raft and setting some of his goods adrift on the raft to be carried by the tides to the shores of Jones's island. On the facts so far stipulated, let's say that justice requires that Smith help Jones. But suppose with certainty that if she sends aid to Jones, the aid will do no good and not help him further any of his goals. Perhaps Jones is clumsy and neglectful and will certainly entirely waste the resources; perhaps Jones has firm religious scruples against using resources that wash ashore on her island. On outcome-oriented principles, Smith's obligation to aid will evaporate in these circumstances. But it might seem that on opportunity-oriented principles, Smith's obligation remains in force. After all, the opportunities are just as good, and just as available to Jones, whether she uses, neglects, or squanders them. If opportunity provision is what fundamentally matters from the standpoint of distributive justice, nothing cancels the obligation to aid. Denying this might be thought tantamount to rejection of opportunity-oriented views of distributive justice; however, the conclusion is premature. At most, the example suggests that the pointless provision of opportunities is not required by justice. Justice is not indifferent to outcomes, we might say, regardless of how the outcomes are produced.

Suppose that after what provisionally seems a fair initial distribution of opportunities, Amanda freely and rationally chooses a course of life that involves a certain sacrifice of her prospects for well-being in order to aid a worthy cause of her choice. Here, as in Bert's case, an initial distribution of opportunities thought to be fair is followed by an imprudent choice by the agent leading to subsequent dismal life prospects. But here, unlike in Bert's case, Amanda's choice (I claim) does not give rise to a case for further redistribution of resources to improve her expectation of personal well-being. A similar judgment (I claim) is appropriate when Cheryl freely and rationally enters into high-stakes gambling immediately after receipt of her canonical moment of fair distribution of resources and emerges the loser of the gamble, with poor prospects for lifetime well-being. Here is a partial characterization of a conception of fair opportunity for well-being that accords with these tentative judgments:

- The measure of interpersonal comparison for distributive justice is the effective opportunity for well-being for the agent that a set of resources provides, the amount of well-being that the resources would provide if the agent conducted herself as prudently as could reasonably be expected in her circumstances.
- Distributive justice requires that resources be set so that at the onset of adulthood each agent faces an array of options that provides

an effective opportunity for well-being such that, for all agents, a function of effective opportunity for well-being is maximized that gives priority to providing gains in well-being to those with less.

- A free and rational choice by an agent to bring about an outcome that provides a low level of well-being for the agent does not bring it about that justice requires further compensation to the agent to increase her well-being.
- A free and rational choice by an agent to undergo a lottery, provided the agent selects it from a set of options that includes acceptable options that would not involve incurring comparable risk, does not bring it about that justice requires further compensation to the agent in the event that the outcome of the lottery is disadvantageous to her.
- Less than fully rational choices by agents may trigger a justice requirement of further compensation to them for misfortune they suffer depending on how faulty their conduct is, fault being assessed according to a fine-grained theory of responsibility.

3.10 Two Rawlsian Rejoinders

Rawls can deploy two powerful rejoinders to this line of thought. One is that the theory of justice must limit its concerns to matters that could feasibly be administered in modern democratic society. But the ideas of individual deservingness and responsibility and individual well-being, even if they could be made clear in principle, cannot conceivably be measured by any institutions we could devise. Since the theory of justice is for men and women, not for angels or for Gods, these indeterminable moral qualities are irrelevant to justice.

We need to know what matters to us morally for its own sake before we can begin to address in a sensible way the issue of how to achieve what matters to the greatest possible extent, given the epistemic and other practical constraints of life as we know it. No doubt the theory of justice is many levels of abstraction removed from the sphere of practical policy determination, but we cannot decide on appropriate proxy measures for the unmeasurable qualities we really care about until we decide what we really care about. At this stage in our inquiry, the appeal to the constraints of feasibility is premature.

Rawls's second powerful rejoinder is that the theory of justice seeks a consensus on fair terms of cooperation that can include all reasonable persons under conditions of pluralism of belief. Pluralism means that reasonable individuals will tend to affirm different and opposed comprehensive conceptions of the good. These opposed conceptions will specify inter alia different and opposed views of human well-being and of human responsibility and deservingness. We simply have to agree to disagree about these matters. To try to base a theory of distributive justice on some particular comprehensive conception of the good is inevitably sectarian and thwarts the aspiration to reasonable consensus. (See also objection 3 and the reply to it in the text.)

The Rawlsian approach to the problem of interpersonal comparison for a theory of justice presumes from the outset a fundamental epistemic asymmetry between ideas of the good and ideas of the right. We have no reason to accept this asymmetry. The ideal coherence test that Rawls proposed and that many others endorse for determining what ethical claims are acceptable does not suggest a reason for supposing that reasoned agreement about the good cannot form part of the moral consensus of a just society. No doubt we face difficult problems of partial commensurability in both domains; that's life.[22]

[22] I wrote this essay in 1995. For my recent thinking on this and related topics, see Richard Arneson, "Desert and Equality," in *Egalitarianism: New Essays on the Nature and Value of Equality*, ed. by Nils Holtug and Kasper Lippert-Rasmussen (Oxford: Oxford University Press, 2007), pp. 262–293; Arneson, "Luck Egalitarianism Interpreted and Defended," *Philosophical Topics* 32, nos. 1 and 2 (2004), 1–20; Arneson, "Justice after Rawls," in *Handbook of Political Theory*, ed. by John Dryzek, Bonnie Honig, and Anne Phillips (Oxford: Oxford University Press, 2006), pp. 45–64; Arneson, "Cracked Foundations of Liberal Equality," in *Ronald Dworkin and His Critics*, ed. by Justine Burley (Oxford: Basil Blackwell, 2004), pp. 79–98; Arneson, "Luck and Equality," *Proceedings of the Aristotelian Society*, supp. vol. 75 (2001), 73–90.

4

Improving Our Ethical Beliefs

James Griffin

I want to raise a subject associated with one of our two distinguished honorands, John Rawls. A proposal of his that has gained almost universal acceptance in the philosophical community is that the way to test and to strengthen ethical beliefs is to bring them into wide reflective equilibrium. It is not that that seems to me wrong. It is just that it has always seemed to me to say so very little. That is the thought I want to develop.[1]

4.1 Piecemeal Appeal to Intuition

Among philosophers, the most common sort of criticism of our ethical standards nowadays is what can be called *piecemeal appeal to intuition*. We are all familiar with how it goes: It follows from your view that it would be all right to do such-and-such, but that's counterintuitive, so your view must be wrong.

Philosophers now pretty much agree that, as criticism, piecemeal appeal to intuition is weak, though, for lack of anything stronger, their belief has not had the revolutionary effect on their practice that one could reasonably expect to follow.

It is not that piecemeal appeal to intuition shows nothing. It is just that the doubts that it raises are not very strong.[2] It may well be that some intuitions

[1] This chapter is adapted with the permission of Oxford University Press from chapter 1 of my book *Value Judgement* (Oxford: Clarendon Press, 1996). Copyright © James Griffin 1996. For Rawls's view, see his *A Theory of Justice* (Oxford: Clarendon Press, 1972), sects. 4, 9; see also his "Outline of a Decision Procedure for Ethics," *Philosophical Review* 60 (1951): 177–197; "The Independence of Moral Theory," *Proceedings and Addresses of the American Philosophical Association* 48 (1974–5): 5–22, esp. sect. 2; "Kantian Constructivism in Moral Theory," *Journal of Philosophy* 77 (1980): 515–572.

[2] The case against piecemeal appeal to intuition has been made powerfully by others, particularly by R. M. Hare and R. B. Brandt. See Hare, "The Argument from Received Opinion," in his *Essays on Philosophical Method* (London: Macmillan, 1971); also his *Moral Thinking*

are as close to sound moral beliefs as we shall ever get. Others, however, clearly are not, and there are no internal marks distinguishing the first lot from the second. Intuitions, despite the misleading suggestion in their name of a special sort of perception into moral reality, are just beliefs. Some of those beliefs have been drummed into us in our youth by authority figures and are no more reliable than those figures were. Some are social taboos that, if we understood their origin, we would see are now obsolete. Some are edicts of the perhaps unfortunate superego that emerged from our private battle with our own aggression. And so on. What slight knowledge we have of the origins of our moral beliefs hardly leads us to grant them, as a kind, much authority.[3] Causal explanations are not equally corrosive, of course. Some leave us hesitant when before we were confident; some make us drop what before we held; some actually strengthen our beliefs. For the most part, though, we simply do not know the causes of our intuitions. Even perfectly natural, nearly universally distributed sentiments and attitudes may not be in order as they are. For instance, it is natural – indeed characteristic of human nature generally – for our sympathies to be warmly engaged by identifiable persons whose lives are at risk and not by merely statistical lives. However, it is not at all clear that governments are right to spend, as they usually do, far more on saving one missing yachtsman than it would take to save dozens of unknown lives through wider detection of cancer. And moral sentiments, attitudes, and beliefs are – like certain observations of supposedly brute facts – theory-laden, probably much more laden with theory than such observations, and the theory can be poor. It is no panacea even to take the most optimistic view about the soundness of some of our

(Oxford: Clarendon Press, 1981), ch. 8. See Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), ch. 1.

[3] On, e.g., psychological causes, see S. Freud, *Civilization and Its Discontents* (London: Hogarth Press, 1957), p. 138:

> Ethics must be regarded . . . as a therapeutic effort: as an endeavour to achieve something through the standards imposed by the super-ego which had not been attained by the work of civilization in other ways. We already know – it is what we have been discussing – that the question is how to dislodge the greatest obstacle to civilization, the constitutional tendency in men to aggressions against one another.

Freud ought to have inserted "in part" between "regarded" and "as a therapeutic effort." But he must have identified here an important cause for why one person inclines to one set of moral views and another to a different set. I suspect that we often find, through the workings of a mechanism of compensation, an overly strict superego associated with relatively unstrict moral intuitions and vice versa. The moral views of any reflective person will be shaped by vastly more sorts of causes than just the ones Freud mentioned. For an example of such psychological speculation, see E. Westermarck, *Ethical Relativity* (London: Kegan Paul, Trench, Trubner, 1932), chs. 8, 9.

intuitions. Even if some are indeed perceptions of moral reality, we have also to wonder whether they are more than the merest glimpse of a fragment of reality, and whether therefore if reality's whole contour were eventually revealed, we might view the fragment quite differently.

That, in very summary form, is the powerful negative case. It prompts the question, Should not the role that intuitions play in moral philosophy be no more nor less than the role that we are content to let them play in other branches of thought, in mathematics, in the natural sciences, and in other parts of philosophy? For instance, Russell's theory of types has strikingly counterintuitive consequences for Boolean class algebra and the definition of numbers. Since the theory restricts a class to members only of one type, it has the result that Boolean algebra can no longer be applied across classes but has to be reproduced within each type, and furthermore that numbers, defined on the basis of certain logical concepts, have similarly to be reduplicated for each type – consequences that W. V. Quine once condemned as "intuitively repugnant."[4] But no logician takes such repugnance as closing an argument. On the contrary, intuitive repugnance is just a spur to start looking for a good argument. Even if the role that intuition should play in ethics is not entirely the same as in other branches of thought – I shall come to that matter in a moment – what I have called the negative case at least reinforces the view that piecemeal appeal to intuition is weak argument. It gives intuitions more weight than they deserve. It is especially in ethics that intuitions have

[4] W. V. Quine, *From a Logical Point of View* (Cambridge, MA: Harvard University Press, 1953), pp. 91 ff. On intuition in the natural sciences, see e.g., W. Newton-Smith, *The Rationality of Science* (London: Routledge and Kegan Paul, 1981), pp. 197, 212–213. On intuition in philosophy, see, e.g., R. Nozick, *Philosophical Explanations* (Oxford: Clarendon Press, 1981), p. 546; R. Rorty, *Philosophy and the Mirror of Nature* (Oxford: Blackwell, 1980), p. 34. We find what seems to be the right sort of ambivalence about intuitions, the right mixture of scepticism and respect, much more commonly in these other departments of thought. On the side of respect, see Jaako Hintikka's Introduction to Jaako Hintikka, ed., *The Philosophy of Mathematics* (Oxford: Oxford University Press, 1967), p. 3:

> An intriguing aspect of the completeness and incompleteness results is that one of their starting-points (viz. our concept of what constitutes completeness) is inevitably an idea which can perhaps be formulated in naive set-theoretic terms but which either is not formulated axiomatically to begin with or which (in the case of incompleteness) cannot even possibly be so formulated. Yet concepts of this kind are most interesting. We seem to have many clear intuitions concerning them, and it is important to develop ways of handling them.

On the side of scepticism, see Daniel Dennett's and Douglas Hofstadter's complaint about the "intuition pump," the use of one sort of example to push our intuitions in a particular direction (say, in a debate about whether computers think), in Hofstadter and Dennett, eds., *The Mind's I* (Harmondsworth: Penguin Books, 1982), pp. 375, 459.

risen so far above their epistemological station. That may be for the reason I mentioned earlier: Where on earth are better arguments going to come from?

4.2 Purist Views

A brave response to the inadequacy of piecemeal appeal to intuition is to become a purist, that is, to forswear all dependence on substantive moral beliefs and to try instead to derive such beliefs from considerations untainted by moral element. But is that possible? To save time, let me just assert that I do not know of any form of purism that works.

Kant is the most famous purist, at least on a common but disputed reading of him.[5] A widespread view, which seems to me right, is that if Kant is read as appealing only to a thin enough conception of rationality to count as a purist, he does not succeed in deriving substantive moral conclusions; and that if he is read as enriching his conception of rationality enough for it to yield some substantive moral conclusions, then he is not a purist. Either way, we cannot point to Kant to show that purism is a live option.

There are modern purists too, R. M. Hare[6] and R. B. Brandt[7] prominent among them, but, to my mind, their forms of purism are too ambitious. I doubt that one can derive substantive moral principles from the logic of key moral terms (Hare) without the help of *some* substantive ethical beliefs. I doubt too that one can choose a reforming definition of ethics that will reduce it to a manageable factual project (Brandt) without *some* substantive ethical beliefs guiding the choice.

[5] I. Kant, *Groundwork of the Metaphysic of Morals*, esp. sect. 2. I explain my views about Kant's categorical imperative and its ethical content somewhat more fully in my book, *Well-Being* (Oxford: Clarendon Press, 1986), ch. 10, sect. 4. For a corrective to this common but oversimplified reading of Kant see, e.g., Onora O'Neill, *Constructions of Reason* (Cambridge: Cambridge University Press, 1989).

[6] Hare hopes to derive ethical principles entirely from the semantics of key moral terms, not local, culture-bound terms like *chastity* or *humility* but global terms like *good* and *ought*. See Hare, *The Language of Morals* (Oxford: Clarendon Press, 1952), ch. 11; *Freedom and Reason* (Oxford: Clarendon Press, 1963), chs. 2, 3, 6, 7; and *Moral Thinking* (Oxford: Clarendon Press, 1981), chs. 1, 2.5, 4.1ff.

[7] The key moral terms of ordinary language are, Brandt thinks, too vague as they stand to allow definite results. See Brandt, *A Theory of the Good and the Right*, ch. 1, and his further thoughts in "The Explanation of Moral Language", in *Morality, Reason and Truth*, ed. D. Copp and D. Zimmerman (Totowa, NJ: Rowman and Allanheld, 1985), pp. 104–119. His solution is for us to adopt a more normative approach, namely, *reforming* definitions. Preempt the term *rational*, he suggests, for *survives maximal criticism by facts and logic*. For our traditional question, What is morally right? substitute the factual question, What is allowed by any moral code that rational persons would want for a society in which they were to spend their lives?

4.3 Have We Been too Hard on Intuitions?

Intuitions may not have great authority in ethics, but that does not suggest that they have no more authority than they do in mathematics or the natural sciences. Part of what ethics seeks to express may be our self-understanding, our characteristic human sense of what matters. In ethics, we ourselves, as we characteristically are, may be one of the central subjects of attention. This suspicion would be reinforced if it turned out, as I think it does,[8] that both reason and characteristic human desire have an important role in giving content to values. This point is related to a point about the human sciences: in the human sciences we are not only interpreting the world – but interpreting that part of it that includes centrally our interpretations of the world.[9] Ethical standards aim partly at giving expression to our sense of what matters, so one would expect the content of ethics already to be embedded in our intuitions – not necessarily in all of them but in some, not necessarily in undistorted form but in some form or other. One would expect ethical standards to display closer connections to our ordinary ethical thought, to our intuitions, than scientific laws need have to our intuitions about the natural world.

Perhaps therefore intuitions should be seen as commonsense beliefs. Some of them will no doubt be faulty, but there may be a core of them that form the unavoidable framework for all our thought. After all, there have been defences of common sense in the case of beliefs about the external world. Why not a similar defence of common sense in ethics?

One plausible defence goes like this. A word has meaning only in virtue of there being rules for its use, rules that settle whether the word is correctly or incorrectly used. And Wittgenstein argues that the rules cannot, in the end, be satisfactorily understood as a mental standard – an image, say, or an articulable formula – but only as part of shared practices. And these shared practices are possible only because of the human beliefs, interests, dispositions, sense of importance, and so on that go to make up what Wittgenstein calls a "form of life".[10] Our form of life provides the setting in which our language develops and only within which its intelligibility is

[8] See my *Value Judgement*, chs. 2–4.

[9] This is what Anthony Giddens has called the *double hermeneutic* feature of the human sciences; see his *Studies in Social and Political Theory* (London: Hutchinson, 1977), p. 12, and his *New Rules of Sociological Method*, 2nd ed. (Cambridge: Polity Press, 1993), introduction to 2nd ed. and conclusion.

[10] See L. Wittgenstein, *Philosophical Investigations*, *passim* but esp. sects. 1–38, 136–156, 167–238; *Zettel*, sects. 338–91. For references to "form of life," see *Philosophical Investigations*, sects. 19, 23, 241; *On Certainty*, sects. 358–359, 559.

possible. And a form of life seems to consist in part in a certain shared set of values. Donald Davidson has a similar argument. We cannot, he thinks, interpret the language that others use without assuming that we have certain basic beliefs and attitudes in common with them – that, for instance, many of our aims, interests, desires, and concerns are the same.[11] If that is right, then general skepticism about commonsense values is self-defeating. The values are embodied in the language we use, which sets for us the bounds of intelligibility.

There is force to these arguments of Wittgenstein and Davidson; the difficulty is to say how far they take us. How many such basic beliefs are there? How much can we mine from them? I shall come back to that later.

4.4 The Coherence Theory

Even so, the negative case against intuitions as a class still stands. So if we are to use intuitions to criticize our ethical beliefs, we shall have an altogether more powerful form of criticism if we can find a way of using intuitions critically, if we can sort the better from the worse. Many people think, John Rawls most prominent among them, that we do that by making our beliefs coherent.

In the natural sciences, we cannot test an hypothesis by seeing whether it squares with pure observation. Observation is not pure in the sense needed; our observations are themselves theory-laden. In the case of a conflict between hypothesis and report of observation, therefore, sometimes the one and sometimes the other should give way. We have to be prepared to adjust each, going back and forth from theory to report, until the set of our beliefs reaches some sort of equilibrium. This procedure is not confined to the natural sciences; it also plays an important part in mathematics and logic. Axiomatic systems face the problem of showing that the axioms themselves are sound. If there can be no doubt about them, if they are, say, self-evident, then one has got a genuinely foundational form of justification: One can

[11] See, e.g., Donald Davidson, "Psychology as Philosophy," p. 237, and "Mental Events," p. 222, both in his *Essays on Actions and Events* (Oxford: Clarendon Press, 1980); also his 1975 Lindley Lecture, "Some Confusions about Subjectivity," in *Freedom and Morality*, ed. John Bricke (Kansas: University of Kansas Press, 1976), pp. 191–208. I slur over the differences between Wittgenstein and Davidson. Wittgenstein's notion of a "form of life" seems to consider local practices as well as universal human features. Davidson's truth-condition semantics locates meaning in the match between sentences and their truth conditions, and the structure of the match between my sentences and some stranger's sentences can occur independently of local practices.

justify certain beliefs by deducing them from sound fundamental beliefs. But in at least much logic and mathematics, the starting points are not beyond doubt. As the system of belief develops, pressures can build up to amend the starting point rather than give up too much of the body of our beliefs. One might even find, in developing theories of meaning and truth, pressures building up to abandon, say, the law of excluded middle.[12] This sort of holism has claims to be the deepest form of rational procedure in all areas of thought.

The best procedure for ethics, it is plausible to think, is a similar one of going back and forth between intuitions about fairly specific situations on the one side and the fairly general principles that we use to try to make sense of our moral practice on the other, adjusting either, until eventually we bring them all into coherence. It would indeed be likely to improve them. But how much?

This brings us on to the well-trodden ground of coherence theories. The *coherence theory of justification* holds that ultimately a belief is justified by, and only by, its being a member of a coherent set of beliefs.

The plausibility of that proposal rests crucially on how demanding "coherence" is taken to be. Nowadays "coherence" is thought of along these lines. "Coherence" cannot demand only consistency; consistency constitutes a quite weak test for a set of beliefs. A conjunction of merely consistent beliefs is no more credible than its least credible member; if a belief of 20 percent credibility is conjoined to a belief of 70 percent credibility, the credibility of the conjunction is 20 percent. The idea of "coherence" has to capture how beliefs can support one another, how in aggregate they can pull up the reliability level of the set of beliefs as a whole.

Coherence cannot be a matter simply of consistency on a *wide* front, either. Adding the requirement that the set of beliefs be comprehensive makes consistency more testing, but the test is fairly weak still. It does not yet capture the bootstrap effect just mentioned. Nor does it meet the regress objection. The regress objection is that the credibility of one belief cannot depend on the credibility of another without end; credibility must come from somewhere. And merely to say that it comes from a set of beliefs as a whole does not meet the objection. If one belief gets its credibility from a

[12] See, e.g., I. Lakatos, "Proofs and Refutations," *British Journal for the Philosophy of Science* 14 (1963–4): 1–25, 120–139, 221–245, 296–342. See also Michael Dummett's argument against the law of excluded middle and in favour of an intuitionist mathematics; for a recent statement of the issue, see his *The Logical Basis of Metaphysics* (London: Duckworth, 1991), pp. 9–11. See also discussion of these matters in Newton-Smith, *The Rationality of Science.*

second, and the second from a third, and so on until eventually, circling back to the starting point, some belief far into the chain of credibility transfers gets its credibility from the first, then we have no explanation of how there is any credibility to be transferred in the first place. The way to meet the regress objection is to reject the foundationalist presupposition at its heart, despite the coherentist language it is sometimes expressed in. Justification, one should insist, is not linear – neither a straight line from a starting point, as foundationalists say, nor a circular line, as coherentists might wrongly be thought to be saying. Coherentists can maintain that some sets of beliefs are not just consistent conjunctions but mutually supporting systems of belief and that it is their systematic connections that make them capable of raising the credibility level of the whole set.

So we should take "coherence" to mean an organization of beliefs in a network of inferential relations. The most justificatorily powerful of such relations are explanatory ones, and the most powerful example of organization that we have found so far is the sort of "systematic unification," as Carl Hempel put it, provided by a natural science.[13] Some contemporary supporters of the coherence theory take such "systematic unification" to be little different from the concept of "coherence."[14] On that interpretation, it is clear how coherence can constitute a test of considerable power. To make beliefs coherent is, in a way familiar from the natural sciences, to verify or falsify them. It is also to test the adequacy of our conceptual framework; in developing an explanatory system, language has often to change.[15]

Still, even on this rich interpretation of "coherence," the coherence theory has problems. For one thing, there are many different coherent sets of beliefs that are incompatible between themselves. How can a coherence theory single out one uniquely justified set?[16] For another, a belief is justified, according to the coherence theory, by its relations to other beliefs, not by its relation to the world. Can a coherence theory find a place for perceptual input from a nonconceptual world? It must; the aim of our empirical beliefs is to describe that world.

[13] Carl Hempel, *Philosophy of Natural Science* (Englewood Cliffs, NJ: Prentice Hall, 1967), p. 83.

[14] E.g., Laurence Bonjour, *The Structure of Empirical Knowledge* (Cambridge, MA: Harvard University Press, 1985), p. 99.

[15] This, I say, is the now dominant interpretation of "coherence." But hardly just now: see F. H. Bradley, *Essays on Truth and Reality* (Oxford: Oxford University Press, 1914), p. 210; B. Blanshard, *The Nature of Thought* (London: Allen and Unwin, 1939), vol. 2, pp. 275–276.

[16] This well-known objection goes back at least to Bertrand Russell, "On the Nature of Truth," *Proceedings of the Aristotelian Society* 7 (1906–7): 28–49.

These two objections are closely linked. There are many, probably infinitely many, possible worlds different from the actual world yet describable equally coherently.[17] We distinguish between the actual and only possible worlds by perception, by input from a reality independent of our thoughts. Our beliefs about the world must be empirically grounded, and this grounding will lead us to a unique, most-justified set.[18]

We believe, on the one hand, that the world around us impinges upon us, causes us to have the perceptions we have. We receive, we register, this independent world. These perceptual beliefs, therefore, have a high credibility that arises from the closeness of their causal connection to this independent world. Yet, on the other hand, we also accept that no perceptual belief merely registers a thought-independent reality. We may to some degree be passive in perception, but we are also partly active: We interpret; we categorize. Our human point of view – our particular sensitivities and interests – are always at work. We have no access to a world behind our experience, entirely innocent of our interpretation, our language. And the justification of a perceptual belief cannot be independent of other beliefs – independent, in particular, of the general principle that the beliefs of direct awareness that arise in certain privileged circumstances are highly credible. What is more, any one perceptual belief is defeasible; whether we stand by it depends on what turns up later.

So, a good coherence theory – indeed any good theory of justification at all – must accommodate both awareness of this independent world and the dependence of perceptual beliefs on other beliefs. What is not easy to decide is whether the theory that manages this can still be a coherence theory,[19] whether it must become a hybrid coherence-foundationalist theory,[20] or indeed whether the question is really now one for linguistic legislation. At the least, the necessary accommodation of empirical input requires the abandonment of a certain simple form of coherentism – the form that

[17] See Bonjour, *The Structure of Empirical Knowledge*, p. 107.

[18] In light of the underdetermination of theory by observation, whether our accommodating perceptual input will, on its own, ensure that we arrive at a *unique* set, may be doubted; but whatever one thinks about the matter of underdetermination it is not a complication that bothers just coherence theories. See J. Dancy, *Introduction to Contemporary Epistemology* (Oxford: Blackwell, 1985), pp. 114–116.

[19] As the following believe: Bonjour, *The Structure of Empirical Knowledge*, pp. 112–119; K. Lehrer, *Theory of Knowledge* (London: Routledge, 1990), pp. 145–146 (though, Lehrer adds, with "elements of foundationalism"); and in the domain of ethics, D. Brink, *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press, 1989), pp. 135–139.

[20] As Susan Haack, thinks; see her *Evidence and Inquiry* (Oxford: Blackwell, 1993), p. 19.

holds, as coherentism in many of its formulations implies, that no one belief starts with greater credibility than any other, that there are no favoured beliefs with an initial credibility independent of coherence with the rest. That accommodation brings with it the abandonment of this pure democracy of beliefs. It may take a large theoretical system to sanction the special authority of certain perceptual beliefs, but what it sanctions is precisely that those beliefs have a special authority because of how they relate to the world and not simply because of how they relate to other beliefs.

It is hard, though, to gauge the consequences of that admission. It moves us to a theory that in some form or other allows nonbelief input (so may not be coherentist) but does not treat perceptual beliefs as the starting points of a one-way route of transfers of credibility (so does not seem to be foundationalist). What is impossible to resist is not coherentism but holism. Holism is the thesis that justification comes only from a whole set of beliefs. Coherentism is just one version of that view: It gives a certain specification of what it is about the set that produces justification, namely, that the beliefs form a system of inferential relations. There is more logical space within holism than coherentism occupies. Both foundationalism and coherentism offer pictures of the overarching structure of credibility transfers. Simple foundationalism has starting points – basic beliefs – of high credibility with one-directional flows of credibility from them. Simple coherentism has beliefs conferring credibility on other beliefs solely through their own relations. The truth is bound to be more complicated than either of these simple pictures capture. What exactly it is I am going to leave aside. My interest is in our ethical beliefs, and the pressing questions about them are quite different.

4.5 A Coherence Theory for Ethics

The defence of the coherence theory that I have just sketched depends on its being a theory about the justification of *empirical* beliefs. It appeals to perceptual input and explanatory system. When we turn to ethical beliefs, we encounter analogues to the problems with coherence theory in science; the question is whether there are also analogues to the solutions. Where are the analogues to perceptual input and explanatory system? Or, if there are no analogues, where are the substitutes that play the same justificatory role?

Of course, we could resign ourselves to only weak analogies and so to a more modest interpretation of "coherence" for ethics. But the less demanding a requirement coherence represents, the less improvement in our beliefs reaching coherence will represent, and, at the extreme, there must sometime come a point at which the improvement is so slight that we should have to stop thinking of ourselves as involved in anything worth calling "justification."

Is there any analogue in ethics to perceptual beliefs? Are there beliefs of high reliability, beliefs of a credibility to some extent independent of their relation to other beliefs? Might the core values that Wittgenstein and Davidson speak of, the values that are part of the framework for intelligibility, be all the highly reliable beliefs that ethics needs? I doubt it. They may be all that ethics has, but they will not get us far. They, I take it, will be confined to a few basic prudential values, for instance, that we want to avoid pain and anxiety, that we have aspirations and attach importance to their being fulfilled, and perhaps also confined to a few basic moral beliefs, for instance, that cruelty is wrong and that we must show respect of some sort for others. But we shall have nothing comparable to the rich set of observations that operate in justifying scientific beliefs. Those few unshakeable prudential and moral beliefs will do no more than rule out highly implausible moral theories. The notion of respect, it is true, is closely connected to more specific concepts, such as some form of loyalty and honesty, but even their addition does not provide much of a test. It could not effectively test the moral views that we now think of as seriously in contention. Those views share most of the same specific ethical concepts; they differ over where in deliberation these concepts figure.[21]

There are various ways of enlarging the set of core beliefs necessary for intelligibility. They are by no means confined to ethical beliefs. Our core values are part of our being able to see others as persons; they are normative constraints on central notions in the philosophy of mind.[22] To see an event as an action, one must be able to see it as intentional, which requires seeing it as aimed at some good or other. But these mental notions are involved in the claims about intelligibility that we have already made.

One might also try adding to the core beliefs various specific ethical notions (what are now often called "thick" concepts), such as "loyal," "honest,"

[21] The same conclusion holds if, instead of considering the conditions for the intelligibility of language generally, we consider those, more specifically, of *ethical* language. For me to see your concerns as ethical, I have to see them as giving human and animal interests a fairly central place, showing respect of some sort for others, and so on. But these constraints rule out only pretty odd views; the views that we seriously wonder about and try to choose between are still left in contention. I say more about the concept of the ethical in *Value Judgement*, ch. 8, sect. 1.

[22] For a good discussion of these issues, see Susan Hurley, *Natural Reasons* (New York: Oxford University Press, 1989).

"just," "chaste," "patriotic," and so on. If they are not quite part of what Wittgenstein calls our form of life, they are anyway much more deeply embedded in a culture, indeed in a particular period of a culture, than thin terms such as "good" and "ought". But thickness is not reliability. Our thick concepts largely define our current commonsense ethical outlook. They are many of the intuitions that I spoke of at the start. They are not the highly reliable beliefs that we hope might be available.

Might we find such beliefs, then, in a different way? We could, as Rawls suggests,[23] put intuitions through an initial sifting, looking for coherence not with all our moral beliefs, no matter how confused or ephemeral, but with only "considered" ones. My considered intuitions, I could say, following John Rawls's lead, are those of which I am confident for a fair amount of time, and which I formed in the absence of conditions likely to corrupt judgement; for example, I was calm, adequately informed, and my self-interest was not aroused.[24] Considered judgements would seem, therefore, to have more weight than unsifted intuitions, and so coherence with them would be more likely to bring improvement.

With perceptual beliefs, we have reason to think that we are to some extent passive recipients of an independent reality. Part of what enters the holistic balance in science is an account of what goes on in observation because that is part of what is to be explained. We test our beliefs about how we are causally connected to what we observe, how we make perceptual errors and correct them. In the natural sciences, part of what is being justified holistically is our belief in there being certain sorts of reliable beliefs. And it receives a lot of justification at quite early stages in our thought about the world, without our needing much help from philosophy – from, say, epistemology and metaphysics. We know that if our eyes and the light are

[23] For Rawls's definition of "considered judgements," see his *A Theory of Justice*, pp. 47–48; see also p. 20.

[24] Rawls's criteria for considered judgements would often leave one with pretty dubious beliefs. Why confine ourselves only to intuitions of which we are relatively *confident*? Confidence in ethics has different psychological explanations, many of them not reassuring. The confident ethical beliefs of a thoroughly comfortable member of a privileged class might be his worst; his best might be his occasional unconfident glimmerings of a different way of life. And why confine ourselves to *calm* judgements? Many people's best moral thinking is reserved for their death bed or their doctor's waiting room (a point made by Norman Daniels, "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *Journal of Philosophy* 76 (1979): 256–282, cf. p. 258). Anyway, to say that we should interest ourselves only in judgements formed in the absence of conditions likely to corrupt judgement begs the important questions. If we knew which conditions did that, and also knew that we were avoiding them, we should indeed be able to isolate a class of especially reliable judgements. But we do not know it.

good, we are close up, and we take a good look, our resultant belief about what we see is especially secure. With considered judgements in ethics we have nothing like as strong an assurance. This is the central point in the case against piecemeal appeal to intuitions. The causal story of our ethical beliefs is generally much more tangled, much less easily established, than the story of our perceptual beliefs. Perhaps *some* value judgements are perception-like and the causal story behind them is relatively simple. There is a lot to be said for that view, I think, in the case of prudential value judgements, judgements about what meets or fails to meet basic human interests.[25] However, complex moral norms, say, about stealing or killing, have highly complex causes.[26] Some of them arise from solutions to cooperation problems that evolve in a society well below the level of conscious decision. Social convergence, convention, myths, taboos, religion, metaphysics, light or dark pictures of human nature, economic conditions, and so on, play an important role, and this highly complex causal background makes ethical beliefs more susceptible to defects. This is not to deny that we can supply a causal account of our normative ethical beliefs or an error theory for them. But neither will be forthcoming until we are able to answer certain metaethical questions in a certain way. Nor is it to deny that most of us must quite naturally get a fair amount right in our ethical beliefs and that those sound beliefs constitute a basis for criticizing our ethical beliefs generally. But it is to deny that those beliefs constitute anything like as large or as readily identifiable a group as the highly reliable beliefs in the natural sciences. This difference is only one of degree, but a difference in degree can turn into a big difference in how rich an interpretation the notion of "coherence" will bear and in how powerful the coherence test will be. In any case, we cannot know what to expect until we know more about the nature of prudential values and of moral norms, and that means broaching some of the central issues in metaethics. Perhaps some value judgements are perception-like, but we need good (metaethical) reasons to accept that conclusion. It is hard to confine the question of justification within the boundaries of normative ethics.

Nor, I should say, is there a strong analogue in ethics to science's goal of explanatory system. We have a much better idea of how to measure the success and the correctness of some set of beliefs, if we know our purpose in forming them. We know about standards of success for a natural science: Does it describe how its chosen part of the world works? The natural world,

[25] See Griffin, *Value Judgement*, chs. 2–4.

[26] See Griffin, *Value Judgement*, chs. 5–7.

as we grasp it, is a network of causes, and this conception of it tends to make our description of it systematic. What, then, is our aim in holding ethical beliefs? The general answer is plain: to decide how to live. But there is nothing in that aim that need take us far down the road to either explanation or system. To decide how to live, we need prudential and moral standards. But they might arise in a piecemeal, unsystematic way. Some of them, as I have just suggested, probably embody solutions to cooperation problems, with different solutions to the same problem evolving in different societies, and solutions to different problems evolving in a single society largely in isolation from one another. The aim of deciding how to live is achieved, in this part of life, once we have tolerably satisfactory solutions to these cooperation problems; not much in the way of system and explanation is required. Our ethical and other beliefs, it is true, do support one another up to a point; it would be an exceptional ethical belief that did not stand in an inferential or evidential relation to other beliefs. There is doubtless some degree of organization to our ethical thought. But that is far from these beliefs' forming, as our scientific beliefs do, a systematic network of credibility transfers. It is hard to see why they should. It is not enough that our ethical principles should form a system in the familiar sense that they are organizable into a structure – of subordination, for instance (as in utilitarianism, all secondary principles are subordinate to the single principle of utility), or of equipollence (as in W. D. Ross's intuitionism, the seven prima facie duties are same-level principles). The sort of system we are looking for now is one not simply of organization into a structure but of organization into a network of credibility transfers that can raise the level of the whole set of beliefs. The first sort of system can lead to the second, but it need not.[27] I am not saying that we can tell, at this early stage, that no system in the second, stronger sense will emerge between ethical beliefs; what I am suggesting is that, at this stage, we have no reason either to assume that it will. Nor is it enough that our ethical beliefs should display system in another familiar sense – that a general principle should throw light on particular cases, often quite difficult ones, and vice versa. This certainly happens – in fact, a bit too readily for it to show much. Utilitarians rightly think that the principle of utility illuminates very many particular cases; deontologists rightly think that the principle of respect for persons, or the doctrine of double effect, does too; and so on. I suspect that it is most often this sort of mutual illumination that philosophers have at the back of their minds when

[27] In any case, I have my doubts about the first sort of system; see Griffin, *Value Judgement*, ch. 7.

they hold that ethics is capable of system. But achieving this sort of system does nothing to discriminate between major moral views. A view would not have become major, I suppose, unless it had a good deal of this power. The explanatory circle, though, is too small. Not much in the way of credibility transfers will flow along these short lines.

4.6 What We Need

What we should do, then, is to start more modestly, not in order to embark on a Cartesian reconstruction of the whole body of our ethical beliefs, which is neither a modest nor a feasible project, but to start more or less where these reasonable, non-Cartesian doubts actually leave us. We should not even assume that "justification," that is, some integrated structure of credibility transfers is appropriate to ethics. There are certainly transfers of credibility between beliefs in ethics here and there, and the local networks of transfers might sometimes become quite extensive. But we should not assume that it is a philosopher's job to find a global network for ethics, modeled on the competing theories of justification for empirical beliefs.

To move things forward, there are two things that we can do. First, we can look for beliefs of high reliability. If, as I suspect, ethical beliefs of high reliability are not confined to those core beliefs necessary for intelligibility, then we must find out what these further beliefs are. Even without beliefs of high reliability, achieving coherence on a wide front provides a test the passing of which confers at least some credit on a set of beliefs. But because of the freedom we should have in arriving at coherence, the credit might be quite modest. A lot would turn on how respectable our initial set of ethical, and other, beliefs happened to be. I include "other" beliefs because clearly it is not just ethical beliefs that can count for or against ethical beliefs; one might include in the final coherence, as Rawls does in his notion of a *wide* equilibrium, any belief relevant to any ethical view.[28] If this large set were fairly respectable, achieving coherence could bring considerable improvement. If it were not, then it might well not. I should not know

[28] Rawls explains wide reflective equilibrium thus:

> There, however, are several interpretations of reflective equilibrium. For the notion varies depending upon whether one is to be presented with only those descriptions which more or less match one's existing judgments except for minor discrepancies, or whether one is to be presented with all possible descriptions to which one might plausibly conform one's judgments together with all relevant philosophical arguments for them [i.e., wide reflective equilibrium].

See his *A Theory of Justice*, p. 49.

what achieving coherence in my own case brought unless I knew something about the respectability of my own initial beliefs. It might, of course, boost my confidence in where I ended up if I found that others were ending up there too. But this is another way in which ethical belief differs importantly from empirical belief. Convergence in belief boosts confidence only if the best explanation of the convergence is of the right sort. With a perceptual belief, if mine differs from everyone else's, the best explanation is that my sense is malfunctioning; if it converges, I can be reassured. It is much less clear what the best explanation of convergence in ethical belief is (think, for instance, of the various convergences we have today), and so how reassuring it is. We should have to know how reliable and how decisive some of the beliefs entering the convergence are. Of course, convergence helps, but for it to do so we also need a fair amount of confidence that it is happening for the right reasons. Can we have that confidence without settling major questions in metaethics? Can we have it without finding some beliefs of high reliability? It may be that there is no stronger test available to us than coherence without beliefs of high reliability, but we should not resign ourselves to the modesty of our critical powers until we are pretty sure we must.[29]

The second thing that we can do is to get a better idea of what ethics can reasonably aspire to be. Can it, for instance, aspire to system? We can do this in part by asking what agents have to be like to be able to live the sort of life that various systematic ethics demand of them.

Both of these projects, especially the first, require broaching some major issues in metaethics. To my mind, normative ethics is not, despite what Rawls says, largely independent of metaethics; neither can be pursued fruitfully for long without attending to the other.

[29] We should not need to look for highly reliable *ethical* beliefs if we could assess competing moral views just by appealing to nonmoral matters of fact. All we should then have to do is to find the relevant highly reliable factual beliefs. And facts about human motivation and about how societies work go a long way toward weeding out unrealistic moral views. This possibility raises a raft of familiar questions about the relation of fact and value, particularly about reductive naturalism – the view that ethical beliefs can be reduced to factual ones, on roughly Hume's understanding of the "factual." I find reductive naturalism implausible. But the facts that do indeed go a long way toward testing moral views, e.g., facts about human motivation and about how societies work, are far from purely factual. For instance, some moral views rest on dubious conceptions of the human will. We cannot determine the limits of the will independently of knowing what are plausible human goals and how inspiring they are. The capacity of the will is partly a function of its goals. So any "fact" likely to get far in testing competing moral views will be partly constituted by beliefs about values; we shall not know whether it is highly reliable without knowing whether its constituent ethical beliefs are too.

There can be no test of much strength for normative ethics without answers to certain key metaethical questions. Rawls's special contribution to the coherence theory has been to maintain that the test for ethical beliefs is largely independent of metaethics.

His reason is this. Metaethics is concerned with such questions as whether and in what sense moral judgements are true, whether they are objective, whether values form an order independent of human belief and attitude, and when they can be known.[30] Normative ethics, by contrast, is the systematic, comparative study of competing general moral views – utilitarianism, Kantianism, virtue theory, and so on. The programme of normative ethics is to develop each view, probably much further than they have yet been developed, then to compare their features, and also importantly, on that basis, to decide on their relative adequacy.[31] For my own part, I decide their adequacy by bringing my own beliefs into wide coherence. Once the rest of you have also done this with your beliefs, we may find ourselves converging on some of the same beliefs. If enough of us converge, then we may be willing to regard the beliefs converged upon as objective.[32] We might then also be in a position to settle issues about the truth of moral judgements, the independent reality of values, and other metaethical difficulties as well.[33] In this way, Rawls argues not just for the independence of normative ethics from metaethics,[34] but also for its priority.[35] At this stage in the history of philosophy, he says, we are not in a position to make much headway with metaethics, but we have just seen ways in which, with advance in normative ethics, we might eventually make advance in metaethics too.

But can we describe a test powerful enough to rank competing normative views, while ignoring metaethical questions about objectivity, truth, or knowledge? The test at work in normative ethics must yield a ranking in a strong sense. It must lead us not merely to a preference between the competing views but to a decision about which has more reason on its side. It must guard against the quite ordinary ways in which our moral beliefs go wrong. It must meet doubts about our beliefs that arise from our own past mistakes – that is, not extreme philosophical doubts about whether we can know

[30] See Rawls, *A Theory of Justice*, pp. 51 ff.; "The Independence of Moral Theory," pp. 5–7; "Kantian Constructivism in Moral Theory," p. 554.

[31] See Rawls, "The Independence of Moral Theory," p. 8.

[32] See Rawls, "The Independence of Moral Theory," p. 9; "Kantian Constructivism in Moral Theory," pp. 554, 570.

[33] See Rawls, "Kantian Constructivism in Moral Theory," pp. 564–565.

[34] See Rawls, "The Independence of Moral Theory," pp. 9, 21.

[35] See Rawls, *A Theory of Justice*, p. 53; "The Independence of Moral Theory," pp. 6, 21.

anything, or at least anything about values, which is a problem that we consign to metaethics, but entirely realistic doubts. Rawls agrees.[36] In describing the ranking, he regularly uses terms that carry considerable epistemic weight. We compare moral views, he says, on the basis of, among other things, how well they accommodate facts about the human psyche and society; that decides what Rawls calls their "feasibility". Then, given their feasibility, we look at their content in wide coherence; that decides their "reasonableness".[37] And, for Rawls, decisions about reasonableness have to come largely from each individual's reaching wide coherence; the further step of convergence between different individuals' beliefs adds little. Lack of convergence can, it is true, serve as a trip wire. My lack of convergence with the rest of you on what I claim to see trips up my claim to see, but whatever special reliability reports of perception have rests primarily on what individual perception is, not on convergence. Similarly, convergence between you and me in ethics matters to the justification of belief only if it is what has been called "principled" convergence, that is, convergence arising from your or my having separately applied standards of reasonableness to the formation of our own beliefs. Rawls agrees with this too.[38]

His agreement just brings us back to old questions. It is not enough to say that putting our beliefs in wide coherence will distinguish the more from the less reasonable. It will do that only if we can identify beliefs of high reliability. As we cannot do that without broaching some key metaethical questions, the independence of normative ethics is seriously compromised.

Given the present state of philosophy, can we make progress in metaethics? Well, we now know so little about the nature and structure of our substantive ethical beliefs that we do not know whether the best moral view will, in the end, recommend itself to us because it meets epistemological standards or because it meets practical ones, such as its meshing effectively with the human will or its providing a much needed social consensus for us here and now. We may find that moral standards are what we agree between us to adopt as such, not what we discover independently to be such. Therefore, we cannot get far with metaethical questions about truth, objectivity, and realism until we have got clearer about the status that moral standards have in what turns out to be the best normative view. This argument of Rawls

[36] See Rawls, *A Theory of Justice*, pp. 50, 53, 121, 452; "The Independence of Moral Theory," pp. 8–9; "Kantian Constructivism in Moral Theory," pp. 534, 568–569.

[37] Rawls, "The Independence of Moral Theory," p. 15; "Kantian Constructivism in Moral Theory," p. 534.

[38] Rawls, "The Independence of Moral Theory," p. 9.

seems to me to have some force. But there is a second argument. For the reasons I have just given, we cannot get far with finding the best normative view until we are clearer about what beliefs are highly reliable, and for that we need answers from metaethics. The combined effect of these two arguments is that sometimes the priority runs one way and sometimes the other. That is why I end by saying that normative ethics and metaethics have to advance together. The first is not independent of the second or, as Rawls allows, the independence he has in mind is not especially strict.[39] There is nothing like the high degree of independence that he suggests.

In any case, it seems to me that, in ethics, coherence is a weak test. So we had better think about what might be stronger.

[39] Rawls, "The Independence of Moral Theory," pp. 5, 6, 21.

PART TWO

HARSANYI'S IMPARTIAL OBSERVER AND SOCIAL AGGREGATION THEOREMS

5

Harsanyi's Impartial Observer Is *Not* a Utilitarian

John E. Roemer[1]

5.1 Introduction

Harsanyi (1953) proposed a veil-of-ignorance argument for concluding that a rational soul, behind the veil of ignorance, would behave like a utilitarian – more precisely, that it would maximize a weighted sum of von Neumann–Morgenstern utilities of individuals. The argument is justly famous, as the first attempt to formalize the idea of the veil of ignorance, using the then recently developed tool of von Neumann–Morgenstern utility, that is, of decision theory under uncertainty. Indeed, Harsanyi used the terminology of the impartial observer (IO), rather than the veil of ignorance, but I shall assume these two metaphors are attempts at capturing the same, ethically correct stance. Weymark (1991) calls the argument Harsanyi's impartial observer theorem. I shall argue that Harsanyi's conclusion is incorrect: It does not follow from his argument that the IO is a utilitarian. The essential point is that utilitarianism requires, for its coherence, a conception of interpersonal comparability of welfare, and no such conception adheres to the concept of von Neumann–Morgenstern utility that Harsanyi invokes.

Let X be the set of social alternatives, or states of the world, and let H be the set of types. Think of X, for instance, as a set of possible income distributions among persons. Define the set of extended prospects as $Y = H \times X$, whose generic member is (h, x). Behind Harsanyi's veil of ignorance, the IO faces the set of extended prospects, where (h, x) is interpreted as meaning "I shall become a type h person in state of the world x." Thus, we must think of x as including a description of how each type, h, fares, so that a prospect (h, x) will be a complete description of how well h's life goes in state x.

[1] I thank Klaus Nehring for useful discussions.

For any set of possible outcomes Z, let $L(Z)$ denote the set of lotteries on Z. Harsanyi assumes that each person has preferences on $L(X)$ that obey the von Neumann–Morgenstern (vNM) axioms and hence can be represented by a vNM utility function defined on X: Call the vNM utility function of a type h person u^h. (Of course, u^h is defined only up to positive affine transformations.) Harsanyi also assumes that (1) the IO has vNM preferences on $L(Y)$ and (2) that the Principle of Acceptance holds, a postulate to be stated in the next section.

Essentially similar arguments to what follows have been presented by Weymark (1991) and Roemer (1996, chapter 4). The key observation was first made by Sen (1976). But the arguments in those places are perhaps too abstract. The purpose of this chapter is to drive the point home with a simple, and I hope compelling, example. In addition, the present statement of Harsanyi's error is slightly different from what Weymark and I wrote previously.

5.2 Harsanyi's Argument

It is worthwhile to reproduce a proof of Harsanyi's impartial observer theorem. This one follows Roemer (1996, chapter 4).

First, some notation. Let X consist of states $x^1, \dots, x^N$, and H of types $1, 2, \dots, T$. Represent a lottery on X as a probability distribution $\pi = (\pi^1, \dots, \pi^N)$ on X, a lottery on H as a probability distribution $\rho = (\rho^1, \dots, \rho^T)$ on H, and a lottery on $Y = H \times X$ as a probability distribution σ on Y, where σ is a $T \times N$ matrix whose hj^{th} element is the probability of the extended prospect (h, x^j).

From the IO's viewpoint, it faces "extended prospects" of the form (h, x), where h is a type and $x \in X$: that is, it could be embodied in a type h person in state of the world x, for any h and x. Harsanyi posits that the IO has preferences over these lotteries that obey the von Neumann–Morgenstern axioms. They can be represented by a von Neumann–Morgenstern utility function ϕ on Y. Denote by Φ the utility function on $L(Y)$ induced by ϕ on Y via the expected utility property. Then by the expected utility property we can write

$$\Phi(\sigma) = \sum \sigma^{hj} \phi(h, x^j) = \sum_h \sum_j \sigma^{hj} \phi(h, x^j). \tag{5.1}$$

Now let us suppose that the lottery σ takes the form $\sigma^{hj} = \rho^h \pi^j$, where ρ is a probability distribution on H and π is a probability distribution on X. This is the kind of lottery that the IO will face, where ρ is the "birth lottery" and π is the lottery persons in the actual world face. We can then further write

$$\Phi(\sigma) = \sum_h \rho^h \sum_j \pi^j \phi(h, x^j). \tag{5.2}$$

Harsanyi argues that on $\{h\} \times L(X)$, the IO should have the same preferences as type h: this is the *Principle of Acceptance.* The functional $\sum_j \pi^j \phi(h, x^j)$, viewed as a function on $L(X)$, has the expected utility property, and this function represents the soul's preferences on $\{h\} \times L(X)$. It follows that

$$\text{for each } h, \quad \exists a^h > 0,\, b^h \text{ s.t. } \phi((h, x)) = a^h u^h(x) + b^h, \tag{5.3}$$

since the vNM utility function on $L(X)$ of h is unique up to positive affine transformations. (Since ϕ is fixed and the U^h are fixed, the choice of (a^h, b^h) is completely determined.) Substituting from Eq. (5.3) into Eq. (5.2) yields:

$$\Phi(\sigma) = \sum_{h,j} \rho^h \pi^j \tilde{u}^h(x^j) + K, \tag{5.4}$$

where $\tilde{u}^h \equiv a^h u^h$ is another vNM utility function for h, and K is a constant. The constant K is immaterial; hence Eq. (5.4) *apparently* says that the IO is a utilitarian: Its preferences are represented by a utility function, which is interpreted as the appropriate probability weighted sum of utilities of types in the actual world.

5.3 Why the IO Is Not a Utilitarian

I shall argue by use of a simple example. Consider an individual, Alicia, who has preferences over cash lotteries. Alicia has a specific, well-defined conception of welfare: explicitly, she favors one lottery over another if and only if the first gives her greater welfare. To simplify the discussion, let us consider only lotteries of the form $((\pi, x_1); (1-\pi, x_2))$ in $L(X)$, which denotes the lottery in which Alicia will receive x_1 in cash with probability π and x_2 with probability $1-\pi$. The fact of the matter is that Alicia will enjoy welfare in the amount $\sqrt{x}$ if she receives x in cash. If Alicia faces the lottery $((\pi, x_1); (1-\pi, x_2))$, her welfare will be $[\pi x_1 + (1-\pi)x_2]^{1/2}$. (If Alicia faces a compound lottery, her welfare level is the square root of the expected value of the lottery.) In particular, the reader can verify that:

$$\forall \pi, x_1, x_2 \quad [\pi x_1 + (1-\pi)x_2]^{1/2} \geq \pi\sqrt{x_1} + (1-\pi)\sqrt{x_2}, \tag{5.5}$$

where strict inequality holds except when $\pi = 0, 1$. Therefore, Alicia's welfare, when facing a lottery, is (generally) greater than the average of the welfare levels she would enjoy from receiving the two cash prizes as sure things, weighted by the appropriate probabilities. Is this crazy? No. The explanation, for instance, could be that Alicia is usually in a state of dysphoria, but when she faces a lottery, she perks up, smiles, and gets excited. These physiological responses increase her welfare.

Now let us postulate that Alicia's preferences over lotteries obey the vNM axioms. Then we can deduce her vNM utility function from Eq. (5.5), for we know her ordinal preferences over lotteries are represented by the utility function

$$\Phi((\pi, x_1);(1-\pi, x_2)) = [\pi x_1 + (1-\pi)x_2]^{1/2}. \qquad (5.6)$$

Any strictly monotone transformation of Φ also represents Alicia's preferences over lotteries. Let us apply the transformation $F(z) = z^2$. Thus, we have that

$$F(\Phi((\pi, x_1);(1-\pi, x_2))) = \pi x_1 + (1-\pi)x_2 \qquad (5.7)$$

is also a utility function over lotteries representing Alicia's preferences. But *this* utility function has the expected utility property, if we define u on amounts of cash by $u(x) = x$, for then Eq. (5.7) takes the form:

$$F(\Phi((\pi, x_1);(1-\pi, x_2))) = \pi u(x_1) + (1-\pi)u(x_2). \qquad (5.8)$$

Thus, Alicia's sole concern with welfare leads her to have risk neutral preferences over cash lotteries. The actual welfare she receives when facing a lottery is given by Eq. (5.6).

Next we introduce Bogdan, who is just like Alicia. His sole concern is with his welfare, and his welfare, when facing lotteries, is given by Eq. (5.6). Bogdan, too, is postulated to have vNM preferences over lotteries. In like manner, we deduce that we can take Bogdan's vNM utility function to be $u(x) = x$.

Now consider the IO who must decide on allocations of cash between Alicia and Bogdan. The observer assigns probability $1/2$ of being in Alicia's (Bodgan's) shoes. If the IO endorses the Principle of Acceptance and has vNM preferences over lotteries, then Harsanyi's theorem tells us it must maximize a utility function of the form

$$ax_A + (1-a)x_B, \qquad (5.9)$$

where a is some number in $[0, 1]$, and where (x_A, x_B) will be the allocation to Alicia and Bogdan of cash, chosen from some feasible set, X, of such allocations. Equation (5.9) follows since we know that $u(x) = x$ is an acceptable vNM utility function for both Alicia and Bogdan on the set of sure prospects.

Let us now suppose that both Bogdan and Alicia agree that the meaningful way to compare the values of their lives is to compare their welfares: Each agrees that welfare is the only important value, and that a unit of welfare for Bogdan has just the same value as a unit of welfare for Alicia. That is, meaningful interpersonal comparisons must be done in units of welfare. Then, at any allocation of cash (x_A, x_B), the average welfare in their world will be

$$\frac{1}{2}\sqrt{x_A} + \frac{1}{2}\sqrt{x_B}. \tag{5.10}$$

Thus a utilitarian, who maximizes average welfare, must choose (x_A, x_B) to maximize Eq. (5.10). Harsanyi's IO, as I've said, maximizes Eq. (5.9), for some fixed number a. But once a is fixed, it is easy to supply feasible sets of cash allocations for which Eq. (5.9) and (5.10) lead to different solutions. Thus, Harsanyi's IO is not a utilitarian.

[One might be tempted to respond that the IO gets to choose a after the set X is revealed. If X is convex, a can always be chosen so that maximization of Eq. (5.9) and (5.10) yield the same allocation. But this response is wrong, for if we allowed the IO to choose a after X is revealed, then there is no content in saying the IO has the objective function given in Eq. (5.9), for we could simply fit Eq. (5.9), by appropriate choice of a, to yield any point on the northeast boundary of the revealed X.]

What is going on here? The combination of axioms (1) and (2) [see Section 5.1] and the view that the IO is utilitarian is inconsistent, for that combination *forces* us to interpret the IO as making interpersonal comparisons between Alicia and Bogdan by transforming their utility, as measured by their vNM utility functions, by a *linear* transformation of units. For *if* we interpret Eq. (5.9) as a utilitarian formula, then we must say that $a/(1-a)$ units of Bogdan's utility are comparable to one unit of Alicia's utility. This *must* be the case if Eq. (5.9) is to be interpreted as an average of the unit-comparable welfare of Alicia and Bogdan. But *we* know that's too restrictive, for the correct way to transform Alicia's welfare at x_A into Bogdan's welfare at x_B is to multiply her welfare at x_A (which is $\sqrt{x_A}$) by $(x_B/x_A)^{1/2}$. In other words, the IO, *if* we insist on thinking of it as a utilitarian, must adopt a false conception of interpersonal comparability.

There are three straightforward alternatives to the Harsanyi view, taking as given that justice should be modeled using the veil of ignorance or impartial observer.[2] The first is to continue to endorse both the Principle of Acceptance and the view that rationality of the IO requires that its preferences obey the vNM axioms, and therefore to accept the conclusion that justice requires maximization of Eq. (5.9) for some a, but to drop the inconsistent claim that Eq. (5.9) consists in utilitarianism. The second alternative is to deny at least one of axioms (1) and (2) but to replace them with axioms which will enable one to infer that the IO is a true utilitarian [in the sense of maximizing Eq. (5.10)]. The third alternative is to deny at least one of the two Harsanyi axioms and to replace them (it) with axioms that do not imply true utilitarianism or Eq. (5.9). I do not pursue this research strategy here.

5.4 The Analogy With Individual Choice

We can see the error of calling the IO utilitarian by an analogy to individual choice. Consider Bogdan and Alicia, now in the actual world, who face the following problem. (These are the same *dramatis personae* as in Section 5.3, with their common conception of welfare.) A cash drop of M will fall on either Alicia or Bogdan, with probability 1/2 that each will receive it: thus, each of them faces the lottery $((1/2, M); (1/2, 0))$. They consider insuring each other. An insurance policy takes the form "he/she who receives the cash drop transfers a sum x to the other person." Alicia and Bogdan each compute the optimal insurance policy by solving:

$$\max_x \frac{1}{2}u(M-x) + \frac{1}{2}u(x). \tag{5.11}$$

Since $u(x) = x$, any value of of x in $[0, M]$ is optimal, naturally, because Alicia and Bogdan are risk neutral. Let us say, then, they choose $x = 0$.

Now one might be tempted to say that Alicia (or Bogdan) is a utilitarian with respect to her (his) own choices across states: that is, Eq. (5.11) appears to say that Alicia is maximizing her average welfare across states. But this view would be incorrect, for Alicia's average welfare across states, at the optimal insurance policy we have chosen, is $\frac{1}{2}\sqrt{M} + \frac{1}{2}0 = \frac{1}{2}\sqrt{M}$. But her average welfare across states is maximized when she chooses $x = M/2$, for that average then becomes $\frac{1}{2}\sqrt{(M/2)} + \frac{1}{2}\sqrt{(M/2)} = \sqrt{(M/2)}$; note that $\sqrt{(M/2)} > \frac{1}{2}\sqrt{M}$.

So it is wrong to say Alicia is behaving like a utilitarian, even though her decision problem (5.11) has the symbolic appearance of being "utilitarian" over states. The resolution is to note that $u(x)$ is not a measure of Alicia's welfare

[2] There are, doubtless, some non-straightforward ways as well.

in units that are interstate comparable. To get interstate comparability, we must measure Alicia's utility in units of welfare. At any solution of Eq. (5.11), Alicia will indeed be maximizing her welfare, for her welfare, when facing any lottery $((1/2, M - x); (1/2, x))$ is $[\frac{1}{2}\ (M - x) + \frac{1}{2}\ x]^{1/2} = \sqrt{(M/2)}$. Thus, the insurance policy Alicia and Bogdan have agreed on gives each of them the highest welfare they can hope for.

5.5 Conclusion

The combination of the Principle of Acceptance and the postulate that the IO has vNM preferences over lotteries implies that the IO maximizes a probability weighted sum of the vNM utilities of persons, where the probabilities are those of the "birth lottery." *If* we, the scientists, wish to call the IO's objective utilitarian, then we are forced to say that the IO makes interpersonal utility comparisons by adjusting the vNM utility of different individuals with simple scale multiples, that is, the IO transforms vNM utility into interpersonally comparable utility by simple linear transformations. But there is no reason that the true interpersonally comparable welfare of the individuals, assuming such exists, can be measured with such simple transformations of vNM units. A (true) utilitarian must use a utility scale that renders the units of utility interpersonally comparable across individuals. We cannot deduce, from knowing that two individuals have (ordinal) preferences over lotteries that obey the vNM axioms, what the measure of interpersonally comparable utility is. Neither does postulating axioms (1) and (2) of Section 5.1 solve the problem of making interpersonal welfare comparisons. The error lies in confusing a mathematical sum of vNM utilities with the substantive view of utilitarianism.

References

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61, 434–435.

Roemer, J. 1996. *Theories of Distributive Justice.* Harvard University Press, Cambridge, MA.

Sen, A. K. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7, 243–262.

Weymark, J. 1991. A reconsideration of the Harsanyi-Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. Roemer. Cambridge University Press, Cambridge, pp. 255–320.

6

Social Aggregation and the Expected Utility Hypothesis

Charles Blackorby, David Donaldson, and John A. Weymark

6.1 Introduction

Harsanyi (1955, 1977) interprets his Social Aggregation Theorem as providing support for weighted utilitarianism – the social ranking of alternatives by a weighted sum of utilities. In the fixed population setting considered by Harsanyi, classical and average utilitarianism coincide and correspond to having identical weights for each individual. The problem Harsanyi considers is one of social choice under uncertainty. Following von Neumann and Morgenstern (1947), he supposes that the set of social alternatives consists of all the lotteries that have a fixed finite set of certain alternatives as possible outcomes once the uncertainty is resolved. Harsanyi requires each individual and society to have preferences that satisfy the expected utility axioms and he represents these preferences by von Neumann–Morgenstern utility functions. His theorem demonstates that the social utility function must be an affine combination of the individual utility functions if society is indifferent between a pair of alternatives whenever all individuals are indifferent (the familiar Pareto Indifference condition). Thus, for any choice of the von Neumann–Morgenstern utility representations, alternatives are ranked socially according to a weighted sum of utilities. The Strong Pareto principle implies that all of the individual weights can be chosen to be positive.

This chapter is a revised version of the paper we presented at the conference on Justice, Political Liberalism, and Utilitarianism held in honour of John Harsanyi and John Rawls at the Université de Caen in June 1996. A version of Theorem 6.7 was first presented to the conference on Social Choice and Welfare in Valencia in July 1989. Discussions over the years with John Broome, Claude d'Aspremont, Philippe Mongin, and John Roemer on the subject of Harsanyi's theorem have been particularly useful in preparing this article. We are grateful to an anonymous referee for his or her comments. Our research has been generously supported by the Social Sciences and Humanities Research Council of Canada.

Sen (1976) has questioned Harsanyi's utilitarian interpretation of his theorem, noting that the weights used in Harsanyi's theorem to aggregate the individual utilities depend on which von Neumann–Morgenstern utility functions are chosen to represent the preferences, whereas with weighted utilitarianism, the weights should be independent of this choice. Further, the expected utility theorem only says that a preference ordering that satisfies the expected utility axioms *can* be represented by a von Neumann–Morgenstern utility function. It does not say that it *must* be so represented – any increasing transform of such a function is an equally good representation. In addition, the expected utility theorem does not imply that von Neumann–Morgenstern utility functions are the relevant representations for welfare analysis. Harsanyi's theorem makes essential use of von Neumann–Morgenstern representations.[1]

In the terminology of social choice theory, Harsanyi's problem is one of single-preference-profile social aggregation; there is one preference ordering for each individual and, correspondingly, one social preference ordering. This is a natural way to model the aggregation problem if the actual preferences are known at the time the aggregation rule is designed. In contrast, in Arrovian social choice theory [Arrow (1951)], the social aggregation procedure is designed before the individual preferences are known. This is a multiprofile aggregation problem in which the social ordering of the alternatives is conditional on the individual preferences. In either case, because information about the individuals' *preferences* is all that is available, no interpersonal comparisons of utility are possible. However, weighted utilitarianism requires interpersonal comparisons of utility gains and losses. This suggests that if Harsanyi's theorem is to have any relevance for utilitarianism, it must be reformulated to permit interpersonal utility comparisons. In such a reformulation, individual utility functions, not preference relations, are the data of the problem. Further, if these utility functions are not known a priori, a multiprofile approach is necessary.

In this chapter, the concept of a social evaluation functional introduced by Sen (1970) is used to model the social aggregation procedure. A social evaluation functional maps each admissible profile of individual utility functions into a social ordering of the alternatives.[2] We consider both single- and multiprofile social choice as well as alternative assumptions concerning the measurability and comparability of individual utilities. These assumptions are formalized by partitioning the set of admissible profiles of utility

[1] See Roemer (1996, 2008) and Weymark (1991) for extended discussions of Sen's critique.

[2] A social evaluation functional is often called a social welfare functional.

functions into information sets within which all profiles are informationally equivalent. Because the available information does not permit distinguishing between profiles in an information set, a social evaluation functional must assign the same social ordering to each profile in a single information set. In Arrovian social choice theory, utilities are ordinally measurable and interpersonally noncomparable and, as a consequence, each information set contains all profiles of utility functions that (person-by-person) represent a single preference profile. For Harsanyi's problem, the domain of the social evaluation functional consists of a single information set in the ordinal noncomparable partition of the set of possible utility profiles. Weighted utilitarianism requires a finer information partition.

Central to Harsanyi's approach to social ethics is his belief that individual and social preferences should satisfy the expected utility hypothesis. Harsanyi models uncertainty using lotteries, but there are other versions of expected utility theory that model uncertainty differently. Versions of Harsanyi's theorem exist for a number of these other expected utility theories.[3] Here, we use state-contingent alternatives with a finite number of states to model uncertainty, as in Arrow (1953, 1964), but with state probabilities that are common to all individuals. Versions of Harsanyi's theorem have been established for state-contingent alternatives by Blackorby, Donaldson, and Weymark (1980, 1999) and Hammond (1981, 1983) using alternative regularity conditions on the profile being aggregated.

On a number of different domains – single utility profile, multiprofile, single information set, and single preference profile – we investigate the implications of requiring the individual utility functions and the social preferences to satisfy the expected utility hypothesis when the social evaluation functional is required to satisfy the Strong Pareto principle and, in some cases, some additional axioms. In particular, we investigate the extent to which our results provide support for weighted utilitarianism.

Because we use individual utility functions rather than preferences, we must determine what utility functions qualify as satisfying the expected utility hypothesis. A utility function is an expected Bernoulli utility function if the utility of an alternative is the expected value of the utility obtained in each state. An expected Bernoulli utility function is the analogue of a von Neumann–Morgenstern utility function for state-contingent alternatives. We argue that any utility function that is an increasing transform of an expected Bernoulli utility function satisfies the expected utility hypothesis. However, because our results are sensitive to this choice, we also consider

[3] See Mongin and d'Aspremont (1998) for a discussion of this literature.

the implications of restricting the domain so that it includes only expected Bernoulli utility functions. Our theorems provide support for weighted utilitarianism only when we make this assumption and utility gains and losses are interpersonally comparable.[4] Because expected utility theory provides no good reason for restricting the domain a priori to profiles of expected Bernoulli utility functions, we conclude that Harsanyi's social aggregation theorem, by itself, does not provide a compelling argument in support of weighted utilitarianism, even when it is reformulated to accommodate interpersonal utility comparisons and, possibly, multiprofile social aggregation.

6.2 State-Contingent Alternatives and Social Evaluation Functionals

As noted in Section 6.1, we use state-contingent alternatives with a finite number of states to model uncertainty, as in Arrow (1953, 1964).[5] There are M states of nature, $m = 1, \ldots, M$, with $M \geq 2$. In state m, the set of feasible alternatives is S and an element $x \in S$ is a *state-contingent alternative*. We assume that S is a bounded connected subset of a finite-dimensional Euclidean space. Many interpretations of the elements in S are possible. For example, x could be a complete specification of how much each individual consumes of each commodity and of how much of each good each firm uses as an input or produces as an output, or it could be the amount spent by a government on various public goods and services. We leave the interpretation of S open. The set of *social alternatives* is S^M, the M-fold Cartesian product of S. A typical element of S^M is $\mathbf{x} := (x_1, \ldots, x_M)$, where x_m is the outcome in state m. Decisions are made before the uncertainty is resolved, so social alternatives are the objects of choice. To emphasize the stochastic aspects of a social alternative $\mathbf{x}$, we sometimes refer to $\mathbf{x}$ as a *prospect*. If a state-contingent alternative x occurs for certain, we denote this by $\mathbf{x}^c := (x, \ldots, x)$.

The probability of state m occurring is $p_m > 0$ and $p := (p_1, \ldots, p_M)$ is the vector of probabilities across all states. Probabilities take on fixed values that are commonly agreed on by all individuals – either there are objective probabilities or everyone's subjective probabilities are the same. Without this agreement, it is easy to construct social choice impossibility theorems (see, for example, Broome (1991) or Mongin and d'Aspremont (1998)).

[4] Our multiprofile theorem for profiles of expected Bernoulli utility functions is closely related to the main theorem in Mongin (1994). See the discussion following Theorem 6.8. for details.

[5] The use of a finite number of states rather than an atomless state space distinguishes this model from that of Savage (1954).

The set of individuals in society is $N := \{1, \ldots, n\}$ with $n \geq 2$. Each person has a utility function $U_i \colon S^M \to \mathbb{R}$, where $\mathbb{R}$ denotes the real numbers. Thus, $U_i(\mathbf{x})$ is the utility person i obtains from prospect $\mathbf{x}$. A *profile* of utility functions is an n-tuple $U := (U_1, \ldots, U_n)$. A profile is, therefore, a vector-valued function $U \colon S^M \to \mathbb{R}^n$ whose value at $\mathbf{x}$ is $U(\mathbf{x}) := (U_1(\mathbf{x}), \ldots, U_n(\mathbf{x}))$. The set of all possible utility functions is $\mathcal{U}$ and the set of all possible profiles is $\mathcal{U}^n$.

We allow for the possibility of interpersonal utility comparisons by using a social evaluation functional to model the social aggregation procedure. The set of *admissible profiles* – the domain of the social evaluation functional – is $\mathcal{D}$. We work with restricted domains, so $\mathcal{D}$ is a subset of $\mathcal{U}^n$. Alternative possibilities for $\mathcal{D}$ are considered in subsequent sections. A *social evaluation* is an ordering R of S^M, the set of prospects.[6] (The corresponding strict preference and indifference relations are P and I, respectively.) The set of all possible orderings of S^M is $\mathcal{O}$. A social evaluation must be chosen from the set of *admissible social evaluations* $\mathcal{R}$, the range of the functional, where $\mathcal{R} \subseteq \mathcal{O}$. The exact specification of $\mathcal{R}$ is considered in Section 6.4. Thus, a social evaluation functional is a mapping $f \colon \mathcal{D} \to \mathcal{R}$. For the profile $U \in \mathcal{D}$, the corresponding social evaluation is $R_U := f(U)$.

We consider both the Strong Pareto principle and the weaker Pareto Indifference condition. *Pareto Indifference* requires any pair of alternatives to be ranked as socially indifferent when every individual is equally well off in each of them.

Pareto Indifference: For all $U \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y} \in S^M$, if $U(\mathbf{x}) = U(\mathbf{y})$, then $\mathbf{x} I_U \mathbf{y}$.

Strong Pareto strengthens Pareto Indifference by also requiring $\mathbf{x}$ to be socially preferred to $\mathbf{y}$ when at least one person is better off with $\mathbf{x}$ than with $\mathbf{y}$ and no one is worse off.

Strong Pareto: For all $U \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y} \in S^M$, (a) if $U(\mathbf{x}) = U(\mathbf{y})$, then $\mathbf{x} I_U \mathbf{y}$ and (b) if $U(\mathbf{x}) \geq U(\mathbf{y})$ and $U_j(\mathbf{x}) > U_j(\mathbf{y})$ for some $j \in N$, then $\mathbf{x} P_U \mathbf{y}$.

Harsanyi (1955, 1977) argues that his axioms imply that the social aggregation functional is weighted utilitarian. In our model, a social evaluation functional f is *weighted utilitarian* if there exist weights $\lambda^1, \ldots, \lambda^n$

[6] An *ordering* is a reflexive, complete, and transitive binary relation.

such that

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i U_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i U_i(\mathbf{y}) \tag{6.1}$$

for all $U \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y} \in S^M$. A weighted utilitarian social evaluation functional satisfies Pareto Indifference. It also satisfes Strong Pareto if all the weights are positive. If the weights are equal and positive, Eq. (6.1) defines the *utilitarian* social evaluation functional for the domain $\mathcal{D}$.

6.3 Interpersonal Utility Comparisons and Information Partitions

The social evaluation functionals that can be considered are limited by our ability to make intrapersonal and interpersonal comparisons of utility. For example, the social evaluations for a utilitarian social evaluation functional are determined by comparing utility sums for different prospects and this requires that interpersonal comparisons of utility gains and losses are possible. The framework introduced in the previous section is flexible enough to allow for various assumptions concerning the measurability and interpersonal comparability of utility. This is accomplished by partitioning the set of admissible profiles $\mathcal{D}$ into equivalence classes called *information sets* within which all profiles are judged to be informationally equivalent. By this we mean that all profiles in the same information set contain the same usable information about individual utilities. Because profiles in the same information set are indistinguishable, they must all be mapped by the social evaluation functional f into the *same* social ordering of the alternatives in S^M. For example, in the problem considered by Arrow (1951), individual utility functions are ordinally measurable and there are no interpersonal comparisons of utilities possible. Thus, if $\bar{U} = (\bar{U}_1, \ldots, \bar{U}_n)$ and $\widehat{U} = (\widehat{U}_1, \ldots, \widehat{U}_n)$ are both in $\mathcal{D}$ and if each $\bar{U}_i$ is an increasing transform of the corresponding $\widehat{U}_i$ (with possibly different transforms used for different individuals), then $R_{\bar{U}} = R_{\widehat{U}}$.

To make these ideas precise, we assume that there is an *information partition* $A := \{A_t \mid t \in \mathcal{T}\}$ of $\mathcal{D}$, where $\mathcal{T}$ indexes the elements of the partition and each A_t is an information set. Utility profiles in different information sets are informationally distinguishable, while profiles in the same information set are not. Let $\mathcal{A}_{\mathcal{D}}$ denote the set of all information partitions of $\mathcal{D}$. A social evaluation functional f must be constant on an information set, a property of f we call *Information Invariance with Respect to the Partition A.*

Information Invariance with Respect to the Partition A**:** For all $\bar{U}, \widehat{U} \in \mathcal{D}$, if $\bar{U}, \widehat{U} \in A_t$ for some $t \in \mathcal{T}$, then $R_{\bar{U}} = R_{\widehat{U}}$.

Consider the partitions A and A' of $\mathcal{D}$ and suppose, for example, that A' is a finer partition than A. There must therefore exist utility profiles that are in different information sets in the partition A' but that are in the same information set in the partition A. When the information partition is A', the two profiles can be informationally distinguished and can be assigned different social evaluations by the social evaluation functional. When the information partition is A, the two profiles cannot be distinguished and must be assigned the same social evaluation. Thus, the restrictiveness of the information invariance condition is inversely related to the fineness of the partition of the domain into information sets.

In Arrovian social choice theory, the only usable utility information in a utility profile U is the information contained in the profile of preference orderings implicitly defined by U. A *preference profile* is an n-tuple $\mathbf{R} := (R_1, \ldots, R_n)$ of individual preference orderings on S^M. A preference ordering R_i on S^M can be obtained from the utility function U_i by setting

$$\mathbf{x} R_i \mathbf{y} \longleftrightarrow U_i(\mathbf{x}) \geq U_i(\mathbf{y}) \tag{6.2}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$.[7] Utility functions are *ordinally measurable and interpersonally noncomparable* if and only if subjecting the individual utility functions to (person-specific) increasing transforms results in a profile that is informationally equivalent. In other words, utility profiles are informationally equivalent if and only if they represent the same preference profile. The corresponding information partition is called the *ordinal noncomparable partition*, denoted A^{ON}, and the social evaluation functional satisfies *Information Invariance with Respect to Ordinal Noncomparable Utilities.*

Ordinal Noncomparable Partition: For all $\bar{U}, \widehat{U} \in \mathcal{D}$, there exists a $t \in \mathcal{T}$ such that $\bar{U}, \widehat{U} \in A_t$ if and only if, for each $i \in N$, there exists an increasing function $\phi_i \colon \mathbb{R} \to \mathbb{R}$ such that $\bar{U}_i(\mathbf{x}) = \phi_i(\widehat{U}_i(\mathbf{x}))$ for all $\mathbf{x} \in S^M$.

Utility functions are *cardinally measurable and interpersonally noncomparable* if and only if applying positive affine transforms to the individual utility functions (with possibly different transforms for different individuals) results in an informationally equivalent profile. This class of transforms preserves intrapersonal comparisons of utility levels and utility differences,

[7] The utility function U_i represents the preference ordering R_i if Eq. (6.2) holds and the utility profile U represents the preference profile $\mathbf{R}$ if Eq. (6.2) holds for all $i \in N$.

but no interpersonal utility comparisons are possible. In this case, the information partition is called the *cardinal noncomparable partition*, denoted A^{CN}, and the social evaluation functional satisfies *Information Invariance with Respect to Cardinal Noncomparable Utilities.*

Cardinal Noncomparable Partition: For all $\bar{U}, \widehat{U} \in \mathcal{D}$, there exists a $t \in \mathcal{T}$ such that $\bar{U}, \widehat{U} \in A_t$ if and only if, for all $i \in N$, there exist scalars α_i and β_i with $\beta_i > 0$ such that $\bar{U}_i(\mathbf{x}) = \alpha_i + \beta_i \widehat{U}_i(\mathbf{x})$ for all $\mathbf{x} \in S^M$.

The utility functions are *cardinally measurable and unit comparable* if subjecting the individual utility functions to positive affine transforms with common unit-scaling parameters results in an informationally equivalent profile. The corresponding information partition is called the *cardinal unit-comparable partition*, denoted A^{CU}, and the social evaluation functional satisfies *Information Invariance with Respect to Cardinal Unit-Comparable Utilities.*

Cardinal Unit-Comparable Partition: For all $\bar{U}, \widehat{U} \in \mathcal{D}$, there exists a $t \in \mathcal{T}$ such that $\bar{U}, \widehat{U} \in A_t$ if and only if there exist scalars $\alpha_1, \ldots, \alpha_n, \beta$ with $\beta > 0$ such that for all $i \in N$, $\bar{U}_i(\mathbf{x}) = \alpha_i + \beta \widehat{U}_i(\mathbf{x})$ for all $\mathbf{x} \in S^M$.

The partitions A^{CU} and A^{CN} are both finer than A^{ON}, and so intrapersonal comparisons of utility levels are meaningful in both cases. With both of these partitions, interpersonal comparisons of utility levels are not meaningful because such comparisons may not be preserved if the utility functions are transformed using positive affine transforms with common unit-scaling parameters. However, intrapersonal and interpersonal comparisons of utility differences (gains and losses) can be made when the partition is A^{CU} because such comparisons are invariant to transforms within this class. Formally, if there exist scalars $\alpha_1, \ldots, \alpha_n, \beta$ with $\beta > 0$ such that for all $i \in N$, $\bar{U}_i = \alpha_i + \beta \widehat{U}_i$, then for all $\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \in S^M$, we have $\bar{U}_i(\mathbf{x}) - \bar{U}_i(\mathbf{y}) \geq \bar{U}_j(\bar{\mathbf{x}}) - \bar{U}_j(\bar{\mathbf{y}})$ if and only if $\widehat{U}_i(\mathbf{x}) - \widehat{U}_i(\mathbf{y}) \geq \widehat{U}_j(\bar{\mathbf{x}}) - \widehat{U}_j(\bar{\mathbf{y}})$ for all $i, j \in N$ (i and j need not be distinct).[8] With the partition A^{CN}, a different β can be used for each individual, and so we can only conclude that these difference comparisons are meaningful for a single individual.

If every utility profile is in a distinct element of the partition, then any two profiles can be distinguished. As a consequence, the numerical values of all

[8] Without further assumptions, the reverse implication need not hold. See Bossert and Weymark (2004) for a discussion of the relationship between affine and difference-preserving transforms as well as for further references to the relevant literature.

utilities are significant and any kind of intrapersonal or interpersonal utility comparison is possible. We call this information partition the *numerically comparable partition*, denoted A^{NC}. In this case, the information invariance assumption is vacuous.

Numerically Comparable Partition: For all $\bar{U}, \widehat{U} \in \mathcal{D}$, there exists a $t \in \mathcal{T}$ such that $\bar{U}, \widehat{U} \in A_t$ if and only if $\bar{U} = \widehat{U}$.

For some results, a precise specification of the information partition is not required. Instead, the ability to discriminate between profiles of utility functions must be at least as good as is possible with the information partition A^{CU}. In other words, we suppose that the actual information partition A is in the collection of information partitions that are refinements (in the weak sense) of A^{CU}. This is the set $\mathcal{A}_{+}^{CU} := \{A \in \mathcal{A}_{\mathcal{D}} \mid A$ is a refinement of $A^{CU}\}$. When we require the information partition A to be in $\mathcal{A}_{+}^{CU}$, we say there is a *cardinal unit-comparable plus partition.*

Cardinal Unit-Comparable Plus Partition: $A \in \mathcal{A}_{+}^{CU}$.

A number of other information partitions have been considered in the literature. For further examples, see the surveys by Blackorby, Donaldson, and Weymark (1984), Bossert and Weymark (2004), d'Aspremont (1985), and Sen (1977).

If the social evaluation functional f is informationally invariant with respect to the partition A, it necessarily is informationally invariant with respect to any finer partition of the domain. As a consequence, it is generally not possible to infer what the information partition is from knowledge of f alone. If f is informationally invariant with respect to the partition A, then all profiles in the same element of the partition must be assigned the same social evaluation. However, this does not preclude assigning the same social evaluation to profiles that are not informationally equivalent. It is this fact that limits our ability to recover the information partition from knowledge of f. For example, if f satisfies Information Invariance with Respect to Cardinal Unit-Comparable Utilities, it may well be the case that the information partition is A^{NC}, but all profiles in the same element of the partition A^{CU} are assigned the same social evaluation, even though the information available permits us to distinguish between some of these profiles and to assign them different social evaluations. Although it may not be possible to infer the actual information partition from f, an upper bound on the coarseness of the partition can be determined by placing profiles in the same element of the partition if and only if they result in the same social evaluation. Thus, if there is a distinct social evaluation for each profile in the domain, the information profile must be A^{NC}, the numerically comparable partition. This is the one case in which it is possible to determine the information partition uniquely from f.

In Eq. (6.1), the social evaluation is unchanged if, for each $i \in N$, U_i is replaced with $\bar{U}_i = \alpha_i + \beta U_i$, where $\beta > 0$. Thus, a weighted utilitarian social evaluation functional satisfies Information Invariance with Respect to Cardinal Unit-Comparable Utilities and, hence, is informationally invariant with respect to any partition in $\mathcal{A}^{CU}_{+}$.

6.4 Expected Utility Theory for State-Contingent Alternatives

We require all individual utility functions and all social evaluations to satisfy the expected utility hypothesis. Expected utility theory was developed as a theory of individual behaviour under uncertainty. Because of this focus on individual behaviour, expected utility theory is ordinal and the primitive of this theory is typically taken to be a preference ordering. In contrast, here, utility has welfare significance. This raises the question: Which utility functions can be said to satisfy the expected utility hypothesis? We consider two possible answers to this question.

The most obvious, but not necessarily the most appropriate, way of answering this question is to say that a utility function satisfies the expected utility hypothesis if the utility of a prospect is the expected value of the utilities obtained in each state. Following Broome (1991), we refer to this as the Bernoulli hypothesis. Formally, for each $i \in N$, the utility function $U_i \in \mathcal{U}$ satisfies the *Bernoulli hypothesis* if there exists a continuous function $V_i\colon S \to \mathbb{R}$ such that

$$U_i(\mathbf{x}) = \sum_{m=1}^{M} p_m V_i(x_m) \tag{6.3}$$

for all $\mathbf{x} \in S^M$. We define

$$EV_i(\mathbf{x}) := \sum_{m=1}^{M} p_m V_i(x_m) \tag{6.4}$$

for all $\mathbf{x} \in S^M$ and all $i \in N$. As in Arrow (1965), we refer to V_i as a *Bernoulli utility function* and EV_i as an *expected Bernoulli utility function*. A Bernoulli utility function is defined on the set of state-contingent alternatives S and is state-independent. Because V_i is continuous on S, EV_i is continuous on S^M. $V_i(x_m)$ is the utility obtained ex post with the state-contingent alternative x_m. Before the uncertainty is resolved, the ex ante utility is given by the expected value of these ex post utilities. We let $\mathcal{B}$ denote the set of all utility functions on S^M that satisfy the Bernoulli hypothesis. The profile U is a *Bernoulli expected utility profile* if $U \in \mathcal{B}^n$.

In the model introduced in Section 6.2, utility is an attribute of a prospect, not of a state-contingent alternative. However, in order for Eq. (6.3) to make sense, utilities must also be well-defined for state-contingent alternatives. If the outcome is $x \in S$ with certainty, Eqs. (6.3) and (6.4) simplify to

$$U_i(\mathbf{x}^c) = EV_i(\mathbf{x}^c) = V_i(x). \tag{6.5}$$

Thus, we can think of the utility associated with $x \in S$ as being the utility of facing the prospect that has outcome x in every state.

In Eq. (6.3), we have supposed that the Bernoulli utility function V_i is continuous and state-independent, and both of these assumptions may seem rather arbitrary. In expected utility theory, a utility function on prospects is not a primitive of the theory as it is here; rather, the theory starts with individual preferences, and utility functions are merely representations of these preferences. In Arrow's version of expected utility theory, preferences have representations of the form given in Eq. (6.4). That is, each individual $i \in N$ has a preference relation (a binary relation) R_i on S^M, and this preference is assumed to satisfy one of a number of equivalent sets of axioms collectively known as the *expected utility axioms.*[9] These axioms imply that each of the preference relations R_i can be represented by a utility function on S^M of the form given in Eq. (6.4); that is, for each $i \in N$, there exists a continuous function $V_i \colon S \to \mathbb{R}$ such that

$$\mathbf{x} R_i \mathbf{y} \longleftrightarrow \sum_{m=1}^{M} p_m V_i(x_m) \geq \sum_{m=1}^{M} p_m V_i(y_m) \tag{6.6}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. If we start with preferences, continuity and state-independence of the Bernoulli utility functions are not assumptions of the theory; they are instead implications of the expected utility axioms.

It might seem that this representation theorem provides a justification for restricting attention to utility functions that satisfy the Bernoulli hypothesis when all individuals have utility functions that satisfy the expected utility hypothesis. However, as with any utility representation theorem, the

[9] Hens (1992) discusses a number of these axiom systems, including one of his own. In these axiomatizations, more structure is placed on the set S than is done here. Broome (1991) provides a detailed defence of the axioms of expected utility theory. For an introduction to axiomatizing expected utility with state-contingent alternatives, see Blackorby, Davidson, and Donaldson (1977).

representation in Eq. (6.6) is not unique. As is well known, if for all $x \in S$, $V_i'(x) = \alpha + \beta V_i(x)$ with $\beta > 0$ (that is, V_i' is a positive affine transform of V_i), then Eq. (6.6) is also satisfied with V_i' replacing V_i. Further, because $V_i(S)$ – the set of ex post utilities attainable with some $x \in S$ – is an interval of $\mathbb{R}$, *only* positive affine transforms of V_i satisfy Eq. (6.6) (unless V_i is a constant-valued function). If V_i' is a positive affine transform of V_i, then EV_i' is the same positive affine transform of EV_i. These observations have often been interpreted as implying that only expected Bernoulli utility functions are admissible representations of R_i. But the axioms of expected utility theory apply to R_i, so if EV_i represents R_i, then so does *any* increasing transform of EV_i, not just any positive affine transform of EV_i.

We start with utility functions, not preferences, but any utility function implicitly defines a preference, as in Eq. (6.2). It is natural to say that the utility function U_i satisfies the expected utility hypothesis if the derived preference relation defined in Eq. (6.2) satisfies the expected utility axioms and so has a representation of the form given in Eq. (6.4). Formally, for each $i \in N$, $U_i \in \mathcal{U}$ satisfies the *expected utility hypothesis* if there exists an increasing function $\mathcal{V}_i : \mathbb{R} \to \mathbb{R}$ and a continuous function $V_i : S \to \mathbb{R}$ such that

$$U_i(\mathbf{x}) = \mathcal{V}_i \left[\sum_{m=1}^{M} p_m V_i(x_m) \right] \tag{6.7}$$

or, equivalently,

$$U_i(\mathbf{x}) = \mathcal{V}_i \left[EV_i(\mathbf{x}) \right] \tag{6.8}$$

for all $\mathbf{x} \in S^M$. In the special case of certain social alternatives, Eq. (6.7) simplifies to

$$U_i(\mathbf{x}^c) = \mathcal{V}_i \left[V_i(x) \right] \tag{6.9}$$

for all $x \in S$. We let $\mathcal{E}$ denote the set of all utility functions on S^M that satisfy the expected utility hypothesis. Clearly, $\mathcal{B}$ is a strict subset of $\mathcal{E}$. A profile $U \in \mathcal{E}^n$ is said to be an *expected utility profile.*

If U_i satisfies the Bernoulli hypothesis, there is only one Bernoulli utility function V_i for which Eq. (6.3) holds. In contrast, if U_i satisfies the expected utility hypothesis, there is no unique way to express U_i in the form Eq. (6.7). If V_i is replaced by an increasing affine transform of itself, we can always adjust the transform $\mathcal{V}_i$ so as to preserve the utility number $U_i(\mathbf{x})$ associated with each prospect $\mathbf{x}$.[10] For any utility function U_i in $\mathcal{E}$, we shall have to choose a particular transform $\mathcal{V}_i$ and a particular Bernoulli utility function V_i so as to be able to express U_i as in Eq. (6.7). Although this choice is arbitrary, nothing of substance depends on the functions that are chosen. If $\mathcal{V}_i$ is an increasing affine transform, then it is possible to apply the inverse of $\mathcal{V}_i$ (which is also an increasing affine transform) to V_i and thereby write U_i as an expected Bernoulli utility function. However, if $\mathcal{V}_i$ is not an affine transform, then U_i cannot be expressed as an expected Bernoulli utility function, and thus U_i does not satisfy the Bernoulli hypothesis.

It is sometimes suggested that the only representations of an expected utility preference R_i that preserve attitudes toward risk are representations of the form Eq. (6.4).[11] If this were the case, we might have an argument for restricting attention to utility functions in $\mathcal{B}$. The argument is that attitudes toward risk are captured by the curvature properties of the Bernoulli utility function V_i and the relevant curvature properties are only preserved by positive affine transforms of V_i. Put somewhat differently, measures of risk-aversion defined using Bernoulli utility functions, such as those of Arrow (1965) and Pratt (1964) for the case in which $S \subseteq \mathbb{R}$, are only well defined if an expected Bernoulli utility function is used to represent R_i because these measures are not invariant to nonaffine transforms of a Bernoulli utility function. However, because the primitive of expected utility theory is a preference relation, all of the meaningful properties of any utility representation, including risk attitudes and the curvature properties of a Bernoulli utility function, must be inherited from properties of the preference relation. Further, whether we use a representation of the form Eq. (6.4) or one of the form Eq. (6.7), a Bernoulli utility function is part of the representation. In either case, the *Bernoulli utility functions* in these representations are unique up to a positive affine transform, and we can use a Bernoulli function to construct a measure of risk aversion that only depends on the properties of the underlying preference relation, even if the actual utility function is non-Bernoulli. Thus, we cannot favour a representation of the form Eq. (6.4) over one of the form Eq. (6.7) by appealing to the desirability of having a utility representation that preserves attitudes toward risk.

In view of the preceding discussion, we see no compelling reason to regard a utility function as satisfying the expected utility hypothesis only

[10] This is not possible if V_i is subjected to a nonaffine transform because, as previously noted, Eq. (6.6) does not hold for nonaffine transforms of V_i.

[11] See Broome (1991, section 4.3), Mongin (1994, section 4), and Mongin and d'Aspremont (1998, section 5.3) for discussions of this point.

if it is in $\mathcal{B}$. This is not to say that there may not be other reasons for restricting attention to utility functions in $\mathcal{B}$. For example, Broome (1991, section 6.5) proposes an interpretation of utility that requires comparing differences in the goodness of outcomes across states, which is only possible in his framework if the Bernoulli hypothesis is satisfied. This interpretation requires utility to be cardinally measurable and therefore goes beyond what expected utility theory can deliver.[12] To allow for these considerations as well as to gain insight into the role the transforms $\mathcal{V}_i$ play in the analysis, we consider utility functions in both $\mathcal{B}$ and $\mathcal{E}$ in the subsequent discussion.

For the profile of Bernoulli utility functions $V := (V_1, \ldots, V_n)$, the set of *feasible Bernoulli utility vectors* is

$$V(S) := \{u \in \mathbb{R}^n \mid u = (V_1(x), \ldots, V_n(x)) \text{ for some } x \in S\}. \tag{6.10}$$

An expected utility profile is *regular* if it can be expressed in terms of a profile of Bernoulli utility functions V for which $V(S)$ is full-dimensioned and well-behaved (in a sense made precise in the following definition).

Regular Expected Utility Profile: An expected utility profile $U \in \mathcal{E}^n$ is regular if Eq. (6.7) holds for each $i \in N$ using a profile of Bernoulli utility functions V for which $V(S)$ has a nonempty connected interior with $V(S)$ contained in the closure of its interior.[13]

Note that whether $V(S)$ satisfies these properties is independent of which Bernoulli utility functions are chosen to express the U_i as in Eq. (6.7). When $V(S)$ has a nonempty interior, for each $i \in N$, there exists a pair of state-contingent alternatives $\{x^i, y^i\}$ such that $V_i(x^i) \neq V_i(y^i)$ and $V_j(x^i) = V_j(y^i)$ for all $j \neq i$. From Eq. (6.7) it then follows that, in comparing the prospects in which either x^i or y^i are obtained for certain, only the utility of individual i is affected. This is a preference diversity assumption.[14] The rest of our regularity assumption rules out profiles that are in some sense pathological.

A social evaluation functional f assigns a social evaluation R_U to each profile $U \in \mathcal{D}$. Each of these social evaluations is required to satisfy the expected utility hypothesis, thereby restricting the range $\mathcal{R}$ of the functional. A

[12] See also Broome (2008) and Weymark (2005).

[13] A regular expected utility profile is called *strongly regular* in Blackorby, Donaldson, and Weymark (1999). They also consider a weaker regularity condition in which the requirement that the interior of $V(S)$ is nonempty is replaced with the requirement that the relative interior of $V(S)$ is nonempty.

[14] If the dimension of S is at least n, as would be the case if there are private goods, it is natural to suppose that this preference diversity condition is satisfied.

social evaluation is a binary relation, not a utility function, and so the identification of which social evaluations satisfy the expected utility hypthosis is straightforward. An ordering $R \in \mathcal{O}$ satisfies the *expected utility hypothesis* (for orderings) if there is a continuous Bernoulli utility function $F: S \to \mathbb{R}$ such that

$$\mathbf{x} R \mathbf{y} \longleftrightarrow \sum_{m=1}^{M} p_m F(x_m) \geq \sum_{m=1}^{M} p_m F(y_m) \tag{6.11}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. We define $EF\colon S^M \to \mathbb{R}$ by setting

$$EF(\mathbf{x}) := \sum_{m=1}^{M} p_m F(x_m) \tag{6.12}$$

for all $\mathbf{x} \in S^M$. Of course, any increasing transform of EF represents R just as well. We let $\mathcal{R}_E$ denote the set of all $R \in \mathcal{O}$ that satisfy the expected utility hypothesis for orderings and we restrict the range of the social evaluation functional to be $\mathcal{R}_E$. When the range is restricted in this way, f has an *unrestricted expected utility range.*

Unrestricted Expected Utility Range: $\mathcal{R} = \mathcal{R}_E$.

6.5 Welfarism

Welfarism requires all social orderings of the alternatives to be determined solely on the basis of the individual utilities obtained with them. Any weighted utilitarian social evaluation functional is welfarist. Axiomatic characterizations of welfarism when the domain of the social evaluation functional f is $\mathcal{U}^n$ (the set of all possible profiles) have been obtained by d'Aspremont and Gevers (1977), Hammond (1979), and Sen (1977). In this section, we show that these characterizations also hold when the domain is either $\mathcal{B}^n$ (the set of all Bernoulli expected utility profiles) or $\mathcal{E}^n$ (the set of all expected utility profiles). We also consider a profile-dependent form of welfarism in which social evaluations are permitted to depend not only on the utilities obtained with the alternatives, but also on the profile that generates them.

An ordering R^* of a set of utility vectors in $\mathbb{R}^n$ is called a *social welfare ordering* and a representation W of R^* (if one exists) is called a *social welfare function.* In our profile-dependent form of welfarism, each social evaluation R_U can be completely determined by a social welfare ordering on the set of utility vectors that are obtainable with the profile U. For any profile $U \in \mathcal{D}$, the set of *feasible utility vectors for* U is

$$U(S^M) := \{u \in \mathbb{R}^n \mid u = U(\mathbf{x}) \text{ for some } \mathbf{x} \in S^M\}. \tag{6.13}$$

The social welfare ordering R_U^* on $U(S^M)$ is *isomorphic* to the social evaluation R_U on S^M if for all $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow U(\mathbf{x}) R_U^* U(\mathbf{y}). \tag{6.14}$$

When Eq. (6.14) is satisfied, the social evaluation R_U can be completely recovered from knowledge of the social welfare ordering R_U^* and the profile U. A social evaluation functional satisfies *Profile-Dependent Welfarism* if Eq. (6.14) holds for every profile in its domain.

Profile-Dependent Welfarism: For all $U \in \mathcal{D}$, there exists a social welfare ordering R_U^* on $U(S^M)$ isomorphic to R_U.

In order for profile-dependent welfarism to be satisfied, the social ordering of any two alternatives must depend only on the utilities obtained from these alternatives and the profile of utility functions that generate them and not on nonutilty information contained in the descriptions of the alternatives. In other words, for any fixed profile, alternatives are treated in a neutral manner, a property we call *Profile-Dependent Strong Neutrality*.

Profile-Dependent Strong Neutrality: For all $U \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \in S^M$, if $U(\mathbf{x}) = U(\bar{\mathbf{x}})$ and $U(\mathbf{y}) = U(\bar{\mathbf{y}})$, then $\mathbf{x} R_U \mathbf{y}$ if and only if $\bar{\mathbf{x}} R_U \bar{\mathbf{y}}$.

Clearly, Profile-Dependent Welfarism implies Profile-Dependent Strong Neutrality. Further, by setting $\bar{\mathbf{x}} = \mathbf{y}$ and $\bar{\mathbf{y}} = \mathbf{x}$, Profile-Dependent Strong Neutrality implies Pareto Indifference. Theorem 6.1 demonstrates that, in fact, these three conditions are equivalent for any social evaluation functional whose domain $\mathcal{D}$ is contained in $\mathcal{U}^n$ and whose range $\mathcal{R}$ is contained in $\mathcal{O}$.

Theorem 6.1: *A social evaluation functional $f: \mathcal{D} \to \mathcal{R}$ satisfies Pareto Indifference if and only if it satisfies Profile-Dependent Strong Neutrality if and only if it satisfies Profile-Dependent Welfarism.*

Proof: See Propositions 1 and 2 in Blackorby, Donaldson, and Weymark (1990).[15]

[15] Blackorby, Donaldson, and Weymark established their propositions for any set of social alternatives containing at least three elements. Our assumptions on S ensure that S^M contains an infinite number of alternatives.

With Profile-Dependent Welfarism, a social ordering of a pair of alternatives need not depend only on the utilities obtained with these alternatives; it can also depend on the profile of utility functions that generates the utilities. Welfarism eliminates any dependence of the social evaluation on the way in which utilities are obtained. When the domain of the social evaluation functional is $\mathcal{D}$, the set of *feasible utility vectors* is

$$\mathbf{U}_{\mathcal{D}} := \bigcup_{U \in \mathcal{D}} U(S^M). \tag{6.15}$$

A vector of utilities u is in $\mathbf{U}_{\mathcal{D}}$ if there is some profile U in the domain $\mathcal{D}$ and some alternative $\mathbf{x}$ in S^M such that $U(\mathbf{x}) = u$. If the domain is sufficiently rich, $\mathbf{U}_{\mathcal{D}}$ can be all of $\mathbb{R}^n$. A social evaluation functional f satisfies *Welfarism* if f can be equivalently described by a *single* social welfare ordering R^* on $\mathbf{U}_{\mathcal{D}}$.

Welfarism: There exists a social welfare ordering R^* on $\mathbf{U}_{\mathcal{D}}$ such that for all $U \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow U(\mathbf{x}) R^* U(\mathbf{y}). \tag{6.16}$$

When Eq. (6.16) is satisfied, the social welfare ordering R^* is *isomorphic* to the social evaluation functional f. Welfarism implies Profile-Dependent Welfarism. For a welfarist social evaluation functional, each of the social welfare orderings R^*_U defined in Eq. (6.14) is simply the restriction of R^* to $U(S^M)$.

Strong Neutrality strengthens Profile-Dependent Strong Neutrality by requiring the social evaluation functional to ignore all nonwelfare characteristics of the alternatives. In other words, it is irrelevant how a vector of utilities is obtained. The only relevant feature of an alternative $\mathbf{x}$ and a profile U is the vector of utilities $u = U(\mathbf{x})$ obtained with U and $\mathbf{x}$.

Strong Neutrality: For all $\bar{U}, \widehat{U} \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \in S^M$, if $\bar{U}(\mathbf{x}) = \widehat{U}(\bar{\mathbf{x}})$ and $\bar{U}(\mathbf{y}) = \widehat{U}(\bar{\mathbf{y}})$, then $\mathbf{x} R_{\bar{U}} \mathbf{y}$ if and only if $\bar{\mathbf{x}} R_{\widehat{U}} \bar{\mathbf{y}}$.

Binary Independence of Irrelevant Alternatives requires the social evaluation of any pair of alternatives to be independent of any utility information about the other alternatives. Binary Independence is simply Strong Neutrality restricted to comparisons of the same pair of alternatives across profiles.

Binary Independence of Irrelevant Alternatives: For all $\bar{U}, \widehat{U} \in \mathcal{D}$ and all $\mathbf{x}, \mathbf{y} \in S^M$, if $\bar{U}(\mathbf{x}) = \widehat{U}(\mathbf{x})$ and $\bar{U}(\mathbf{y}) = \widehat{U}(\mathbf{y})$, then $\mathbf{x} R_{\bar{U}} \mathbf{y}$ if and only if $\mathbf{x} R_{\widehat{U}} \mathbf{y}$.

In general, this independence axiom is weaker than the standard Arrovian independence axiom, which replaces the antecedent with the condition that the individual *rankings* of $\mathbf{x}$ and $\mathbf{y}$ are the same in both profiles. With our independence axiom, the actual vectors of utility numbers obtained with $\mathbf{x}$ and $\mathbf{y}$ must be the same in both profiles for the axiom to apply. If the social evaluation functional satisfies Information Invariance with Respect to Ordinal Noncomparable Utilities (so we are in the Arrow framework), then our independence axiom is satisfied if and only if the usual Arrovian independence axiom is satisfied.

For any domain and range, it is easy to verify that Welfarism implies Strong Neutrality and Strong Neutrality implies both Pareto Indifference and Binary Independence of Irrelevant Alternatives. We now consider the reverse implications when the domain is either $\mathcal{B}^n$ (the set of all Bernoulli expected utility profiles) or $\mathcal{E}^n$ (the set of all expected utility profiles) and the social evaluation functional has an unrestricted expected utility range. When the domain is $\mathcal{B}^n$ (resp. $\mathcal{E}^n$), there is an *unrestricted Bernoulli expected utility domain* (resp. an *unrestricted expected utility domain*).

Unrestricted Bernoulli Expected Utility Domain: $\mathcal{D} = \mathcal{B}^n$.

Unrestricted Expected Utility Domain: $\mathcal{D} = \mathcal{E}^n$.

For a social evaluation functional f with domain $\mathcal{U}^n$ (the unrestricted domain) and range $\mathcal{O}$ (the set of all orderings of the alternatives), Strong Neutrality is equivalent to the joint satisfaction of Pareto Indifference and Binary Independence of Irrelevant alternatives. (See Theorem 2.3 in d'Aspremont (1985).) Theorem 6.2 shows that this equivalence also holds when there is either an unrestricted Bernoulli expected utility domain or an unrestricted expected utility domain and f has an unrestricted expected utility range.

Theorem 6.2: *If a social evaluation functional $f : \mathcal{D} \to \mathcal{R}$ has either an unrestricted Bernoulli expected utility domain or an unrestricted expected utility domain and has an unrestricted expected utility range, then f satisfies Pareto Indifference and Binary Independence of Irrelevant Alternatives if and only if it satisfies Strong Neutrality.*

Proof: It is clear that Strong Neutrality implies both Pareto Indifference and Binary Independence, so we only need to consider the reverse implication. First, suppose that the domain is $\mathcal{E}^n$. Consider any $\bar{U}, \widehat{U} \in \mathcal{E}^n$ and any $\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \in S^M$ for which $\bar{U}(\mathbf{x}) = \widehat{U}(\bar{\mathbf{x}})$ and $\bar{U}(\mathbf{y}) = \widehat{U}(\bar{\mathbf{y}})$. Let $u = \bar{U}(\mathbf{x}) = \widehat{U}(\bar{\mathbf{x}})$ and $v = \bar{U}(\mathbf{y}) = \widehat{U}(\bar{\mathbf{y}})$. Because S has an infinite number of elements, we can find a state-contingent alternative $z \in S$ such that z is distinct from any state-contingent alternative that is a possible outcome with $\mathbf{x}$, $\mathbf{y}$, $\bar{\mathbf{x}}$, or $\bar{\mathbf{y}}$. We want to construct profiles $U^1, U^2, U^3 \in \mathcal{E}^n$ such that (i) $U^1(\mathbf{x}) = U^1(\mathbf{z}^c) = u$ and $U^1(\mathbf{y}) = v$, (ii) $U^2(\bar{\mathbf{x}}) = U^2(\mathbf{z}^c) = u$ and $U^2(\bar{\mathbf{y}}) = v$, and (iii) $U^3(\mathbf{z}^c) = u$ and $U^3(\mathbf{y}) = U^3(\bar{\mathbf{y}}) = v$.

Consider any $i \in N$. We first construct a utility function U_i^1 with the requisite properties. Because $\bar{U}_i$ is in $\mathcal{E}$, $\bar{U}_i = \bar{\mathcal{V}}_i(E\bar{V}_i)$ for some Bernoulli utility function $\bar{V}_i$ on S and some increasing function $\bar{\mathcal{V}}_i \colon \mathbb{R} \to \mathbb{R}$. Similarly, we can express the utility function being constructed as $U_i^1 = \mathcal{V}_i^1(EV_i^1)$ for some Bernoulli utility function V_i^1 on S and some increasing function $\mathcal{V}_i^1 \colon \mathbb{R} \to \mathbb{R}$. In this construction, let $\mathcal{V}_i^1 = \bar{\mathcal{V}}_i$ and, for all $m = 1, \ldots, M$, let $V_i^1(x_m) = \bar{V}_i(x_m)$ and $V_i^1(y_m) = \bar{V}_i(y_m)$. Hence, $U_i^1(\mathbf{x}) = u$ and $U_i^1(\mathbf{y}) = v$. Because z is not one of the outcomes in $\mathbf{x}$ or $\mathbf{y}$, the value of V_i^1 at z has not yet been specified. Letting $V_i^1(z) = \mathcal{V}_i^{1^{-1}}(u_i)$, we have $U_i^1(\mathbf{z}^c) = u_i$, as desired. The values of V_i^1 have only been specified at a finite number of points in S; therefore, it is possible to define the other values of V_i^1 so that V_i^1 is a continuous function. The construction of U_i^2 can be dealt with in a similar fashion.

The function U_i^3 is chosen so that it is in $\mathcal{B}$. Let V_i^3 be the corresponding Bernoulli utility function on S. For all $m = 1, \ldots, M$, letting $V_i^3(y_m) = V_i^3(\bar{y}_m) = v_i$, we have $U_i^3(\mathbf{y}) = U_i^3(\bar{\mathbf{y}}) = v_i$. The value of $V_i^3(z)$ has not yet been determined. Setting $V_i^3(z) = u_i$, we have $U_i^3(\mathbf{z}^c) = u_i$. As in the preceding argument, the other values of V_i^3 can be defined so that V_i^3 is a continuous function.

By Binary Independence, $\mathbf{x} R_{\bar{U}} \mathbf{y} \leftrightarrow \mathbf{x} R_{U^1} \mathbf{y}$. Pareto Indifference and the transitivity of R_{U^1} imply that $\mathbf{x} R_{U^1} \mathbf{y} \leftrightarrow \mathbf{z}^c R_{U^1} \mathbf{y}$. A similar argument shows that $\mathbf{z}^c R_{U^1} \mathbf{y} \leftrightarrow \mathbf{z}^c R_{U^3} \mathbf{y} \leftrightarrow \mathbf{z}^c R_{U^3} \bar{\mathbf{y}}$. Applying the same argument once again, we have $\mathbf{z}^c R_{U^3} \bar{\mathbf{y}} \leftrightarrow \mathbf{z}^c R_{U^2} \bar{\mathbf{y}} \leftrightarrow \bar{\mathbf{x}} R_{U^2} \bar{\mathbf{y}}$. By Binary Independence, $\bar{\mathbf{x}} R_{U^2} \bar{\mathbf{y}} \leftrightarrow \bar{\mathbf{x}} R_{\widehat{U}} \bar{\mathbf{y}}$. We have thus shown that $\mathbf{x} R_{\bar{U}} \mathbf{y} \leftrightarrow \bar{\mathbf{x}} R_{\widehat{U}} \bar{\mathbf{y}}$, which completes the proof for the domain $\mathcal{E}^n$.

The proof for the domain $\mathcal{B}^n$ is the same, but with all the transforms chosen to be the identity function.[16]

When the domain is $\mathcal{U}^n$, $\mathcal{E}^n$, or $\mathcal{B}^n$, the set of feasible utility vectors $\mathbf{U}_{\mathcal{D}}$ is all of $\mathbb{R}^n$. As a consequence, if Welfarism is satisfied, there is a social

[16] Our proof of Theorem 6.2 is based on the proof in d'Aspremont (1985) of the corresponding result for the domain $\mathcal{U}^n$ and range $\mathcal{O}$. Only the transitivity of the social evaluations is used in the proof, so no modification to d'Aspremont's proof is needed to accommodate our range restriction. However, the requirement that the auxiliary profiles U^1, U^2, and U^3 must all be in either $\mathcal{E}^n$ or $\mathcal{B}^n$ presents complications that are not present when the domain is unrestricted.

welfare ordering R^* on all of $\mathbb{R}^n$ that is isomorphic to the social evaluation functional f. If the domain is $\mathcal{U}^n$ and the range $\mathcal{O}$, Welfarism and Strong Neutrality are equivalent restrictions on f. (See Theorem 2.2 in Blackorby, Donaldson, and Weymark (1984).) The same equivalence holds when the domain and range are restricted as in Theorem 6.2.

Theorem 6.3: *If a social evaluation functional $f : \mathcal{D} \to \mathcal{R}$ has either an unrestricted Bernoulli expected utility domain or an unrestricted expected utility domain and has an unrestricted expected utility range, then f satisfies Strong Neutrality if and only if it satisfies Welfarism.*

Proof: We only show that Strong Neutrality implies Welfarism as the reverse implication is trivial. The same proof applies to both of our domains. For any $u, v \in \mathbb{R}^n$, there exist $\mathbf{x}, \mathbf{y} \in S^M$ and there exists a $\bar{U} \in \mathcal{D}$ such that $\bar{U}(\mathbf{x}) = u$ and $\bar{U}(\mathbf{y}) = v$. For example, let $\mathbf{x} = \mathbf{x}^c$ and $\mathbf{y} = \mathbf{y}^c$ for any distinct $x, y \in S$ and, for each $i \in N$, let $\bar{U}_i = E\bar{V}_i$ for a Bernoulli utility function $\bar{V}_i$ on S for which $\bar{V}_i(x) = u_i$ and $\bar{V}_i(y) = v_i$. We let $uR^*v \leftrightarrow \mathbf{x}R_{\bar{U}}\mathbf{y}$ and $vR^*u \leftrightarrow \mathbf{y}R_{\bar{U}}\mathbf{x}$. For any other pair of alternatives $\bar{\mathbf{x}}, \bar{\mathbf{y}} \in S^M$ and any other profile $\widehat{U} \in \mathcal{D}$ such that $\widehat{U}(\bar{\mathbf{x}}) = u$ and $\widehat{U}(\bar{\mathbf{y}}) = v$, by Strong Neutrality, we obtain the same ordering of u and v, so R^* is well defined. Thus, there is a reflexive and complete binary relation R^* on $\mathbb{R}^n$ satisfying Eq. (6.16).

Now consider any $u, v, w \in \mathbb{R}^n$ with uR^*v and vR^*w. Either of our domains is rich enough to ensure that we can find three alternatives $\mathbf{x}, \mathbf{y}, \mathbf{z} \in S^M$ and a profile $U \in \mathcal{D}$ such that $U(\mathbf{x}) = u$, $U(\mathbf{y}) = v$, and $U(\mathbf{z}) = w$. By Eq. (6.16), we have $\mathbf{x}R_U\mathbf{y}$ and $\mathbf{y}R_U\mathbf{z}$. Transitivity of R_U then implies that $\mathbf{x}R_U\mathbf{z}$. Using Eq. (6.16) once more, we conclude that uR^*w, and so R^* is transitive.[17]

Combining Theorems 6.2 and 6.3, we see that Welfarism and Strong Neutrality are each equivalent to the joint satisfaction of Pareto Indifference and Binary Independence of Irrelevant Alternatives if the domain of the social evaluation functional is either $\mathcal{B}^n$ or $\mathcal{E}^n$ and its range is $\mathcal{R}_E$.[18]

In Section 6.8, we consider domains that are subsets of $\mathcal{B}^n$ or $\mathcal{E}^n$, but which include more than one profile. These domains may not be rich enough to

[17] This proof is essentially the same as in the case of an unrestricted domain. We have included it here to show how to construct the social welfare ordering R^*.

[18] Mongin (1994) has established versions of Theorems 6.2 and 6.3 when the set of alternatives is a convex subset of a vector space, the domain of the social evaluation functional is the set of all mixture-preserving utility functions on this set, and the range is the set of social evaluations that satisfy the mixture-set version of the expected utility hypothesis. Harsanyi's lottery set is a convex set. On a convex set of lotteries, a mixture-preserving utility function is a von Neumann–Morgenstern utility function.

ensure that the auxiliary profiles used in the proof of Theorem 6.2 exist, and so Strong Neutrality does not necessarily follow from Pareto Indifference and Binary Independence of Irrelevant Alternatives. However, these domains have the property that there is some profile U in the domain $\mathcal{D}$ for which the set of feasible utility vectors for U is all of the utility vectors feasible with the domain $\mathcal{D}$. This is a sufficiently rich domain for the equivalence between Strong Neutrality and Welfarism to hold.

Theorem 6.4: *If a social evaluation functional $f : \mathcal{D} \rightarrow \mathcal{R}$ has a domain contained in $\mathcal{E}^n$ that includes a profile U for which $U(S^M) = \mathbf{U}_{\mathcal{D}}$ and has an unrestricted expected utility range, then f satisfies Strong Neutrality if and only if it satisfies Welfarism.*

Proof: The proof is the same as the proof of Theorem 6.3 except (i) $\mathbf{U}_{\mathcal{D}}$ may be a strict subset of $\mathbb{R}^n$ and (ii) the profile U described in the theorem statement is used to define R^* on $\mathbf{U}_{\mathcal{D}}$ and to show that R^* is transitive. □

In the rest of this chapter, we use Strong Pareto, not Pareto Indifference. With Strong Pareto, each of the social welfare orderings R^*_U and R^* considered in this section is strictly monotonic; that is, u is socially prefered to v if $u_i \geq v_i$ for all $i \in N$, with a strict inequality for a least one individual.

6.6 Single-Profile Aggregation

In this section, we consider single-profile aggregation and suppose that the actual profile of utility functions is known. In order for this to be the case, utilities must be numerically comparable, so there are no informational invariance restrictions imposed on the social evaluation functional. We also assume that the profile satisfies either the expected utility or Bernoulli hypothesis and is regular. We identify all of the social evaluation functionals that satisfy Strong Pareto and have an unrestricted expected utility range when the domain is restricted to this single profile. Because there is only one profile U in the domain, a social evaluation functional is completely characterized by a single social evaluation – the social evaluation R_U assigned to U.

This problem differs from the aggregation problem considered by Harsanyi (1955, 1977) in a number of respects. First, Harsanyi aggregates a profile of individual preference orderings, not a profile of individual utility functions, into a social preference ordering. Second, we determine a social preference, whereas Harsanyi starts with a social preference and shows how it relates to the individual preferences. Third, we use state-contingent alternatives to model uncertainty, not lotteries.

We first suppose that the domain of the social evaluation functional is a regular expected utility profile. A complete characterization of the social evaluation functionals with an unrestricted expected utility range that satisfy Strong Pareto is provided by Theorem 6.5, which is a simple corollary of a result established by Blackorby, Donaldson, and Weymark (1999).

Theorem 6.5: *Suppose that $f : \mathcal{D} \rightarrow \mathcal{R}$ is a social evaluation functional with an unrestricted expected utility range and a domain consisting of the single regular expected utility profile $U \in \mathcal{E}^n$. For each $i \in N$, suppose that U_i is expressed as in Eq. (6.7) using the Bernoulli utility function V_i and the transform $\mathcal{V}_i$. Then, f satisfies Strong Pareto if and only if there exists a vector $\lambda \gg 0_n$, unique up to a positive factor of proportionality, such that for all $\mathbf{x}, \mathbf{y} \in S^M$,*

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(U_i(\mathbf{x})) \geq \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(U_i(\mathbf{y})).^{19} \tag{6.17}$$

Proof: Clearly, if Eq. (6.17) is satisfied and $\lambda \gg 0_n$, then f satisfies Strong Pareto, so we only need to consider the reverse implication. Because R_U satisfies the expected utility hypothesis, R_U can be represented as in Eq. (6.12) by an expected Bernoulli utility function $EF : S^M \rightarrow \mathbb{R}$ for some Bernoulli utility function $F : S \rightarrow \mathbb{R}$. By Theorem 6.4 in Blackorby, Donaldson, and Weymark (1999), there exists a unique vector $\lambda \gg 0_n$ and a unique scalar μ such that

$$EF(\mathbf{x}) = \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(U_i(\mathbf{x})) + \mu \tag{6.18}$$

for all $\mathbf{x} \in S^M$. In moving from Eq. (6.18) to Eq. (6.17), μ is eliminated, and so λ can be multiplied by a positive scalar without affecting the inequality in Eq. (6.17). □

In Eq. (6.17), utilities are first transformed using the functions $\mathcal{V}_i^{-1}$ before taking a weighted sum. A social evaluation of this form is a *transformed utilitarian* social evaluation. Such orderings may bear little resemblance to any weighted utilitarian ordering. For example, if $U(\mathbf{x}) \gg 0_n$ for all $\mathbf{x} \in S^M$ and $\mathcal{V}_i^{-1}$ is the natural logarithmic function for each $i \in N$, then prospects are being ordered in Eq. (6.17) by a Cobb–Douglas function of the individual utilities.

[19] 0_n is the vector of n zeros.

By substituting Eq. (6.8) into Eq. (6.17), we see that for all $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i EV_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i EV_i(\mathbf{y}). \tag{6.19}$$

Thus, the social evaluation is a weighted utilitarian rule (with positive weights) in terms of the expected Bernoulli utility functions. Although the expected Bernoulli utility functions can be used to predict behaviour under uncertainty, they have no welfare significance if the profile is not a Bernoulli expected utility profile. It is the profile of utility functions U that has welfare significance and in terms of these utility functions, the ordering R_U is not weighted utilitarian if U is not a Bernoulli expected utility profile.

The weights λ^i depend both on the profile U and the choice of the profile of Bernoulli utility functions V. For fixed U, if V_i is replaced by $V_i' = \alpha_i + \beta_i V_i$ where $\beta_i > 0$, then in order to preserve the equivalence in Eq. (6.19) with the other weights held fixed, λ^i / β_i must be substituted for λ^i. The same substitution is made in Eq. (6.17). In addition, the transform $\mathcal{V}_i^{-1}$ must be adjusted in Eq. (6.17) to maintain the equality in Eq. (6.7). The net effect of these two changes is the addition of an irrelevant constant to all values of the function $\lambda^i \mathcal{V}_i^{-1}$.

If the profile is a regular Bernoulli expected utility profile, Theorem 6.5 implies that the social evaluation functional must be weighted utilitarian (with positive weights).

Theorem 6.6: *Suppose that $f : \mathcal{D} \to \mathcal{R}$ is a social evaluation functional with an unrestricted expected utility range and a domain consisting of the single regular Bernoulli expected utility profile $U \in \mathcal{B}^n$. For each $i \in N$, suppose that U_i is expressed as in Eq. (6.3) using the Bernoulli utility function V_i. Then, f satisfies Strong Pareto if and only if there exists a vector $\lambda \gg 0_n$, unique up to a positive factor of proportionality, such that for all* $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i U_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i U_i(\mathbf{y}). \tag{6.20}$$

Theorems 6.5 and 6.6 make it clear that the expected utility hypothesis and Strong Pareto are not sufficient for utilitarianism. To obtain utilitarianism, the profile must satisfy the Bernoulli hypothesis, not just the expected utility hypothesis.

Because there is only the one profile U in the domain, it follows that $\mathbf{U}_{\mathcal{D}} = U(S^M)$, Profile-Dependent Welfarism is equivalent to Welfarism, and Profile-Dependent Strong Neutrality is equivalent to Strong Neutrality. By

Theorem 6.1, we know that the social evaluations in Eqs. (6.17) and (6.20) may be equivalently described by social welfare orderings on $U(S^M)$. In the case of Eq. (6.17), the corresponding social welfare ordering R_U^* is given by

$$uR_U^* v \longleftrightarrow \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(u_i) \geq \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(v_i) \tag{6.21}$$

for all $u, v \in U(S^M)$. This ordering can be represented by the social welfare function W on $U(S^M)$ defined by setting

$$W(u) = \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(u_i) \tag{6.22}$$

for all $u \in U(S^M)$. A social welfare function that ranks utility vectors as in Eq. (6.22) is a *transformed utilitarian* social welfare function. Similarly, the social welfare ordering R_U^* corresponding to Eq. (6.20) is given by

$$uR_U^* v \longleftrightarrow \sum_{i=1}^{n} \lambda^i u_i \geq \sum_{i=1}^{n} \lambda^i v_i \tag{6.23}$$

for all $u, v \in U(S^M)$ and this ordering can be represented by the social welfare function W defined by setting

$$W(u) = \sum_{i=1}^{n} \lambda^i u_i \tag{6.24}$$

for all $u \in U(S^M)$. This is a *weighted utilitarian* social welfare function. Although the social welfare function in Eq. (6.24) is weighted utilitarian, the social welfare function in Eq. (6.22) is weighted utilitarian only if the functions $\mathcal{V}_i^{-1}$ are affine, and this is the case only if U is a Bernoulli expected utility profile.

We are using the ex ante approach to social evaluation. In the ex post approach, there is an ex post social welfare function defined on the utilities obtained ex post and prospects are ordered using the ex post social welfare function. Hammond (1981, 1983) has shown that when the Bernoulli hypothesis is satisfied, the ex ante and ex post approaches to social evaluation coincide for a utilitarian. This equivalence does not hold if the expected utility hypothesis is satisfied but the Bernoulli hypothesis is not satisfied. To illustrate this point, suppose that there are two people and two states with $p_1 = p_2 = 1/2$. The outcome in state m is $x_m = (x_{m1}, x_{m2})$ where x_{mi} is person i's consumption of a single good (x_{mi} can be any nonnegative number). The utility functions are $U_i(\mathbf{x}) = (p_1 x_{1i} + p_2 x_{2i})^{1/2}$, $i = 1, 2$. Note that U_i is in $\mathcal{E}$ but not in $\mathcal{B}$. Prospect $\mathbf{y}$ is described by $y_1 = (100, 10)$ and

$y_2 = (10, 100)$, while prospect $\mathbf{z}$ is described by $z_1 = z_2 = (50, 50)$. The ex ante utilities are $U_1(\mathbf{y}) = U_2(\mathbf{y}) = 7.42$ and $U_1(\mathbf{z}) = U_2(\mathbf{z}) = 7.07$. The ex ante utilities obtained with $\mathbf{y}$ Pareto dominate the ex ante utilities obtained with $\mathbf{z}$, so an ex ante utilitarian prefers $\mathbf{y}$ to $\mathbf{z}$. Now suppose that utilitarianism is used to rank the prospects on an ex post basis by computing the expected value of total utility in each state. For $\mathbf{y}$, this number is 13.16, while for $\mathbf{z}$, it is 14.14. Hence, an ex post utilitarian prefers $\mathbf{z}$ to $\mathbf{y}$.[20]

6.7 Multiprofile Aggregation

The theorems on single-profile aggregation in the preceding section can be restated as multiprofile propositions. In the case of Theorem 6.5, the multiprofile analogue would say that if all profiles in the domain of the social evaluation functional f are regular expected utility profiles and f has an unrestricted expected utility range, then Strong Pareto is satisfied if and only if there are profile-dependent positive weights (unique up to a factor of proportionality) for which Eq. (6.17) holds for each profile in the domain. Similarly, the analogue of Theorem 6.6 would say that if all profiles in the domain of f are regular Bernoulli expected utility profiles and f has an unrestricted expected utility range, then Strong Pareto is satisfied if and only if there are profile-dependent positive weights (unique up to a factor of proportionality) for which Eq. (6.22) holds for each profile in the domain. Because there are no assumptions placing cross-profile restrictions on f, the social evaluations for different profiles can be chosen independently, and we have, in effect, a set of single-profile results.

Binary Independence of Irrelevant Alternatives is an interprofile condition that limits our ability to choose the social evaluations for different profiles independently. If both Binary Independence and Pareto indifference are assumed, we know from Theorems 6.2 and 6.3 that f also satisfies Strong Neutrality and Welfarism if the domain of f is either $\mathcal{E}^n$ (the set of all expected utility profiles) or $\mathcal{B}^n$ (the set of all Bernoulli expected utility profiles) and f has an unrestricted expected utility range. Welfarism by itself is compatible with such diverse approaches to social evaluation as utilitarianism and leximin (the lexicographic version of the maximin utility rule).[21] However, we show in this section that when combined with the assumption that each social evaluation must satisfy the expected utility hypothesis, Welfarism is extremely restrictive if the domain of f is either $\mathcal{E}^n$ or $\mathcal{B}^n$. We

[20] Roemer (2008) considers a similar example.

[21] See Bossert and Weymark (2004) for a wide range of examples of welfarist social evaluation functionals.

assume throughout this section that the set of state-contingent alternatives S is such that $\mathcal{B}^n$ contains a regular profile.

For the domain $\mathcal{E}^n$, Theorem 6.7 shows that requiring the social evaluation functional to have an unrestricted expected utility range is incompatible with satisfying both Strong Pareto and Binary Independence. Consequently, on this domain, it is impossible to satisfy the range restriction, Welfarism, and Strong Pareto.

Theorem 6.7: *Suppose that $\mathcal{B}^n$ contains a regular profile. Then, there is no social evaluation functional $f : \mathcal{D} \to \mathcal{R}$ with an unrestricted expected utility range and an unrestricted expected utility domain that satisfies both Strong Pareto and Binary Independence of Irrelevant Alternatives.*

Proof: On the contrary, suppose that f satisfies Strong Pareto and Binary Independence. By Theorems 6.2 and 6.3, f satisfies Welfarism and there is a social welfare ordering R^* on $\mathbb{R}^n$ isomorphic to f. Consider an arbitrary regular profile $\bar{U} \in \mathcal{B}^n$, the Bernoulli expected utility domain. By considering the restriction of the domain to $\bar{U}$, it follows from Theorem 6.6 that there exists a vector $\bar{\lambda} \gg 0_n$, unique up to a factor of proportionality, such that

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \bar{\lambda}^i \bar{U}_i(\mathbf{x}) \geq \sum_{i=1}^{n} \bar{\lambda}^i \bar{U}_i(\mathbf{y}) \tag{6.25}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. The corresponding social welfare ordering $R^*_{\bar{U}}$ is given by

$$u R^*_{\bar{U}} v \longleftrightarrow \sum_{i=1}^{n} \bar{\lambda}^i u_i \geq \sum_{i=1}^{n} \bar{\lambda}^i v_i \tag{6.26}$$

for all $u, v \in \bar{U}(S^M)$.

The domain $\mathcal{E}^n$ is rich enough that we can find a profile $\widehat{U}$ in $\mathcal{E}^n \setminus \mathcal{B}^n$ such that the interiors of $\bar{U}(S^M)$ and $\widehat{U}(S^M)$ have a nonempty intersection. For each $i \in N$, suppose that $\widehat{U}_i$ is written as in Eq. (6.7) using the Bernoulli utility function $\widehat{V}_i$ and the transform $\widehat{\mathcal{V}}_i$. Note that $\widehat{\mathcal{V}}_j$ must be nonaffine for some $j \in N$. By considering the restriction of the domain to $\widehat{U}$, it follows from Theorem 6.5 that there exists a vector $\hat{\lambda} \gg 0_n$, unique up to a factor of proportionality, such that

$$\mathbf{x} R_{\widehat{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \hat{\lambda}^i \widehat{\mathcal{V}}_i^{-1}(\widehat{U}_i(\mathbf{x})) \geq \sum_{i=1}^{n} \hat{\lambda}^i \widehat{\mathcal{V}}_i^{-1}(\widehat{U}_i(\mathbf{y})) \tag{6.27}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. The corresponding social welfare ordering $R^*_{\widehat{U}}$ is given by

$$uR^*_{\widehat{U}}v \longleftrightarrow \sum_{i=1}^{n} \hat{\lambda}^i \widehat{\mathcal{V}}_i^{-1}(u_i) \geq \sum_{i=1}^{n} \hat{\lambda}^i \widehat{\mathcal{V}}_i^{-1}(v_i) \tag{6.28}$$

for all $u, v \in \widehat{U}(S^M)$.

Welfarism requires the orderings $R^*_{\bar{U}}$ and $R^*_{\widehat{U}}$ to coincide on $\bar{U}(S^M) \cap \widehat{U}(S^M)$. For some $j \in N$, $\widehat{\mathcal{V}}_j^{-1}$ is nonaffine because the inverse of a nonaffine function is nonaffine. Because both $\bar{\lambda}$ and $\hat{\lambda}$ are strictly positive and because the interiors of $\bar{U}(S^M)$ and $\widehat{U}(S^M)$ have a nonempty intersection, it then follows from Eqs. (6.26) and (6.28) that $R^*_{\bar{U}}$ and $R^*_{\widehat{U}}$ do not coincide on $\bar{U}(S^M) \cap \widehat{U}(S^M)$, a contradiction. Hence, it is not possible for f to satisfy both Strong Pareto and Binary Independence. □

No information invariance assumption is used in Theorem 6.7. As a consequence, we have an impossibility theorem for any information partition of the domain. In particular, we have an impossibility result even if utilities have numerical significance.

The intuition for Theorem 6.7 is quite straightforward. By requiring the social evaluation functional to have an unrestricted expected utility range, Strong Pareto implies that the social evaluation for any Bernoulli expected utility profile is isomorphic to a weighted utilitarian social welfare ordering. With these same assumptions, the social evaluation for any non-Bernoulli expected utility profile is isomorphic to a transformed utilitarian social welfare ordering that is not weighted utilitarian. Provided that the utility vectors that are feasible for these profiles have sufficient overlap, the two social welfare orderings must differ in how they rank some pairs of utility vectors, and this is incompatible with Welfarism, and hence with Binary Independence.

This argument is illustrated in Figure 6.1 for a two-person society. The social welfare ordering $R^*_{\bar{U}}$ for the Bernoulli expected utility profile $\bar{U}$ is assumed to be utilitarian on $\bar{U}(S^M) = \mathbb{R}^2$. The dashed lines in the diagram are indifference curves of $R^*_{\bar{U}}$. For $i = 1, 2$, if $\widehat{U}_i(\mathbf{x}) = \exp(\bar{U}_i(\mathbf{x}))$ for all $\mathbf{x} \in S^M$, then $\widehat{U}(S^M) = \mathbb{R}^2_{++}$, the positive orthant of $\mathbb{R}^2$. The social welfare ordering $R^*_{\widehat{U}}$ can be represented by a social welfare function that orders utility vectors by taking a positive weighted sum of the logarithms of the individual utilities. This is a Cobb–Douglas social welfare function. Because we have not required that $R_{\bar{U}}$ equal $R_{\widehat{U}}$, the weights for the two individuals need not be equal. Some indifference curves for $R^*_{\widehat{U}}$ are shown by solid lines

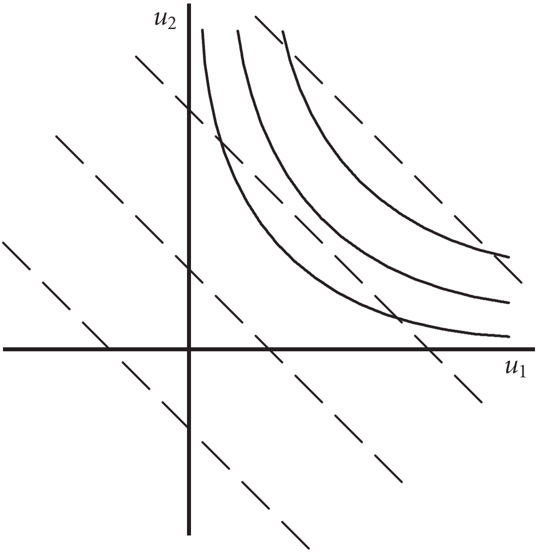


Figure 6.1. A violation of Welfarism.

in the diagram. Because the indifference curves of $R^*_{\bar{\bar{U}}}$ and $R^*_{\bar{\bar{U}}}$ are not the same on $\mathbb{R}^2_{++}$, Welfarism is violated.

Suppose now that the domain is restricted to $\mathcal{B}^n$, so that each profile is a Bernoulli expected utility profile. With this domain and an unrestricted expected utility range, a social evaluation functional satisfies Strong Pareto and Binary Independence if and only if it is weighted utilitarian with positive weights, provided that there is a Cardinal Unit-Comparable Plus Information Partition.

Theorem 6.8: *Suppose that $\mathcal{B}^n$ contains a regular profile and that $f : \mathcal{D} \to \mathcal{R}$ is a social evaluation functional with an unrestricted expected utility range and an unrestricted Bernoulli expected utility domain. Further suppose that the information partition A is a cardinal unit-comparable plus partition of $\mathcal{B}^n$. Then, f satisfies Strong Pareto, Binary Independence of Irrelevant Alternatives, and Information Invariance with Respect to A if and only if f is weighted utilitarian for a vector of weights $\lambda \gg 0_n$ that is unique up to a positive factor of proportionality.*

Proof: A weighted utilitarian social evaluation functional clearly satisfies Strong Pareto, Binary Independence of Irrelevant Alternatives, and

Information Invariance with respect to a Cardinal Unit-Comparable Plus Partition of $\mathcal{B}^n$ if the weights are all positive. Now suppose that f satisfies these three assumptions. By Theorems 6.2 and 6.3, f satisfies Welfarism and there is a social welfare ordering R^* on $\mathbb{R}^n$ isomorphic to f. Choose a regular profile $\bar{U} \in \mathcal{B}^n$ for which $0_n \in \bar{U}(S^M)$. By restricting the domain of f to $\bar{U}$, Theorem 6.6 implies that there exists a vector of weights $\lambda \gg 0_n$, unique up to a positive factor of proportionality, such that

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i \bar{U}_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i \bar{U}_i(\mathbf{y}) \tag{6.29}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. Hence, by Welfarism,

$$u R^* v \longleftrightarrow \sum_{i=1}^{n} \lambda^i u_i \geq \sum_{i=1}^{n} \lambda^i v_i \tag{6.30}$$

for all $u, v \in \bar{U}(S^M)$.

Now consider any $u, v \in \mathbb{R}^n$. If both u and v are in $\bar{U}(S^M)$, the social welfare ordering of u and v is given by Eq. (6.30). If not, we can find an $a > 0$ such that $\widehat{U}(S^M)$ contains both u and v where $\widehat{U} = a\bar{U}$. Applying the same argument to $\widehat{U}$ as we used for $\bar{U}$, it follows that Eq. (6.30) holds for all $u, v \in \widehat{U}(S^M)$, but with a possibly different set of weights $\hat{\lambda}$. However, because $\bar{U}(S^M)$ has a nonempty interior and is contained in $\widehat{U}(S^M)$, $\hat{\lambda}$ must be proportional to λ, and can be set equal to λ with no loss of generality. Welfarism then implies that f satisfies Eq. (6.1) with the weights λ; that is, f is weighted utilitarian. □

Theorem 6.6 tells us that the social evaluation assigned to any regular Bernoulli expected utility profile must be weighted utilitarian with positive weights in order to satisfy strong Pareto when the social evaluation is required to satisfy the expected utility hypothesis. If the domain of the social evaluation functional consists of only this profile, any positive set of weights will do. If Binary Independence is not assumed, different weights can be chosen for profiles in distinct information sets. However, with the domain $\mathcal{B}^n$, combining Strong Pareto with Binary Independence implies Welfarism, and Welfarism eliminates the freedom to choose different weights (once some normalization rule is adopted) for different regular profiles. The argument is illustrated in Figure 6.2 for a two-person society. If $\bar{U}$ and $\widehat{U}$ were the only two profiles in the domain, then the relative weights used to order $\bar{U}(S^M)$ can be different from the relative weights used to order $\widehat{U}(S^M)$. However, with the domain $\mathcal{B}^n$, we can always find a regular profile $\tilde{U}$ for which $\tilde{U}(S^M)$

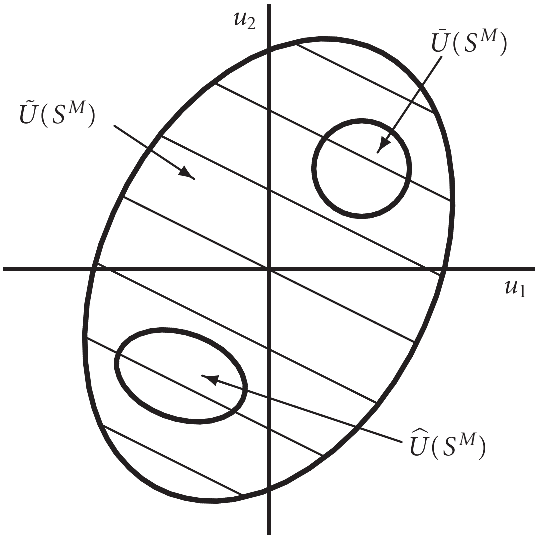


Figure 6.2. Welfare weights are profile independent.

is a superset of both $\bar{U}(S^M)$ and $\widehat{U}(S^M)$. Welfarism then requires that the same relative weights must be used for all three profiles, as shown in the diagram. Once the weights are chosen for one regular profile – and any positive weights can be chosen for this profile – the social welfare ordering isomorphic to the social evaluation functional is completely determined on all of $\mathbb{R}^n$. The impossibility of Theorem 6.7 is avoided because with no non-Bernoulli expected utility profile in the domain, we never have to use a social welfare ordering that is not weighted utilitarian to order some profile's feasible utility vectors.

An impossibility would reemerge if the information partition were too coarse. For example, suppose we have the cardinal noncomparable partition A^{CN} of $\mathcal{B}^n$. If $\bar{U}_1 = 2\widehat{U}_1$ and $\bar{U}_i = \widehat{U}_i$ for all other i, then $\bar{U}$ and $\widehat{U}$ are in the same element of A^{CN}, and so must be assigned the same social evaluation. If both of these profiles are regular Bernoulli expected utility profiles and the interiors of the feasibles sets of utility vectors for these profiles intersect (which will certainly be the case if $\bar{U}(S^M)$ contains a neighbourhood of the origin), then the social welfare orderings for these two profiles must use different weights in order for the social evaluation to be the same in both cases – the ratio of person 1's weight to any other person's weight must be twice as large with the profile $\widehat{U}$ as it is with the profile $\bar{U}$. With the cardinal unit-comparable partition of the domain (or any finer partition), these profiles are in different elements of the partition, and we are no longer constrained to assign the same social evaluation to both profiles.

An analogue of Theorem 6.8 has been established by Mongin (1994) for the set of all profiles of mixture-preserving utility functions on a convex set of a vector space when the social evaluations in the range of the social evaluation functional are required to satisfy the mixture-set version of the expected utility hypothesis.[22] This is a multiprofile version of Harsanyi's theorem because, as noted earlier, Harsanyi's lottery set is a convex set and on a convex set of lotteries, a mixture-preserving utility function is a von Neumann–Morgenstern utility function. Mongin does not *assume* that there is a cardinal unit-comparable partition of his domain. Instead, he notes that because his axioms imply that the social evaluation functional is weighted utilitarian, it must satisfy Information Invariance with Respect to Cardinal Unit-Comparable utilities in order for his axioms to be consistent. This information invariance condition is compatible with any cardinal unit-comparable plus partition.

The characterization of weighted utilitarian social evaluation functionals in Theorem 6.8 holds for *any* information partition that is a refinement of the cardinal unit-comparable partition. This characterization makes essential use of our assumptions that we have an unrestricted Bernoulli expected utility domain and an unrestricted expected utility range. Multiprofile characterizations of utilitarian and weighted utilitarian social evaluation functionals have also been obtained by d'Aspremont and Gevers (1977) and Roberts (1980) for the cardinal unit-comparable information partition of the set of all possible profiles. These characterizations do not suppose that there is any uncertainty about the outcome that results from the choice of an alternative and therefore cannot require the utility functions or the social evaluations to satisfy the expected utility hypothesis. Because the range is less restricted than is the case here, there is no incompatibility between Strong Pareto and Binary Independence when the domain is unrestricted. However, these characterization theorems are quite sensitive to the assumption made about the information partition. For sufficiently fine information partitions, further social evaluation functionals are possible.[23]

[22] Coulhon and Mongin (1989) had previously established a related result but for a social welfare functional whose range is a set of mixture-preserving *functions*.

[23] See Bossert and Weymark (2004) for details. See also the related theorems in Deschamps and Gevers (1977) and Maskin (1978). Deschamps and Gevers consider social choice under uncertainty interpretations of their theorem.

6.8 Single-Information-Set Aggregation

In constructing an information partition, it is natural to suppose that different preference profiles can be distinguished. Henceforth, we assume that any two utility profiles in the same element of an information partition must represent the same preference profile. When designing the social aggregation rule, if the actual information set is known, then the domain of the social evaluation functional is a single information set for some information partition that is no coarser than the ordinal noncomparable partition.

The theorems in Section 6.6 are single-information-set aggregation theorems for numerically comparable utilities. Because there is only one utility profile in the domain, the social evaluation functionals identified in these theorems are strongly neutral. For other domains and other information partitions, Pareto Indifference and Binary Independence do not, in general, imply either Strong Neutrality or Welfarism. However, if there is sufficient preference diversity (for example, if the assumptions of Theorem 6.4 are satisfied), then Strong Neutrality implies Welfarism. A single-information-set version of the multiprofile weighted utilitarian theorem (Theorem 6.8) can be obtained by substituting Strong Neutrality for Binary Independence and by supposing that the domain is a single information set in a cardinal unit-comparable plus partition of $\mathcal{B}^n$ that contains a profile U for which $U(S^M) = \mathbf{U}_{\mathcal{D}}$.[24] To be in this domain, a profile $\bar{U}$ must not only be a Bernoulli expected utility profile, in addition, for each $i \in N$, $\bar{U}_i$ must be an increasing affine transform of U_i with the unit-scaling parameters common to all of the transforms. With the cardinal noncomparable (or any coarser) partition of $\mathcal{B}^n$, it is a simple matter to construct a single-information-set impossibility for a strongly neutral social evaluation functional similar to the impossibility result for the domain $\mathcal{B}^n$ described informally at the end of the preceding section.

These observations suggest that if the information partition is too coarse, then it is not possible to satisfy both Strong Pareto and Strong Neutrality when there is an unrestricted expected utility range and the domain is a single information set, provided the domain satisfies some regularity condition. In this section, we consider single-information-set aggregation and investigate the nature of the restrictions that must be imposed on the information set if the social evaluation functional is to have an unrestricted expected

[24] The proof of this result is the same as the proof of Theorem 6.8 except that Theorem 6.4 is used to show that f satisfies Welfarism.

utility range and satisfy both Strong Pareto and Strong Neutrality.[25] As in the preceding sections, we also characterize the social evaluation functionals that satisfy these assumptions given our domain restrictions.

Consider any regular profile $U \in \mathcal{E}^n$. We let $\langle U \rangle$ denote the information set containing U. As in Eq. (6.8), for each $i \in N$, there is an increasing function $\mathcal{V}_i : \mathbb{R} \to \mathbb{R}$ and a Bernoulli utility function $V_i : S \to \mathbb{R}$ such that

$$U_i(\mathbf{x}) = \mathcal{V}_i(EV_i(\mathbf{x})) \tag{6.31}$$

for all $\mathbf{x} \in S^M$. In order to simplify the statement of our theorems, we suppose that $U(S^M) = EV(S^M) = \mathbb{R}^n$.[26] We discuss the implications of relaxing this assumption informally later in this section. Because all of the profiles in $\langle U \rangle$ represent the same preference profile, any $\bar{U} \in \langle U \rangle$ is also in $\mathcal{E}^n$. Further, for any $\bar{U} \in \langle U \rangle$, for each $i \in N$, there exists an increasing function $g_i : \mathbb{R} \to \mathbb{R}$ such that for all $\mathbf{x} \in S^M$,

$$\bar{U}_i(\mathbf{x}) = g_i(U_i(\mathbf{x})). \tag{6.32}$$

The information set is characterized by specifying the n-tuples $(g_1, \ldots, g_n)$ that are admissible. Substituting Eq. (6.31) into Eq. (6.32), we obtain

$$\bar{U}_i(\mathbf{x}) = \bar{\mathcal{V}}_i(EV_i(\mathbf{x})) \tag{6.33}$$

for all $\mathbf{x} \in S^M$ and all $i \in N$, where

$$\bar{\mathcal{V}}_i = g_i \circ \mathcal{V}_i. \tag{6.34}$$

Thus, every profile in the information set can be written using a *single* Bernoulli utility function for each individual.

The nature of our results depends on whether the reference profile U is a Bernoulli expected utility profile or not. In Theorem 6.9, we suppose that U is a regular profile with $U(S^M) = EV(S^M) = \mathbb{R}^n$ and consider the general case in which U can be either a Bernoulli or a non-Bernoulli expected utility profile.

Theorem 6.9: *Suppose that $f : \mathcal{D} \to \mathcal{R}$ is a social evaluation functional with an unrestricted expected utility range. Suppose that A is an information partition of $\mathcal{E}^n$ and that the domain of f is the single information set $\langle U \rangle$ where*

[25] Samuelson (1977) has been quite critical of the use of Strong Neutrality as an axiom in single-preference-profile social choice. His argument, however, makes essential use of the assumption that no interpersonal comparisons of utility are possible. See Blackorby, Donaldson, and Weymark (1990) for a discussion of Samuelson's argument.

[26] Because U is continuous and S is bounded, we are therefore implicitly assuming that the set of alternatives S^M is not closed.

(a) U is a regular profile in $\mathcal{E}^n$, (b) $U(S^M) = \mathbb{R}^n$, (c) for each $i \in N$, U_i is expressed as in Eq. (6.31) using the Bernoulli utility function V_i and the transform $\mathcal{V}_i$, and (d) $EV(S^M) = \mathbb{R}^n$. For each $\bar{U} \in \langle U \rangle$ and each $i \in N$, suppose that $\bar{U}_i$ is expressed as in Eq. (6.33) using the Bernoulli utility function V_i and the transform $\bar{\mathcal{V}}_i$. Then, f satisfies Strong Pareto, Strong Neutrality, and Information Invariance with Respect to A if and only if (i) there exists a vector $\lambda \gg 0_n$, unique up to a positive factor of proportionality, such that for all $\bar{U} \in \langle U \rangle$ and all $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i \bar{\mathcal{V}}_i^{-1}(\bar{U}_i(\mathbf{x})) \geq \sum_{i=1}^{n} \lambda^i \bar{\mathcal{V}}_i^{-1}(\bar{U}_i(\mathbf{y})) \tag{6.35}$$

and (ii) for each $\bar{U} \in \langle U \rangle$, there exist scalars $\gamma, \varepsilon_1, \ldots, \varepsilon_n$ with $\gamma > 0$ such that Eq. (6.32) holds for each $i \in N$ with

$$g_i(\tau) = \mathcal{V}_i(\gamma \mathcal{V}_i^{-1}(\tau) + \varepsilon_i) \tag{6.36}$$

for all $\tau \in \mathbb{R}$.

Proof: Suppose that Eqs. (6.35) and (6.36) are satisfied. Substituting Eq. (6.33) into Eq. (6.35), we obtain

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i EV_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i EV_i(\mathbf{y}), \tag{6.37}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$ and all $\bar{U} \in \langle U \rangle$. Hence, $R_{\bar{U}} = R_{\widehat{U}}$ for all $\bar{U}, \widehat{U} \in \langle U \rangle$, and so f satisfies the information invariance assumption. Clearly, if Eq. (6.35) is satisfied with $\lambda \gg 0_n$, then f satisfies Strong Pareto.

Substituting $U_i(\mathbf{x})$ for τ in Eq. (6.36) and using Eqs. (6.31) and (6.32), we obtain

$$\bar{U}_i(\mathbf{x}) = \mathcal{V}_i(\gamma\, EV_i(\mathbf{x}) + \varepsilon_i) \tag{6.38}$$

for all $\mathbf{x} \in S^M$ and all $i \in N$. Solving Eq. (6.38) for $EV_i(\mathbf{x})$, we find that

$$EV_i(\mathbf{x}) = [\mathcal{V}_i^{-1}(\bar{U}_i(\mathbf{x})) - \varepsilon_i]/\gamma \tag{6.39}$$

for all $\mathbf{x} \in S^M$ and all $i \in N$. Next substitute Eq. (6.39) into Eq. (6.37). After cancelling the profile-dependent parameters $\gamma, \varepsilon_1, \ldots, \varepsilon_n$, we conclude that for all $\bar{U} \in \langle U \rangle$ and all $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(\bar{U}_i(\mathbf{x})) \geq \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(\bar{U}_i(\mathbf{y})). \tag{6.40}$$

It immediately follows from Eq. (6.40) that Strong Neutrality is satisfied.

Now suppose that f satisfies Strong Pareto, Strong Neutrality, and Information Invariance with Respect to A. Consider any $\bar{U} \in \langle U \rangle$. Applying Theorem 6.5 to U and $\bar{U}$ separately, Strong Pareto implies that there exist weights $\lambda_U \gg 0_n$ and $\lambda_{\bar{U}} \gg 0_n$, each unique up to a positive factor of proportionality, such that for all $\mathbf{x}, \mathbf{y} \in S^M$,

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda_U^i \mathcal{V}_i^{-1}(U_i(\mathbf{x})) \geq \sum_{i=1}^{n} \lambda_U^i \mathcal{V}_i^{-1}(U_i(\mathbf{y})) \tag{6.41}$$

and

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda_{\bar{U}}^i \bar{\mathcal{V}}_i^{-1}(\bar{U}_i(\mathbf{x})) \geq \sum_{i=1}^{n} \lambda_{\bar{U}}^i \bar{\mathcal{V}}_i^{-1}(\bar{U}_i(\mathbf{y})). \tag{6.42}$$

Using Eqs. (6.31) and (6.33) to eliminate U_i and $\bar{U}_i$ in Eqs. (6.41) and (6.42), respectively, we obtain

$$\mathbf{x} R_U \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda_U^i EV_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda_U^i EV_i(\mathbf{y}) \tag{6.43}$$

and

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda_{\bar{U}}^i EV_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda_{\bar{U}}^i EV_i(\mathbf{y}) \tag{6.44}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. The information invariance assumption implies that $R_U = R_{\bar{U}}$. Because $EV(S^M) = \mathbb{R}^n$, Eqs. (6.43) and (6.44) then imply that λ_U is proportional to $\lambda_{\bar{U}}$. By letting $\lambda^i = \lambda_{\bar{U}}^i$, Eq. (6.35) then follows from Eq. (6.42).

By Theorem 6.4, f satisfies Welfarism. Hence, f is isomorphic to a social welfare ordering R^* on $\mathbf{U}_{\mathcal{D}} = \mathbb{R}^n$. By Eq. (6.35), R^* can be represented by the social welfare function $W_U \colon \mathbb{R}^n \to \mathbb{R}$ given by

$$W_U(u) = \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(u_i) \tag{6.45}$$

for all $u \in \mathbb{R}^n$ and R^* can be represented on $\bar{U}(S^M)$ by the social welfare function $W_{\bar{U}} \colon \bar{U}(S^M) \to \mathbb{R}$ given by

$$W_{\bar{U}}(u) = \sum_{i=1}^{n} \lambda^i \bar{\mathcal{V}}_i^{-1}(u_i) \tag{6.46}$$

for all $u \in \bar{U}(S^M)$. On $\bar{U}(S^M)$, W_U and $W_{\bar{U}}$ must be ordinally equivalent. Thus, there exists an increasing function $\Psi\colon W_{\bar{U}}(\bar{U}(S^M)) \to \mathbb{R}$ such that

$$\Psi\left[\sum_{i=1}^{n} \lambda^i \bar{\mathcal{V}}_i^{-1}(u_i)\right] = \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}(u_i) \tag{6.47}$$

for all $u \in \bar{U}(S^M)$. For all $i \in N$ and all $u_i \in \bar{U}_i(S^M)$, let

$$\xi_i = \lambda^i \bar{\mathcal{V}}_i^{-1}(u_i). \tag{6.48}$$

Solving Eq. (6.48) for u_i, we find that

$$u_i = \bar{\mathcal{V}}_i(\xi_i/\lambda^i). \tag{6.49}$$

Using Eqs. (6.48) and (6.49) in Eq. (6.47), we obtain

$$\Psi\left[\sum_{i=1}^{n} \xi_i\right] = \sum_{i=1}^{n} \lambda^i \mathcal{V}_i^{-1}[\bar{\mathcal{V}}_i(\xi_i/\lambda^i)] \tag{6.50}$$

for all $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$.[27] For each $i \in N$, let $\phi_i\colon \mathbb{R} \to \mathbb{R}$ be defined by setting

$$\phi_i(\xi_i) = \lambda^i \mathcal{V}_i^{-1}[\bar{\mathcal{V}}_i(\xi_i/\lambda^i)] \tag{6.51}$$

for all $\xi_i \in \mathbb{R}$. Because $\lambda^i > 0$ and both $\mathcal{V}_i^{-1}$ and $\bar{\mathcal{V}}_i$ are increasing, ϕ_i is an increasing function. Using Eq. (6.51), Eq. (6.50) can be rewritten as

$$\Psi\left[\sum_{i=1}^{n} \xi_i\right] = \sum_{i=1}^{n} \phi_i(\xi_i) \tag{6.52}$$

for all $\xi \in \mathbb{R}^n$.

Equation (6.52) is a Pexider equation. By Theorem 3.1.1 and Corollary 3.1.9 in Eichhorn (1978), there exist scalars $\gamma, \bar{\varepsilon}_1, \ldots, \bar{\varepsilon}_n$ such that for each $i \in N$,

$$\phi_i(\xi_i) = \gamma \xi_i + \bar{\varepsilon}_i \tag{6.53}$$

for all $\xi_i \in \mathbb{R}$. Because each of the functions ϕ_i is increasing, γ must be positive. Equating Eqs. (6.51) and (6.53), for each $i \in N$,

$$\lambda^i \mathcal{V}_i^{-1}[\bar{\mathcal{V}}_i(\xi_i/\lambda^i)] = \gamma \xi_i + \bar{\varepsilon}_i \tag{6.54}$$

[27] Because $EV(S^M) = \mathbb{R}^n$, for any $\xi \in \mathbb{R}^n$, there exists an $\bar{\mathbf{x}} \in S^M$ such that $(\xi_1/\lambda^1, \ldots, \xi_n/\lambda^n) = EV(\bar{\mathbf{x}})$. From Eqs. (6.33) and (6.48), it then follows that any $\xi \in \mathbb{R}^n$ is attainable with the profile $\bar{U}$.

for all $\xi_i \in \mathbb{R}$. Substituting Eq. (6.48) into Eq. (6.54), for each $i \in N$,

$$\lambda^i \mathcal{V}_i^{-1}(u_i) = \gamma \lambda^i \bar{\mathcal{V}}_i^{-1}(u_i) + \bar{\varepsilon}_i \tag{6.55}$$

for all $u_i \in \bar{U}_i(S^M)$. From Eqs. (6.31), (6.32), and (6.33), it follows that for all $i \in N$,

$$\bar{U}_i(\mathbf{x}) = g_i(U_i(\mathbf{x})) = \bar{\mathcal{V}}_i[\mathcal{V}_i^{-1}(U_i(\mathbf{x}))] \tag{6.56}$$

for all $\mathbf{x} \in S^M$. Letting $u_i = \bar{U}_i(\mathbf{x})$ in Eq. (6.55) and using Eq. (6.56), for all $i \in N$ and all $\mathbf{x} \in S^M$,

$$\mathcal{V}_i^{-1}[g_i(U_i(\mathbf{x}))] = \gamma \mathcal{V}_i^{-1}(U_i(\mathbf{x})) + \varepsilon_i, \tag{6.57}$$

where $\varepsilon_i = \bar{\varepsilon}_i / \lambda^i$ for each $i \in N$. Because $U(S^M) = \mathbb{R}^n$, Eq. (6.36) then follows from Eq. (6.57). □

Before discussing this result, it is useful to consider the special case in which the reference profile U is any regular Bernoulli expected utility profile for which $U(S^M) = \mathbb{R}^n$, as the statement of the theorem can then be greatly simplified. Although the reference profile is assumed to be in $\mathcal{B}^n$, we do not assume a priori that the other profiles in the information set are all in $\mathcal{B}^n$ as well. Because $U \in \mathcal{B}^n$, for each $i \in N$, there is a Bernoulli utility function $V_i \colon S \to \mathbb{R}$ such that

$$U_i(\mathbf{x}) = EV_i(\mathbf{x}) \tag{6.58}$$

for all $\mathbf{x} \in S^M$. From Eq. (6.58), it trivially follows that $EV(S^M) = \mathbb{R}^n$ when $U(S^M) = \mathbb{R}^n$.

Theorem 6.10: *Suppose that $f \colon \mathcal{D} \to \mathcal{R}$ is a social evaluation functional with an unrestricted expected utility range. Suppose that A is an information partition of $\mathcal{E}^n$ and that the domain of f is the single information set $\langle U \rangle$, where (a) U is a regular profile in $\mathcal{B}^n$, (b) $U(S^M) = \mathbb{R}^n$, and (c) for each $i \in N$, U_i is expressed as in Eq. (6.58) using the Bernoulli utility function V_i. For each $\bar{U} \in \langle U \rangle$ and each $i \in N$, suppose that $\bar{U}_i$ is expressed as in Eq. (6.33) using the Bernoulli utility function V_i and the transform $\bar{\mathcal{V}}_i$. Then, f satisfies Strong Pareto, Strong Neutrality, and Information Invariance with Respect to A if and only if (i) f is weighted utilitarian for a vector of weights $\lambda \gg 0_n$ that is unique up to a positive factor of proportionality and (ii) for each $\bar{U} \in \langle U \rangle$, there exist scalars $\gamma, \varepsilon_1, \ldots, \varepsilon_n$ with $\gamma > 0$ such that Eq. (6.32) holds for each $i \in N$ with*

$$g_i(\tau) = \gamma \tau + \varepsilon_i \tag{6.59}$$

for all $\tau \in \mathbb{R}$.

Proof: By letting $\mathcal{V}_i$ be the identity function in Eq. (6.36), we obtain Eq. (6.59). It then follows from Eq. (6.34) that for any $\bar{U} \in \langle U \rangle$, $\bar{\mathcal{V}}_i$ is an increasing affine function for all $i \in N$. Hence, for each $\bar{U} \in \langle U \rangle$, there exist profile-dependent scalars $\alpha, \beta_1, \ldots, \beta_n$ with $\alpha > 0$ such that for all $i \in N$,

$$\bar{\mathcal{V}}_i^{-1}(u_i) = \alpha u_i + \beta_i \tag{6.60}$$

for all $u_i \in \bar{U}_i(S^M)$. Substituting this expression for $\bar{\mathcal{V}}_i^{-1}$ in Eq. (6.35), on cancelling the parameters that appear in Eq. (6.60) we obtain

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i \bar{U}_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i \bar{U}_i(\mathbf{y}) \tag{6.61}$$

for all $\bar{U} \in \langle U \rangle$ and all $\mathbf{x}, \mathbf{y} \in S^M$, and so f is weighted utilitarian. □

Theorem 6.10 provides a single-information-set characterization of weighted utilitarianism. It follows from Eq. (6.59) that for any $\bar{U}, \widehat{U} \in \langle U \rangle$, there exist scalars $\alpha, \beta_1, \ldots, \beta_n$ with $\alpha > 0$ such that

$$\widehat{U}_i(\mathbf{x}) = \alpha \bar{U}_i(\mathbf{x}) + \beta_i \tag{6.62}$$

for all $\mathbf{x} \in S^M$ and all $i \in N$. Hence, the domain of the social evaluation functional must be an information set in a cardinal unit-comparable plus partition of $\mathcal{E}^n$. Because the reference profile U is a Bernoulli expected utility profile, this implies that *all* profiles in the domain must be Bernoulli expected utility profiles as well and that all of these profiles must embody the same intrapersonal and interpersonal comparisons of utility gains and losses.

The intuition for this result is quite straightforward. By Theorem 6.6, we know that the social evaluation for the reference profile U must be isomorphic to a weighted utilitarian social welfare ordering on all of $\mathbb{R}^n$. Because the social evaluation functional is welfarist, it must therefore be weighted utilitarian. Theorem 6.5 then implies that each profile $\bar{U}$ in the domain is a Bernoulli expected utility profile. Hence, because U and $\bar{U}$ represent the same preference profile, for each $i \in N$, $\bar{U}_i$ is an increasing affine transform of U_i. Further, each of these affine transforms must use the same unit-scaling parameter, otherwise it would not be possible to compare utility differences interpersonally as required by weighted utilitarianism.

Theorem 6.9 generalizes Theorem 6.10 by permitting the reference profile to be non-Bernoulli. For a social evaluation functional to satisfy the assumptions of Theorem 6.9, it must be transformed utilitarian. The domain restriction Eq. (6.36) implies that for any profile $\bar{U}$ in the domain, $\bar{U}_i$ is an expected Bernoulli utility function *if and only if* the reference utility function U_i is also an expected Bernoulli utility function. If even a single individual's reference utility function does not satisfy the Bernoulli hypothesis, then the social evaluation functional is not weighted utilitarian and the information set need not be an element of a cardinal unit-comparable plus partition of $\mathcal{E}^n$.[28] In this case, because the social evaluation functional is not weighted utilitarian, there may be no need for utility differences to be interpersonally comparable.

The restriction in Eq. (6.36) can be interpreted using Eq. (6.38). By Eq. (6.31), for each $i \in N$, i's reference utility function U_i can be expressed using the expected Bernoulli utility function EV_i and the transform $\mathcal{V}_i$. Consider any information set in a cardinal unit-comparable plus partition of $\mathcal{E}^n$ containing the profile EV. If we then map *each* profile in this set into a new profile by applying the transform $\mathcal{V}_i$ to i's utility function, Eq. (6.38) tells us that we obtain an information set satisfying Eq. (6.36) in another partition of $\mathcal{E}^n$. Further, only sets that can be constructed in this way satisfy our domain restriction. In effect, we have a bijection between a weighted utilitarian rule on an information set satisfying Eq. (6.59) and a transformed utilitarian rule on an information set satisfying Eq. (6.36).

Further intuition for the domain restriction can be obtained from Eqs. (6.45) and (6.46). As we have seen in the single-profile case, for each profile in our domain, the social evaluation is transformed utilitarian. Because all of the profiles in the domain represent the same preference profile, they can all be written as transforms of a common Bernoulli expected utility profile. In terms of this Bernoulli profile, each of the social evaluations is weighted utilitarian, but, in principle, the weights can depend on the profile from the information set being considered. However, because all profiles in the domain are assigned the same social evaluation, in fact, the same (relative) weights must be used for all profiles. These are the weights λ^i used in Eqs. (6.45) and (6.46). Welfarism requires the social welfare functions in Eqs. (6.45) and (6.46) to be ordinally equivalent on the intersection of their domains. Clearly, this is the case if there exist scalars $\alpha, \beta_1, \ldots, \beta_n$ with $\alpha > 0$ such that for all u in the domain of both functions and all $i \in N$,

$$\mathcal{V}_i^{-1}(u_i) = \alpha \bar{\mathcal{V}}_i^{-1}(u_i) + \beta_i. \tag{6.63}$$

Because the common part of the domain has a nonempty interior, Eq. (6.63) is also necessary for Eqs. (6.45) and (6.46) to be ordinally equivalent. The restriction in Eq. (6.36) follows from this observation after some simple algebra.

[28] Of course, if the information set only contains a single profile, it trivially is an information set for a cardinal unit-comparable plus partition of $\mathcal{E}^n$.

In Theorems 6.9 and 6.10, we have assumed that there is a profile U in the domain for which $U(S^M) = EV(S^M) = \mathbb{R}^n$. The most important implication of this assumption is that for any other profile $\bar{U}$ in the domain, $\bar{U}(S^M) \subseteq U(S^M)$. If this is not the case, our theorems require some qualification. For example, suppose that U and $\bar{U}$ are both regular profiles in the information set and $\Delta := U(S^M) \cap \bar{U}(S^M) \neq \varnothing$, but neither of these two sets of feasible utility vectors is contained in the other. As in the proof of Theorem 6.9, the weights in Eq. (6.43) are proportional to the weights in Eq. (6.44).[29] Strong Neutrality implies that the social welfare orderings R^*_U and $R^*_{\bar{U}}$ must coincide on Δ. In general, the Pexider equation in Eq. (6.52) only holds on a restricted domain, not all of $\mathbb{R}^n$. Provided that the domain of the Pexider equation is contained in the closure of its interior and the interior is a connected set, the functions ϕ_i must have the functional forms given in Eq. (6.53).[30] To establish this result, we must appeal to Corollary 3 in Radó and Baker (1987) to solve the Pexider equation, instead of the theorems in Eichhorn (1978) used in the proof of Theorem 6.9.[31] When U and $\bar{U}$ are expressed as in Eqs. (6.31) and (6.32), the transforms $\mathcal{V}_i$ and g_i must satisfy Eq. (6.57) for all $\mathbf{x} \in S^M$ for which $\bar{U}(\mathbf{x}) \in \Delta$. In Theorems 6.9 and 6.10, all $\mathbf{x} \in S^M$ have this property.

Care must be taken in interpreting our theorems when $\bar{U}(\mathbf{x}) \notin \Delta$. For example, if U and $\bar{U}$ are both in $\mathcal{B}^n$, then the same relative weights must be used to rank utility vectors in $U(S^M)$ as are used to rank utility vectors in $\bar{U}(S^M)$, as illustrated in Figure 6.3. However, we cannot conclude that the same absolute weights are used in both cases. As a consequence, we cannot conclude that the u and v shown in Figure 6.3 are socially indifferent, as would be the case if the social evaluation functional were weighted utilitarian.[32]

6.9 Single-Preference-Profile Aggregation

Harsanyi aggregates a single-preference profile into a social ordering of the alternatives. Single-preference-profile aggregation is appropriate when the actual preference profile is known at the time the aggregation rule is proposed and no utility information is available other than what is contained

[29] This argument does not require that either $U(S^M) = EV(S^M)$ or $EV(S^M) = \mathbb{R}^n$.

[30] Because both U and $\bar{U}$ are regular, these are mild restrictions on the domain of the Pexider equation.

[31] The Pexider equation is first solved on the interior of its domain and then continuity is used to extend the solution to the whole domain. This two-step procedure is used to solve the Pexider equation in Blackorby, Donaldson, and Weymark (1999).

[32] See Wakker (1993) for a more detailed discussion of this problem in a related example.

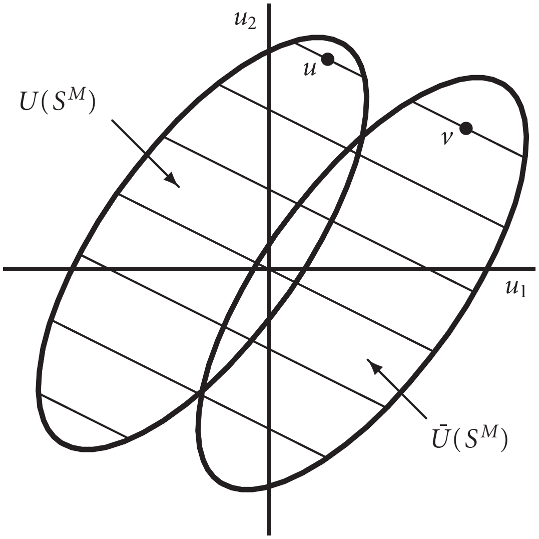


Figure 6.3. Different absolute welfare weights may be used in $U(S^M)$ and $\bar{U}(S^M)$.

in the individual preferences. Single-preference-profile aggregation can be modeled in our framework by supposing that the social welfare functional satisfies Information Invariance with Respect to Ordinal Noncomparable Utilities and the domain is a single information set in the ordinal noncomparable partition of $\mathcal{U}^n$. Provided that this information set contains a regular expected utility profile, no strongly neutral social evaluation functional with an unrestricted expected utility range satisfes Strong Pareto and Information Invariance with Respect to Ordinally Noncomparable Utilities on this domain. This impossibility result can be established using the same argument as is used to prove Theorem 6.7 because the two profiles used in that proof can be chosen to represent the same preference profile. Harsanyi does not assume Strong Neutrality and so avoids this impossibility. In this section, we consider nonneutral single-preference-profile aggregation, as in Harsanyi (1955, 1977), but for state-contingent alternatives.

Suppose that our single preference profile is **R** and that each of the individual preferences in **R** satisfies the expected utility hypothesis. Consider any Bernoulli expected utility profile U that represents **R**, and let $[U]$ denote the element of the ordinal noncomparable partition of $\mathcal{U}^n$ containing U. Thus, $[U]$ is the set of all utility profiles that represent **R** and a profile $\bar{U}$ is in $[U]$ if and only if, for each $i \in N$, $\bar{U}_i$ is an increasing transform of U_i. The set $[U]$ is the domain of the social evaluation functional when the single preference profile **R** is being aggregated. Each of the utility profiles in $[U]$ is in $\mathcal{E}^n$. We suppose that U and, hence, every other profile in $[U]$ is regular. As in Eq. (6.33), all of these profiles can be expressed in terms of a single Bernoulli utility function for each individual.

The analogue to Theorems 6.5 and 6.6 for single-preference-profile aggregation is provided by Theorem 6.11.

Theorem 6.11: *Suppose that $f : \mathcal{D} \to \mathcal{R}$ is a social evaluation functional with an unrestricted expected utility range and the domain $[U]$ for some regular Bernoulli expected utility profile $U \in \mathcal{B}^n$. For each $i \in N$, suppose that U_i is expressed as in Eq. (6.3) using the Bernoulli utility function V_i. For each $\bar{U} \in [U]$ and each $i \in N$, suppose that $\bar{U}_i$ is expressed as in Eq. (6.7) using the Bernoulli utility function V_i and the transform $\bar{\mathcal{V}}_i$. Then, f satisfies Strong Pareto and Information Invariance with Respect to Ordinal Noncomparable Utilities if and only if there exists a vector $\lambda \gg 0_n$, unique up to a positive factor of proportionality, such that Eq. (6.35) holds for all $\bar{U} \in [U]$ and all* $\mathbf{x}, \mathbf{y} \in S^M$.

Proof: The proof that Strong Pareto and the information invariance assumption follow from Eq. (6.35) is the same as the first part of the proof of Theorem 6.9. To establish the reverse implication, we first note that by applying Theorem 6.5 to each profile in $[U]$ using U as the reference Bernoulli expected utility profile in each case, Strong Pareto implies that for any $\bar{U}, \widehat{U} \in [U]$, there exist weights $\lambda_{\bar{U}} \gg 0_n$ and $\lambda_{\widehat{U}} \gg 0_n$ (each unique up to a positive factor of proportionality) such that

$$\mathbf{x} R_{\bar{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i_{\bar{U}} U_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i_{\bar{U}} U_i(\mathbf{y}) \tag{6.64}$$

and

$$\mathbf{x} R_{\widehat{U}} \mathbf{y} \longleftrightarrow \sum_{i=1}^{n} \lambda^i_{\widehat{U}} U_i(\mathbf{x}) \geq \sum_{i=1}^{n} \lambda^i_{\widehat{U}} U_i(\mathbf{y}) \tag{6.65}$$

for all $\mathbf{x}, \mathbf{y} \in S^M$. By information invariance with respect to ordinally measurable utilities, $R_{\bar{U}} = R_{\widehat{U}}$. Because $U(S^M)$ has a nonempty interior, it follows that $\lambda_{\widehat{U}}$ must be proportional to $\lambda_{\bar{U}}$, and so without loss of generality can be set equal to $\lambda_{\bar{U}}$. □

When Eq. (6.35) is satisfied, for each $\bar{U} \in [U]$, the social welfare function $W_{\bar{U}}$ on $\bar{U}(S^M)$ corresponding to $R_{\bar{U}}$ is given by

$$W_{\bar{U}}(u) = \sum_{i=1}^{n} \lambda^i \bar{\mathcal{V}}_i^{-1}(u_i) \tag{6.66}$$

for all $u \in \bar{U}(S^M)$. These are transformed utilitarian social welfare functions. Because all profiles are expressed as transforms of the reference profile U, the same weights λ are used in Eq. (6.66) for each profile in $[U]$. This restriction is needed to ensure that the same social evaluation is assigned to every profile in the domain. Because $\bar{\mathcal{V}}_i^{-1}$ is profile dependent, the social evaluation functional f satisfies Profile-Dependent Welfarism and Profile-Dependent Strong Neutrality, but it does not satisfy either Welfarism or Strong Neutrality.

If $\bar{U}$ is in $\mathcal{B}^n$, $\bar{\mathcal{V}}_i^{-1}$ is an increasing affine transform and $W_{\bar{U}}$ is ordinally equivalent to a weighted utilitarian social welfare function. As in Section 6.6, the ith weight is $\lambda^i/\bar{\beta}_i$, where $\bar{\beta}_i$ is the parameter that is used to scale units in $\bar{\mathcal{V}}_i$. If $\bar{U}$ is not in $\mathcal{B}^n$, then $W_{\bar{U}}$ is a transformed utilitarian social welfare function that uses a nonaffine function to transform at least one person's utility before aggregating. Because the social evaluation functional f is not welfarist, little can be inferred from the functional forms of these profile-dependent social welfare functions. In particular, it is illegitimate to argue that f is weighted utilitarian because by choosing the utility profile $\bar{U}$ that represents **R** to be a Bernoulli expected utility profile, $R^*_{\bar{U}}$ can be represented by a weighted utilitarian social welfare function. For f to be weighted utilitarian, there must be a single set of weights $\lambda^1, \ldots, \lambda^n$ such that Eq. (6.1) holds for all utility profiles in the domain, and this is not the case here.[33] In order to talk meaningfully about weighted utilitarianism, it must be possible to make interpersonal comparisons of utility gains and losses. Because we only have the preference profile **R** at our disposal, no interpersonal utility comparisons are feasible. Weighted utilitarianism is meaningful in the single-utility-profile case considered in Section 6.6 because utilities must have numerical significance when there is only one utility profile in the domain of the social evaluation function.

[33] Sen (1976) uses their observation in his argument that Harsanyi's single-preference-profile social aggregation theorem does not provide an argument in support of utilitarianism. For a detailed discussion of this issue, see Weymark (1991). See also Blackorby, Donaldson, and Weymark (1980) and Roemer (1996, 2008).

6.10 Concluding Remarks

Strong Pareto ensures that every individual's utility counts positively. However, in the original statement of Harsanyi's theorem, the social aggregation procedure is merely required to satisfy Pareto Indifference. If Pareto Indifference is assumed instead of Strong Pareto in our single-profile, single-preference-profile, or Bernoulli expected utility multiprofile theorems, the only consequence is that the individual weights would no longer be restricted in sign. The proof of the multiprofile impossibility theorem for the domain $\mathcal{E}^n$ requires the social welfare function isomorphic to the social evaluation functional to be nonconstant in the utilities of at least two individuals. With Pareto Indifference, this is not guaranteed and so dictatorial, antidictatorial, and null social evaluation functionals are then possible.[34] Note that if there is a dictator or an antidictator, a transformed utilitarian social evaluation functional is also weighted utilitarian because transforming this individual's utility has no effect on any social evaluation. In the single-information-set case, dictatorial, antidictatorial, and null social evaluation functionals are consistent with Pareto Indifference, and these social aggregation procedures place no restrictions on the information set (other than our maintained assumption that all profiles represent the same preference profile). In addition, it is possible for at least two of the weights in λ to be nonzero, but not all. In this case, the restrictions on the information set only apply to the individuals who receive a nonzero weight.

At many points in our argument, we have made use of the fact that the set of feasible utilities for some profile is of full dimension, a property of any regular profile. This assumption corresponds to the assumption needed in Harsanyi's lottery model of uncertainty to ensure that the social utility function can be *uniquely* expressed as an affine combination of the individual utility functions.[35] For our single-profile aggregation theorems, if "interior" is replaced by "relative interior" in the definition of a regular expected utility profile, the only consequence of this change is that the weights λ in Theorems 6.5 and 6.6 need not be unique up to a factor of proportionality.[36] Full dimensionality of sets of feasible utilities is used in a more fundamental way in some of our other theorems. For example, in the proof of

[34] A social evaluation functional is *antidictatorial* if the alternatives are always socially ranked in the reverse order of the dictator's preference and it is *null* if for each profile in the domain, all alternatives are socially indifferent.

[35] See Coulhon and Mongin (1989) and Weymark (1991) for discussions of the uniqueness issue in Harsanyi's social aggregation problem.

[36] See Blackorby, Donaldson, and Weymark (1999) for details.

Theorem 6.7, the full dimensionality of the sets of feasible utilities for the two profiles considered is needed to obtain the contradiction that establishes the theorem.

Utilitarianism requires that all individuals' interests receive the same weight. In the multiprofile characterization of weighted utilitarianism (Theorem 6.8), the individual weights are all equal if we add an *anonymity* assumption that requires the social evaluation to be invariant to any permutation of the profile. Anonymity is vacuous on our single-profile and single-information-set domains. Consequently, it is not possible to obtain characterizations of utilitarianism by adding anonymity to the assumptions of Theorems 6.6 and 6.10.[37]

We have used state-contingent alternatives to model uncertainty. We could equally well have used Harsanyi's lottery model. Some of the technical details would be different (for example, with lotteries it is not possible to have $\mathbb{R}^n$ as the feasible set of utilities for a single profile), but lottery analogues exist for each of our theorems.

The results of this article show that, for a strongly Paretian social evaluation functional, several factors are critical in determining the implications of requiring the individual utility functions and the social evaluations to satisfy the expected utility hypothesis. These factors are (i) whether only utility functions that satisfy the Bernoulli hypothesis are regarded as satisfying the expected utility hypothesis; (ii) whether the domain is single profile, single-preference profile, single information set, or multiprofile; (iii) which information partition is employed; and (iv) whether the social evaluation functional is required (either directly or indirectly) to be strongly neutral.

For Bernoulli expected utility profiles, the social evaluation functional is weighted utilitarian in the single-profile, single-information-set, and multiprofile cases, provided that in the latter two cases there is a cardinal unit-comparable plus information partition and the social evaluation functional is strongly neutral. For expected utility profiles that are not Bernoulli, the social evaluation functional is transformed utilitarian in the single-profile and single-information-set cases; that is, each social evaluation is obtained by taking a weighted sum of transformed utilities, where the transforms are the inverses of the individual transforms required to convert a Bernoulli representation of an individual's preferences into his or her actual utility function. For the single-information-set result, the social evaluation functional is assumed to be strongly neutral and the information set must be

[37] Coulhon and Mongin (1989) and Mongin (1994) have made a similar observation about Harsanyi's theorem.

obtainable from an information set in a cardinal unit-comparable plus information partition by applying the individual transforms just described to each of the profiles in this set. In the multiprofile case with an unrestricted expected utility domain, an impossibility results no matter how much utility information is available when the social evaluation functional is strongly neutral. Without Strong Neutrality, the axioms are consistent on a single-preference-profile domain. However, because it is not welfarist, it is not possible to interpret the social evaluation functional as being either transformed or weighted utilitarian.

We have argued that expected utility theory provides no good reason for restricting the domain a priori to only profiles of expected Bernoulli utility functions. We therefore conclude that Harsanyi's idea of requiring individual and social preferences to satisfy the expected utility hypothesis does not, by itself, provide a defensible argument for weighted utilitarianism. But, as we have already noted, there may be other reasons, such as the ones provided by Broome (1991), for only considering Bernoulli expected utility profiles. Broome identifies utilities with individual good or well-being. This permits him to distinguish well-being from preferences and to apply the requirements of expected utility theory normatively. In that context, he provides additional arguments to support the claim that individual good is the expected value of Bernoulli utilities at "best-information" probabilities that are common to all individuals. In such a setting, if Broome's arguments are found to be compelling, our theorems *do* provide support for weighted utilitarianism and, in the multiprofile case, for utilitarianism itself if individuals are required to be treated impartially.

Another possible response to our findings is to question the use of ex ante Pareto principles. Critics of the ex ante approach to social evaluation argue that individual and social rationality are different and, at times, incompatible when alternatives are uncertain. However, the ex post approach to social evaluation is not without its own difficulties.[38]

References

Arrow, K. J. 1951. *Social Choice and Individual Values.* Wiley, New York.

Arrow, K. J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. In *Econométrie.* Centre National de la Recherche Scientifique, Paris, pp. 41–47, translated as Arrow (1964).

Arrow, K. J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, 91–96.

[38] See Hammond (1981, 1983) for discussions of ex ante and ex post social choice.

Arrow, K. J. 1965. *Aspects of the Theory of Risk-Bearing*. Academic Bookstore, Helsinki.

Blackorby, C., Davidson, R., and Donaldson, D. 1977. A homiletic exposition of the expected utility hypothesis. *Economica* 44, 351–358.

Blackorby, C., Donaldson, D., and Weymark, J. A. 1980. On John Harsanyi's defences of utilitarianism. Discussion Paper No. 8013, Center for Operations Research and Econometrics, Université Catholique de Louvain.

Blackorby, C., Donaldson, D., Weymark, J. A. 1984. Social choice with interpersonal utility comparisons: A diagrammatic introduction. *International Economic Review* 25, 327–356.

Blackorby, C., Donaldson, D., Weymark, J. A. 1990. A welfarist proof of Arrow's theorem. *Recherches Economiques de Louvain* 56, 259–286.

Blackorby, C., Donaldson, D., Weymark, J. A. 1999. Harsanyi's social aggregation theorem for state-contingent alternatives. *Journal of Mathematical Economics* 32, 365–387.

Bossert, W., and Weymark, J. A. 2004. Utility in social choice. In *Handbook of Utility Theory*, Vol. 2. *Extensions,* ed. S. Barberà, P. J. Hammond, and C. Seidl, Kluwer Academic, Boston, pp. 1099–1177.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty and Time.* Basil Blackwell, Oxford.

Broome, J. 2008. Can there be a preference-based utilitarianism? In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark. Cambridge University Press, Cambridge, pp. 221–238.

Coulhon, T., and Mongin, P. 1989. Social choice theory in the case of von Neumann–Morgenstern utilities. *Social Choice and Welfare* 6, 175–187.

d'Aspremont, C., 1985. Axioms for social welfare orderings. In *Social Goals and Social Organization: Essays in Memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein. Cambridge University Press, Cambridge, pp. 19–76.

d'Aspremont, C., and Gevers, L., 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44, 199–209.

Deschamps, R., and Gevers, L., 1977. Separability, risk-bearing and social welfare judgements. *European Economic Review* 10, 77–94.

Eichhorn, W. 1978. *Functional Equations in Economics.* Addison-Wesley, Reading, MA.

Hammond, P. J. 1979. Equity in two-person situations: Some consequences. *Econometrica* 47, 1127–1135.

Hammond, P. J. 1981. *Ex-ante* and *ex-post* welfare optimality under uncertainty. *Economica* 48, 235–250.

Hammond, P. J. 1983. Ex-post optimality as a dynamically consistent objective for collective choice under uncertainty. In *Social Choice and Welfare*, ed. P. K. Pattanaik and M. Salles. North-Holland, Amsterdam, pp. 175–205.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, Cambridge.

Hens, T. 1992. A note on Savage's theorem with a finite number of states. *Journal of Risk and Uncertainty* 5, 63–71.

Maskin, E. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45, 93–96.

Mongin, P. 1994. Harsanyi's aggregation theorem: Multi-profile version and unsettled questions. *Social Choice and Welfare* 11, 331–354.

Mongin, P., and d'Aspremont, C. 1998. Utility theory and ethics. In *Handbook of Utility Theory*, Vol. 1: *Principles,* ed. S. Barberà, P. J. Hammond, and C. Seidl. Kluwer Academic, Boston, pp. 371–481.

Pratt, J. W. 1964. Risk aversion in the small and in the large. *Econometrica* 32, 122–136.

Radó, F., and Baker, J. A. 1987. Pexider's equation and aggregation of allocations. *Aequationes Mathematicae* 32, 227–239.

Roberts, K. W. S. 1980. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47, 421–439.

Roemer, J. E. 1996. *Theories of Distributive Justice.* Harvard University Press, Cambridge, MA.

Roemer, J. E. 2008. Harsanyi's impartial observer is *not* a utilitarian. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls,* ed. M. Fleurbaey, M. Salles, and J. A. Weymark. Cambridge University Press, Cambridge, pp. 129–135.

Samuelson, P. A. 1977. Reaffirming the existence of "reasonable" Bergson–Samuelson social welfare functions. *Economica* 44, 81–88.

Savage, L. J. 1954. *The Foundations of Statistics.* Wiley, New York.

Sen, A. K. 1970. *Collective Choice and Social Welfare.* Holden-Day, San Francisco.

Sen, A. K. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision 7,* 243–262.

Sen, A. K. 1977. On weights and measures: Informational constraints in social welfare analysis. *Econometrica* 45, 1539–1572.

von Neumann, J., and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd ed. Princeton University Press, Princeton, NJ.

Wakker, P. 1993. Additive representations on rank-ordered sets. II: The topological approach. *Journal of Mathematical Economics* 22, 1–26.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, pp. 255–320.

Weymark, J. A. 2005. Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare* 25, 527–555.

7

A Welfarist Version of Harsanyi's Aggregation Theorem

Claude d'Aspremont and Philippe Mongin

7.1 Introduction

The Aggregation Theorem is one of the main arguments used by Harsanyi in support of utilitarian ethics. It was first presented in his 1955 article and further developed in chapter 4 of his book in 1977. Since then, several authors[1] have constructed alternative proofs of this theorem in more general settings. It is generally presented as relating a "single profile" of individual utility functions $\{U_i\}$, to the utility function W of a moral observer by means of the Pareto Indifference rule. In this context, the theorem states that if all utility functions (including the moral observer's) are von Neumann–Morgenstern (VNM), then the moral observer's utility is an affine transformation of the individual utilities, that is, $W = \sum \beta_i U_i + \gamma$.

The relevance of this result in giving proper foundations to utilitarianism has been questioned on several grounds. First, the weights $\{\beta_i\}$ are not necessarily positive, and hence the welfare of some individuals might not affect, or worse, might negatively affect total welfare. This first problem can be solved quite naturally by strengthening Pareto Indifference into the

[1] See Domotor (1979), Border (1981, 1985), Fishburn (1984), Selinger (1986), Coulhon and Mongin (1989), Hammond (1992), and Weymark (1993).

This chapter was first presented at the conference on Justice, Political Liberalism, and Utilitarianism held in honour of John Harsanyi and John Rawls at the University of Caen, June 1996. We are grateful to Charles Blackorby, Vincenzo Denicolo, David Donaldson, Marc Fleurbaey, Louis Gevers, and Serge Kolm for useful discussions and comments. Special thanks are due to John Weymark for suggestions and encouragement in the writing of this final version. This text presents research results of the Belgian Program on Interuniversity Poles of Attraction (Science Policy Programming, Prime Minister's Office). Scientific responsibility is assumed by its authors.

Strong Pareto condition; the latter implies that all weights are positive.[2] A second problem is that the weights might not be uniquely defined, creating an indeterminacy. This further problem can be solved by introducing an additional condition, called Independent Prospects, which says that for every individual there exists a pair of lotteries for which that individual alone is not indifferent.[3] The third problem, which is one of the main objections formulated by Sen (1986) against the Aggregation Theorem as an axiomatisation of utilitarianism, is that the weights cannot be determined independently of the utility functions to be aggregated; indeed, if the β_i's are functions of the U_i's, the formula is different from a utilitarian rule. A related issue is how to obtain, in the context of Harsanyi's theorem, the pure classical utilitarian rule, with all weights equal to 1 (Bentham's sum rule) or to $1/n$ (as in the average utility rule). To determine the weights independently from the given utilities, and eventually to get equal weights by introducing a symmetry condition, one needs to consider a more general framework, allowing the utility profiles to vary. As suggested in Coulhon and Mongin (1989) and Mongin (1994), this can be conveniently done in Sen's (1970) framework of social welfare functionals (SWFLs). The Aggregation Theorem can then be reformulated so as to give an axiomatisation which, at least formally, relates to Utilitarianism.[4]

This chapter elaborates on this reformulation. It will not, though, start from Sen's multiprofile approach – with SWFLs defined on some universal domain – but instead from the "enlarged" single-profile approach used in Roberts (1980a) and d'Aspremont (1985), with SWFLs being defined on a restricted domain. More specifically, we will closely follow Harsanyi in assuming a *single profile of individual VNM preferences* and allow for *multiple profiles of VNM utility functions*, representing these given preferences. Following an argument in d'Aspremont (1985), this will be sufficient to obtain a VNM version of Welfarism and, thus, to introduce conditions that are usually stated in the multi-profile approach. One such condition is Anonymity, which will imply a symmetric formula. Another is Cardinality and Unit Comparability, an invariance axiom that allows for (some version of) interpersonal comparisons of utility differences. Following Mongin (1994), this cardinal condition will be shown to result from assigning VNM utility

[2] See Domotor (1979), Weymark (1993), and De Meyer and Mongin (1995). This was already suggested by Harsanyi (1955).

[3] See Fishburn (1984) and Coulhon and Mongin (1989). This condition was used implicitly, as a structural assumption, by Harsanyi in the proof of the Aggregation Theorem. Domotor (1979) and Border (1981) showed that it was not needed.

[4] See also Mongin and d'Aspremont (1998).

functions to the individuals and VNM preferences to Harsanyi's moral observer. Our results are closely related to the ones given by Blackorby, Donaldson, and Weymark (2008). They also investigate how the expected utility hypothesis, combined with a Paretian condition, can provide support for utilitarianism. However, their investigation is done for other domains than the one considered here.

This chapter is organized as follows. In the next section, we define a social welfare functional restricted to the domain of all VNM representations of a given single profile of VNM individual preferences, and state the corresponding Aggregation Theorem. Then, in Section 7.3, we derive Welfarism, prove the theorem, and derive the VNM characterizations of pure and generalized utilitarian rules. Finally, in the concluding section, we show that, under the VNM domain restriction adopted here, two standard cardinality notions are equivalent.

7.2 A SWFL Version of the Aggregation Theorem for a Single Profile of VNM Preferences

The social choice problem to be considered here is defined by a set of individuals $N = \{1, 2, \dots, n\}$, a set of social states X, and a "moral observer." According to Harsanyi's approach, the moral observer is any individual, adopting a moral point of view and forming moral preferences (to be distinguished from this individual's personal preferences). But the moral rule to be finally determined should, in principle, be the same for each individual. The objective is to derive (some version of) the utilitarian rule. The set X of social states is not precisely interpreted. Following Harsanyi, who claims to be a rule-utilitarian, it could be the set of all possible rules to constrain individual behavior (including all sorts of possible amendments), or, more specifically, some given set of possible rules and all probability mixtures (i.e., lotteries) on this set. Formally, X is supposed to be a convex subset (which is not a singleton) of some vector space: for any $x, y \in X$ and any $\lambda \in [0, 1]$, the convex combination (or mixture) $[\lambda x + (1 - \lambda)y]$ is also in X.

For any set Ξ (which may be X or some other convex set that will be introduced in the sequel), a *preference ordering* R on Ξ is a reflexive, complete and transitive binary relation on Ξ. Moreover, it is a *VNM preference ordering on* Ξ if it satisfies in addition:

Continuity: $\forall a, b, c \in \Xi$, the sets $\{\lambda \in [0, 1] : cR[\lambda a + (1 - \lambda)b]\}$ and $\{\lambda \in [0, 1] : [\lambda a + (1 - \lambda)b]\, Rc\}$ are closed in $[0, 1]$.

Independence: $\forall a, b, c \in \Xi, \forall \lambda \in]0, 1], aRb \Leftrightarrow [\lambda a + (1 - \lambda)c]\, R[\lambda b + (1 - \lambda)c]$.

A VNM preference ordering R can always be represented by a utility function υ defined on Ξ: $\forall a, b \in \Xi, aRb \Longleftrightarrow \upsilon(a) \geq \upsilon(b)$. Moreover, in this framework, every utility representation of R is either *mixture-preserving*, that is,

$$\forall a, b \in \Xi, \forall \lambda \in [0, 1], \upsilon(\lambda a + (1 - \lambda)b) = \lambda \upsilon(a) + (1 - \lambda)\upsilon(b),$$

or a monotone transformation of a mixture-preserving utility function. A mixture-preserving utility representation is called a *VNM utility function.*

We start with a given *single profile* of individual preference orderings $(\bar{R}_i)_{i \in N}$. This will remain fixed throughout. Our first assumption is that each $\bar{R}_i$ is a VNM preference ordering on the set X of social states, which is *nontrivial* in the sense of being different from total indifference (for each $i \in N$, there exist $x, y \in X$ such that i strictly prefers x to y: $x \bar{P}_i y$). A *social welfare functional* (SWFL) is a function F associating to each utility profile $U = (U_1, U_2, ..., U_n)$ in some admissible domain $\mathcal{D}$, a preference ordering $F(U)$ on X. The objective here is to associate to the single profile of VNM preferences $(\bar{R}_i)_{i \in N}$ a particular SWFL $\bar{F}$, satisfying a set of conditions. The first three conditions are directly linked to Harsanyi's basic axioms: The first determines the domain of the SWFL, the second fixes its range, and the third is a strenghtening of Pareto Indifference. The last axiom is a weakening of the structural assumption used by Harsanyi.

VNM-Utility Domain (VNM-D): For every $i \in N$, $\bar{R}_i$ is a nontrivial VNM preference ordering on X. The domain of $\bar{F}$ is the set $[\bar{U}]$ of all vectors of possible individual VNM utility representations of $(\bar{R}_i)_{i \in N}$; that is, $\mathcal{D} = [\bar{U}]$.

VNM-Range (VNM-R): For any $U \in [\bar{U}]$, the moral observer's preference ordering $\bar{F}(U)$ satisfies continuity (**VNM1-R**) and independence (**VNM2-R**).

These first two conditions reflect Harsanyi's commitment to the VNM preference axioms as a norm of rationality for both the personal and moral preferences. It will become clear that the restriction on the domain, as well as the restriction on the range of the SWFL, plays an important role in moving away from an ordinal noncomparable framework and in giving some ethical relevance to the rules that will be derived. Indeed, Harsanyi's choice to restrict consideration to the class of VNM (i.e., mixture-preserving) utility representations of each individual preference is used in the next section to transpose the VNM-Range condition (both continuity and independence) to the welfarist framework (obtained after the last two conditions have been introduced). Then, eventually, the welfarist version of

VNM-independence will be shown equivalent to cardinality and interpersonal unit comparability. This fact gives some foundation to Harsanyi's claim [or Vickrey's (1945)] for basing the determination of moral preferences on individual attitudes toward risk or, more precisely, on the various factors explaining these attitudes. In other terms, VNM representations of individual preferences provide cardinal information to a VNM rational moral observer.

This claim should not be interpreted as meaning that an individual's risk attitudes are not already contained in his VNM preferences (the primitive of expected utility theory) and cannot be represented by nonmixture preserving utility functions. This is justly stressed by Sen (1986), Weymark (1991), and Blackorby, Donaldson, and Weymark (2008). For example, taking a single-dimension outcome space and any one particular VNM utility representation of some individual's VNM preference relation, we can compute the corresponding Arrow-Pratt (absolute or relative) measure of risk aversion. It is a consequence of the VNM theorem that this piece of information about this individual's risk attitude can be recovered from any other utility representation of his VNM preferences because any such representation has to be a monotone transformation of the VNM utility used in computing the measure. However, it is only when this other representation is itself VNM (i.e., taking the transformation to be positive affine) that one can recompute directly (without making some preliminary ordinal retransformation) the Arrow-Pratt measure and get the same number. In other words, the Arrow-Pratt measure is only invariant to positive affine transformations. It is this particular invariance property, holding within the class of all VNM representations, that is exploited in Harsanyi's approach (as described for instance by our four axioms), the objective being not simply to get an evaluation of each single individual's risk attitude but to allow for interpersonal comparisons of risk attitudes, that is, to make sense of statements such as "individual i is more risk averse than individual j, in the Arrow-Pratt sense."

What we want to stress here, though, is that such an interpersonal cardinalization is not a consequence of just the domain restriction but of a combination of this restriction and of the one limiting the range of the SWFL to VNM preference orderings on X. Moreover, we need the other two conditions.

The third one replaces Harsanyi's Pareto Indifference. Having adopted an *enlarged* single-profile approach, Pareto Indifference needs to be strengthened. It is replaced by a neutrality condition, restricted to the set of VNM utility representations.

Relative Neutrality (RN): For any $U, U' \in [\bar{U}]$, any two pairs $\{x, y\}$ and $\{x', y'\}$, if $U(x) = U'(x')$ and $U(y) = U'(y')$, then $x\bar{R}y \Leftrightarrow x'\bar{R}'y'$, with $\bar{R} = \bar{F}(U)$ and $\bar{R}' = \bar{F}(U')$.

To see that RN implies Pareto Indifference, it is enough to put $U' = U$, $x = y'$, and $x' = y$. One needs, finally, a structural assumption not directly imposed on $\bar{F}$ but on the set X and on the given single profile of individual preferences. It ensures that any three vectors in the n-dimensional utility space (the real Euclidean space indexed by the names of the individuals), denoted by E^N, is attainable.[5]

Relative Attainability (RA): For any $u, v, w \in E^N$, there are $x, y, z \in X$ and $U \in [\bar{U}]$ such that $U(x) = u$, $U(y) = v$, and $U(z) = w$.

This assumption is weaker than Harsanyi's own structural assumption, Independent Prospects. The latter could be rephrased as saying that, for each VNM utility profile, the range of that profile has full dimension, hence that any n-tuple of vectors in E^N can be attained from $[\bar{U}]$.

The following is now a version of the Aggregation Theorem, adapted to the present enlarged single-profile approach.

Theorem 7.1: *If the SWFL $\bar{F}$ satisfies conditions VNM-D, VNM-R, RN, and RA, then there is a real vector of weights $(\beta_1, \ldots, \beta_n)$, unique up to a positive scale factor, such that for all $U \in [\bar{U}]$, for all $x, y \in X$, and $\bar{R} = \bar{F}(U)$,*

$$x\bar{R}y \Leftrightarrow \sum_{i=1}^{n} \beta_i U_i(x) \geq \sum_{i=1}^{n} \beta_i U_i(y).$$

The proof is delayed until the next section. There, it will mainly be argued that Harsanyi's theorem is best seen as a welfarist result. Under Welfarism, another version of the Aggregation Theorem will be stated and proved. This version will have the advantage of making clear the cardinal content of the theorem. Also, this further version will incorporate two additional assumptions directly stated in welfarist terms, Strict Pareto and Anonymity, to ensure, respectively, that the weights (β_i) are all positive and that they are all equal.

[5] In d'Aspremont (1985), this condition is called "Unrestricted Individual Utility Profile" (UP).

7.3 A Welfarist Version of the Aggregation Theorem

The first result in this section shows that the SWFL $\bar{F}$ restricted to $[\bar{U}]$, as in Theorem 7.1, can be used to derive *welfarism.* This property means that the preference ordering of the moral observer $\bar{F}(U)$ on X can be translated, whatever $U \in [\bar{U}]$, into a *social welfare ordering* (SWO), that is, a preference ordering R^* defined on the n-dimensional utility space E^N. This moral observer is then truly consequentialist in the sense of taking into account only the utility consequences of all social states and not the social states themselves. In addition, he will be a VNM decision maker because R^* will be shown to satisfy:

VNM-Social Welfare Ordering (VNM-R*): The moral observer's preference ordering R^* defined on E^N satisfies continuity (**VNM1***) and independence (**VNM2***).

The following lemma combines results from d'Aspremont (1985) and Mongin (1994).

Lemma 7.1 (VNM Welfarism): *If the SWFL $\bar{F}$ satisfies VNM-D, RN, and RA, then there exists a SWO R^* defined on E^N such that: For any $x, y \in X$, for any $U \in [\bar{U}]$, and $\bar{R} = \bar{F}(U)$, if $U(x) = u$ and $U(y) = v$, then $uR^*v \Leftrightarrow x\bar{R}y$. Moreover, if $\bar{F}$ also satisfies VNM-R, then R^* satisfies VNM-R*.*

Proof: The first step in the proof consists in constructing a binary relation R^* on E^N. By RA, we may take, for every $u, v \in E^N$, some $x, y \in X$ and $U \in [\bar{U}]$ such that $U(x) = u$ and $U(y) = v$ and put: $uR^*v \Leftrightarrow x\bar{F}(U)y$. The relation R^* is well-defined by RN: for any other profile $U' \in [\bar{U}]$ and pair $\{x', y'\} \subset X$ such that $U'(x') = u$ and $U'(y') = v$, $x'\bar{F}(U')y' \Leftrightarrow x\bar{F}(U)y$. It is complete and transitive because of, respectively, the completeness and the transitivity of $\bar{F}(U)$ for any $U \in [\bar{U}]$.

The second step is to show that VNM-R implies VNM-R*. Consider VNM-independence first. We have to get the conclusion that VNM-2* holds, namely, that $\forall u, v, w \in E^N$, $\forall \lambda \in [0, 1]$,

$$uR^*v \Leftrightarrow [\lambda u + (1-\lambda)w]\, R^*[\lambda v + (1-\lambda)w].$$

By RA (or by the definition of R^*), there are x, y, and z in X and $U \in [\bar{U}]$ such that $U(x) = u$, $U(y) = v$, and $U(z) = w$, and using VNM1-R,

$$x\bar{F}(U)y \Leftrightarrow (\lambda x + (1-\lambda)z)\bar{F}(U)(\lambda y + (1-\lambda)z).$$

So, by the definition of R^*, we get

$$U(x)R^*U(y) \Leftrightarrow U(\lambda x + (1-\lambda)z)R^*U(\lambda y + (1-\lambda)z).$$

Since VNM-D holds, U is mixture-preserving, that is,

$$U(\lambda x + (1-\lambda)z) = \lambda U(x) + (1-\lambda)U(z),$$
$$U(\lambda y + (1-\lambda)z) = \lambda U(y) + (1-\lambda)U(z).$$

The conclusion follows. To derive VNM-1*, a similar argument can be used. □

From now on, we may as well assume that the moral observer's preferences are given by a VNM social welfare ordering R^* on E^N (which amounts to assuming VNM-R, VNM-D, RA, and RN) and introduce additional axioms directly on this R^*. But first, let us prove Theorem 7.1.

Proof of Theorem 7.1: Because the preference ordering R^*, defined on the convex set E^N, satisfies VNM-R*, it has a VNM utility representation W. This mixture-preserving function is affine on E^N, that is, for all $u \in E^N$, $W(u) = \sum_{i \in N} \beta_i u_i + \gamma$, for some vector $(\beta_1, \ldots, \beta_n)$ and some scalar γ (for the equivalence of mixture-preserving and affine functions on convex sets, see, e.g., Coulhon and Mongin, 1989). Moreover, any other VNM representation, with weights $(\beta'_1, \ldots, \beta'_n)$, must be such that $\beta'_i = \lambda \beta_i$, for some $\lambda > 0$ and all $i \in N$. □

To understand better the ethical relevance of this result, another observation is in order. This is the logical equivalence between the independence axiom (VNM2*) and a well-known invariance property of the SWO R^*, stating the minimal kinds of measurability (cardinality) and interpersonal comparability (unit comparability), which are compatible with utilitarianism.

Cardinality and Unit Comparability (**CU***): For any $u, v \in E^N$, any vector $(\alpha_1, \ldots, \alpha_n)$, and any $\beta > 0$, if $u'_i = \alpha_i + \beta u_i$ and $v'_i = \alpha_i + \beta v_i$ for all $i \in N$, then $uR^*v \Leftrightarrow u'R^*v'$.

The following argument, given in Mongin (1994), is close to the one used by Harsanyi to show the linear homogeneity of the function W representing moral preferences (1977, chapter 4, Lemma 4).

Lemma 7.2 (VNM cardinality): *A SWO R^* on E^N satisfies CU* if and only if it satisfies VNM2*.*

Proof: Suppose first that CU* holds. We want to show that VNM2* holds, that is, that, for $u, v \in E^N$, uR^*v if and only if for any $\lambda \in]0, 1]$ and $w \in E^N$, $[\lambda u + (1-\lambda)w]R^*[\lambda v + (1-\lambda)w]$. Taking $\alpha_i = (1-\lambda)w_i$, for every i, and $\beta = \lambda$, this equivalence immediately follows from CU*. Second, to prove the converse, suppose that VNM2* holds and take any vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta > 0$. If $\beta < 1$, we can simply put $w = \alpha/(1-\beta)$ and $\lambda = \beta$, then apply VNM2*. If $\beta > 1$, clearly $uR^*v \Leftrightarrow \frac{1}{2\beta}(2\beta u)R^*\frac{1}{2\beta}(2\beta v)$, which by VNM2* is equivalent to $(2\beta u)R^*(2\beta v)$ [letting $w \equiv 0$ and $\lambda = 1/(2\beta)$]. To get the conclusion, it is then enough to let $\lambda = 1/2$ and $w = 2\alpha$, and apply VNM2* again. □

This shows that VNM2* (hence, granting welfarism, VNM2-R) implies the axiom that traditionally formalizes the possibility of making interpersonal comparisons of utility differences. If two utility vectors $u, v \in E^N$ are transformed into two vectors $u', v' \in E^N$ according to CU*, then for any $i, j \in N$

$$u_i - v_i \geq u_j - v_j \Leftrightarrow u'_i - v'_i \geq u'_j - v'_j.$$

This invariance property is clearly important from a moral point of view. However, it might be objected that this property is here only a necessary condition, not a sufficient one. We shall come back to this objection (which is relevant to both VNM2* and CU*) in the next section. We now pursue the task of deriving an improved version of the Aggregation Theorem from an ethical point of view.

It seems also important that all individuals be given positive weights. This is ensured by adding the following condition.

Strict Pareto (S-P*): If $u, v \in E^N$ are such that $u_i \geq v_i$, for all $i \in N$, and $u_j > v_j$, for some $j \in N$, then uP^*v.

In conjunction with Pareto Indifference (which is satisfied by construction in a welfarist framework), this principle is equivalent to the usual Strong Pareto principle.

To give positive weight to each individual might even be considered as insufficient. It is an advantage of our welfarist approach – as opposed to the initial Harsanyi approach where only a single profile of individual preferences was considered – to make it possible to formulate an anonymity axiom.

This axiom will make the chosen weights definitely independent from the single profile fixed at the outset.

Anonymity (A*): For all $u \in E^N$, and any permutation σ of N,

$$uI^*(u_{\sigma_1}, \dots, u_{\sigma_n}).$$

We may, finally, state the two welfarist versions of Harsanyi's Aggregation Theorem characterizing utilitarian rules, one of which is anonymous and the other not. These theorems can be seen as alternative versions of already known welfarist characterizations of utilitarianism.

Theorem 7.2 (Pure Utilitarianism): *If the SWO R* satisfies S-P*, A*, and VNM2*, then for all* $u, v \in E^N$,

$$uR^*v \Leftrightarrow \sum_{i=1}^{n} u_i \geq \sum_{i=1}^{n} v_i.$$

Several proofs of this theorem are available, bearing in mind that in a context of cardinal comparisons, Anonymity implies the suitable notion of continuity for the SWO. More precisely, in the presence of CU* (or, equivalently, VNM2*), A* implies VNM1*. One proof relies on Theorem 7.1 (as in Mongin, 1994). Another uses the equivalence between VNM2* and CU*, as well as the characterization of the pure utilitarian rule in terms of the latter condition (see d'Aspremont and Gevers, 1977). In either case, axiom A* is to be part of the conditions.

Theorem 7.3 (Generalized Utilitarianism): *If the SWO R* satisfies S-P* and VNM-R*, then there is a real vector of positive weights* $(\beta_1, \dots, \beta_n)$, *such that for all* $u, v \in E^N$,

$$uR^*v \Leftrightarrow \sum_{i=1}^{n} \beta_i u_i \geq \sum_{i=1}^{n} \beta_i v_i.$$

Again, this result can be seen as a corollary to Theorem 7.1. Alternatively, the proof can use a theorem characterizing "weak utilitarianism" (i.e., $\sum_{i=1}^{n} \beta_i u_i > \sum_{i=1}^{n} \beta_i v_i \Rightarrow uR^*v$ for some positive weights $\beta_1, \dots, \beta_n$) in terms of S-P*, VNM1*, and CU*; see, e.g., Blackwell and Girschik (1954), Roberts (1980b), and d'Aspremont (1985). Using this last reference (Theorem 3.3.5), it is easy to get generalized utilitarianism by showing that the continuity condition VNM1* implies the following: for any $i, j \in N$, there exist $u, v \in E^N$ such that $u \neq v$, $u_h = v_h$ for $i \neq h \neq j$, and vI^*u. To show this property (called *Weak Anonymity*), it is enough to pick $u, a, b \in E^N$, with $a_h = b_h = u_h$, for $i \neq h \neq j$, such that u is not a convex combination of a and b, but aR^*uR^*b. Then, by VNM1*, for some $\lambda \in [0, 1]$ and $v = [\lambda a + (1 - \lambda)b]$, we get vI^*u.

7.4 Concluding Remarks: More on SWFLs and Cardinality

Once stated in an appropriate framework, that is, the welfarist framework, the Aggregation Theorem performs no worse and no better, from an ethical point of view, than existing characterizations of the utilitarian rule. It offers an alternative but equivalent axiomatization. This results from the equivalence between the VNM-independence axiom (VNM2*) and the cardinality-with-unit-comparability axiom (CU*), as imposed on the social welfare ordering. In our presentation, VNM2* was taken to be the welfarist translation of VNM2-R, the VNM-independence axiom imposed on the SWFL $\bar{F}$. Knowing now this equivalence, VNM2* can as well be viewed as the translation of an axiom of interpersonal utility comparison, which would be imposed from the start on $\bar{F}$. More formally, under RN and RA, the condition CU* is equivalent to the following:

Cardinality and Unit Comparability (CU): For any $U \in [\bar{U}]$, any vector $\alpha = (\alpha_1, \ldots, \alpha_n)$, and any $\beta > 0$, if $U' = \beta U + \alpha$, then $\bar{F}(U) = \bar{F}(U')$.

This is cardinality in a specific sense, to be compared with cardinality in the larger and more meaningful sense of preserving interpersonal utility differences. This other axiom is (see Bossert and Weymark, 1997):

Interpersonal Difference Comparability (IRDC): For any $U, U' \in [\bar{U}]$, if, for all $x, y, x', y' \in X$ and all $i, j \in N$,

$$U_i(x) - U_i(y) \geq U_j(x') - U_j(y') \Leftrightarrow U'_i(x) - U'_i(y) \geq U'_j(x') - U'_j(y'),$$

then $\bar{F}(U) = \bar{F}(U')$.

In general, conditions on the individual utility functions are needed to get the equivalence between these two cardinality principles. An interesting fact, in the context of Harsanyi's Aggregation Theorem, is that one such sufficient condition is VNM-D.

Lemma 7.3 (Cardinality): *If the SWFL $\bar{F}$ satisfies VNM-D, then CU is equivalent to IRDC.*

Proof: That IRDC implies CU is easily verified. For any $U \in [\bar{U}]$, any vector $\alpha = (\alpha_1, \ldots, \alpha_n)$, and any $\beta > 0$, if $U' = \beta U + \alpha$, then obviously, for all $x, y, x', y' \in X$ and all $i, j \in N$, differences orderings are preserved, that is,

$$U_i(x) - U_i(y) \geq U_j(x') - U_j(y') \Leftrightarrow U'_i(x) - U'_i(y) \geq U'_j(x') - U'_j(y'),$$

so that $\bar{F}(U) = \bar{F}(U')$ by IRDC.

For the reverse implication, select any $U, U' \in [\bar{U}]$ preserving all differences orderings. By VNM-D, U and U' are nontrivial and there are $x^i, y^i, z^i \in X$ and $\lambda^i \in [0, 1]$, for every $i \in N$, such that

$$z^i = \lambda^i x^i + (1 - \lambda^i) y^i,$$
$$U_i(x^i) - U_i(z^i) = U_j(x^j) - U_j(z^j) > 0,$$

hence,

$$U'_i(x^i) - U'_i(z^i) = U'_j(x^j) - U'_j(z^j) > 0,$$

for all $j \in N$. But, by VNM-D again, for every $i \in N$, we must have $U'_i = \alpha_i + \beta_i U_i$ for some α_i and some $\beta_i > 0$. Using the above equalities, we obtain $\beta_i = \beta_j$, for all $i, j \in N$. By CU, it implies $\bar{F}(U) = \bar{F}(U')$. □

In other words, to restrict individual utility functions, as Harsanyi does, to nontrivial VNM representations entails equivalence between the two definitions of cardinality. This conclusion holds more generally in a multiprofile approach, for a SWFL F defined on a domain $\mathcal{D}$ of profiles of mixture-preserving individual utility functions (not all trivial). The conditions CU and IRDC have simply to be rephrased by substituting F for $\bar{F}$ and $\mathcal{D}$ for $[\bar{U}]$.

This chapter has shown that an "enlarged" single-profile approach leads to reformulating Harsanyi's Aggregation Theorem in welfarist terms and thus turns it into an alternative characterization of utilitarianism, along standard lines in social choice theory. The theorem may now include an anonymity condition and seems compatible with meaningful comparisons of cardinal utility functions. Whatever ethical content it has depends essentially on the following three assumptions: to consider only VNM representations of the individual preferences, to strengthen Pareto Indifference so as to get welfarism, and to impose VNM-independence on the moral observer's preferences. These three conditions appear to constitute the proper content of Harsanyi's particular approach to utilitarianism.

References

Blackorby, C., Donaldson, D., and Weymark, J. A. 2008. Social aggregation and the expected utility hypothesis. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark. Cambridge University Press, Cambridge, pp. 136–183.

Blackwell, D., and Girshick, M. A. 1954. *Theory of Games and Statistical Decisions.* Wiley, New York.

Border, K. 1981. Notes on von Neumann-Morgenstern social welfare functions. Division of the Humanities and Social Sciences, California Institute of Technology (unpublished).

Border, K. 1985. More on Harsanyi's utilitarian cardinal welfare function. *Social Choice and Welfare* 1, 279–281.

Bossert, W., and Weymark, J. A. 2004. Utility in social choice. In *Handbook of Utility Theory,* Vol. 2: *Extensions*, ed. S. Barberà, P. J. Hammond, and C. Seidl. Kluwer Academic, Boston, pp. 1099–1177.

Coulhon, T., and Mongin, P. 1989. Social choice theory in the case of von Neumann-Morgenstern utilities. *Social Choice and Welfare* 6, 175–187.

d'Aspremont, C. 1985. Axioms for social welfare orderings. In *Social Goals and Social Organizations: Essays in Memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein. Cambridge University Press, Cambridge, pp. 19–76.

d'Aspremont, C., and Gevers, L. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44, 199–209.

De Meyer, B., and Mongin, P. 1995. A note on affine aggregation. *Economics Letters* 47, 177–183.

Domotor, Z. 1979. Ordered sum and tensor product of linear utility structures. *Theory and Decision* 11, 375–399.

Fishburn, P. C. 1984. On Harsanyi's utilitarian cardinal welfare theorem. *Theory and Decision* 17, 21–28.

Hammond, P. J. 1992. Harsanyi's utilitarian theorem: A simpler proof and some ethical considerations. In *Rational Interactions: Essays in Honor of John C. Harsanyi*, ed. R. Selten. Springer-Verlag, Berlin, pp. 305–319.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, Cambridge.

Mongin, P. 1994. Harsanyi's aggregation theorem: Multi-profile version and unsettled questions. *Social Choice and Welfare* 11, 331–354.

Mongin, P., and d'Aspremont, C. 1998. Utility theory and ethics. In *Handbook of Utility Theory,* Vol. 1: *Principles*, ed. S. Barberà, P. J. Hammond, and C. Seidl. Kluwer Academic, Boston, pp. 371–481.

Roberts, K. W. S. 1980a. Social choice theory: The single-profile and multiprofile approaches. *Review of Economic Studies* 47, 441–450.

Roberts, K. W. S. 1980b. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47, 421–439.

Selinger, S. 1986. Harsanyi's aggregation theorem without selfish preferences. *Theory and Decision* 20, 53–62.

Sen, A. K. 1970. *Collective Choice and Social Welfare*. Holden-Day, San Francisco.

Sen, A. K. 1986. Social choice theory, in *Handbook of Mathematical Economics*, Vol. 3, eds. K. J. Arrow and M. D. Intriligator. North-Holland, Amsterdam, pp. 1073–1181.

Weymark, J. A. 1993. Harsanyi's social aggregtion theorem and the weak Pareto principle. *Social Choice and Welfare* 10, 209–221.

Weymark, J. A. 1994. Harsanyi's social aggregation theorem with alternative Pareto principles. In *Models and Measurement of Welfare and Inequality*, ed. W. Eichhorn. Springer-Verlag, Berlin, pp. 869–887.

8

Preference Aggregation after Harsanyi

Matthias Hild, Richard Jeffrey, and Mathias Risse

8.1 Introduction

Consider a group of people whose preferences satisfy the axioms of one of the current versions of utility theory, such as von Neumann–Morgenstern (1944), Savage (1954), or Bolker (1965) and Jeffrey (1965). There are political and economic contexts in which it is of interest to find ways of aggregating these individual preferences into a group preference ranking. The question then arises of whether methods of aggregation exist in which the group's preferences also satisfy the axioms of the chosen utility theory, while at the same time the aggregation process satisfies certain plausible conditions (e.g., the Pareto conditions introduced later).

The answer to this question is sensitive to details of the chosen utility theory and method of aggregation. Much depends on whether uncertainty, expressed in terms of probabilities, is present in the framework and, if so, on how the probabilities are aggregated. The goal of this chapter is (a) to provide a conceptual map of the field of preference aggregation – with special emphasis, prompted by the occasion, on Harsanyi's aggregation result and its relations to other results – and (b) to present a new problem ("flipping"), which leads to a new impossibility result.

The story begins with some bad news, roughly fifty years old, about "purely ordinal" frameworks, in which probabilities play no role.[1]

Arrow's General Possibility Theorem (1950, 1951, 1963): *No universally applicable nondictatorial method of aggregating individual preferences into group preferences can satisfy both the Pareto Preference condition (unanimous*

[1] Sen (1970), chapter 3, provides an excellent exposition.

Richard Jeffrey died on November 9, 2002.

individual preferences are group preferences) and the condition of Independence of Irrelevant Alternatives (group preference between two prospects depends only on individual preferences between those same prospects).

But for nearly as long we have had some good news about the vN–M (von Neumann–Morgenstern) framework, in which probabilities play an essential role:[2]

Harsanyi's Representation Theorem (1955): *If individual and group preferences all satisfy the vN–M axioms, if ("Pareto Indifference") the group is indifferent whenever all individuals are, and if ("Strong Pareto") group preference agrees with that of an individual whenever no individual has the opposite preference, then group utility is a linear function W of individual utilities.*

Both news items are accurate. Their differences stem from differences in the requirements they place on utility functions that count as representing a given preference ordering. Arrow's framework was purely ordinal in the sense that for a utility function to count as a representation of a preference ordering he only required the numerical ordering of utilities to agree with the given preference ordering of prospects. In the von Neumann–Morgenstern framework, where the agent is assumed to have preferences between lotteries that yield particular outcomes with particular numerical probabilities, there is a second requirement: The place of a lottery in the preference ranking must correspond to the *eu* (the expected utility, the probability-weighted sum) of the utilities of its possible outcomes. In the vN–M framework, utilities of outcomes and *eu*'s of lotteries are uniquely determined by the preference ranking once a zero and a unit have been chosen.

Actual personal probabilities play no part in Harsanyi's aggregation process. Even though individuals may have personal probabilities and use them to solve their own decision problems, the process does not aggregate these into group probabilities; it is only personal utilities for outcomes that are aggregated. These will determine social *eu*'s for chancy prospects in which outcomes are assigned definite numerical probabilities. Harsanyi's result will be our main concern in Section 8.2.

In various other frameworks, for example, Savage's (1954), Bolker's (1965), and Jeffrey's (1965), personal probabilities as well as utilities are deducible from preferences. If both group and individual preferences are

[2] Here and in Section 8.2 we draw on Weymark's (1991) reconstruction of the Harsanyi theorem.

to be placed in these frameworks, we need to decide how to use personal probabilities as well as personal utilities in the aggregation process – a decision that does not arise in the von Neumann–Morgenstern framework. There are two ways to go: ex ante and ex post. (Harsanyi's own method of aggregation falls into neither of these categories because personal probabilities have no place in his vN–M framework.) Both methods of aggregation face serious problems.

In ex ante aggregation (Section 8.3) group *eu* is a function, say, W, of individual *eu*'s. Here the question arises: under what conditions is the aggregate $W(eu_1, \ldots, eu_I)$ of individual *eu*'s itself an *eu*? The answer is bad news for those who hope to use aggregation as a way of arriving at compromises among conflicting judgments of fact or value:[3]

Generic Ex Ante Impossibility Theorem: *In general, ex ante aggregation is possible only for groups that are highly homogeneous in their probability judgments or in their value judgments.*

In ex post aggregation (Section 8.4) individual *eu*'s are first disintegrated into utilities and probabilities. These are then aggregated separately into group utilities and group probabilities, which are finally reintegrated into group *eu*'s. This blocks the difficulty that led to the generic ex ante possibility theorem. Later in Section 8.4 we announce some new bad news for the ex post approach:

Flipping Theorem: *In ex post aggregation, utility and probability profiles for individuals exist relative to which group preference between some pair of options reverses repeatedly or even endlessly as the analysis is refined, although individual preferences remain constant throughout these analyses.*

Finally, Harsanyi's good news is not vitiated by the flipping phenomenon, and we suggest a connection between that fact and a certain sort of individualism.

8.2 Harsanyi's Utilitarianism

In "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility" (1955), Harsanyi challenged Arrow's (1951, p. 9) thesis "that interpersonal comparison of utilities has no meaning and, in fact, that there

[3] Among the bearers of bad tidings have been Broome (1987, 1990), Seidenfeld, Kadane, and Schervish (1989), and Mongin (1995).

is no meaning relevant to welfare comparisons in the measurability of individual utility." Both saw themselves as responding to Bergson's (1938, 1948) challenge "to establish an ordering of social states which is based on indifference maps of individuals."[4] But their responses were radically different, with Arrow reaffirming the ordinalism of the 1930s, and Harsanyi rejecting it in favor of von Neumann and Morgenstern's revived cardinalism, which he applied to social as well as individual preferences.

Needing cardinal utilities for game theory, von Neumann and Morgenstern had turned the tables on ordinalists who had argued that the significance of a numerical utility function for prospects $X, Y, \ldots$ is exhausted by the corresponding order relation ($\succeq$) of preference-or-indifference on those prospects:

$$u(X) \geq u(Y) \text{ iff } X \succeq Y. \tag{8.1}$$

By replacing the old prospects $X, Y, \ldots$ by the set $\mathcal{G}$ of all gambles among them[5] and replacing the utilities $u(X), u(Y), \ldots$ by expected utilities $eu(P), eu(Q), \ldots$ relative to $P, Q, \ldots \in \mathcal{G}$, they obtained a preference relation with definite cardinal significance:

$$eu(P) \geq eu(Q) \text{ iff } P \succeq Q. \tag{8.2}$$

Here $eu(P) = u(X)P(X) + u(Y)P(Y) + \ldots$, and similarly for $eu(Q)$. In the presence of Eq. (8.2), the full set of monotone increasing transformations of u under which Eq. (8.1) is preserved shrinks to its positive affine subset.

It is not eu's (or their ratios or differences) that are invariant, but ratios of differences, ratios of "preference intensities":[6]

$$\frac{eu(P) - eu(Q)}{eu(R) - eu(S)} = \frac{\text{intensity of preference for P over Q}}{\text{intensity of preference for R over S}} \tag{8.3}$$

Harsanyi used Marschak's formulation of the vN–M theory. In Marschak's framework the outcomes $X, Y, \ldots$ of gambles are offstage; it is only members of the set $\mathcal{G}$ that appear on stage. However, each outcome offstage is represented on stage by the member of $\mathcal{G}$ that assigns probability 1 to it and 0 to all the others.

[4] The words are Arrow's (1951, p. 9).

[5] In these gambles, the probabilities of outcomes must be specified explicitly in numerical form, e.g., "Victory with probability .1, defeat with probability .9." The contrast is with specifications in terms of events for which different individuals might have different probabilities, e.g., "Victory if Ruritania joins us, defeat if it does not."

[6] See remark 3 at the end of this section.

Marschak's Postulates:[7] For $P, Q, R, S \in \mathcal{G}$ and $x, \tilde{x} \in [0, 1]$, where $\tilde{x} = 1 - x$:

M_1 $\succeq$ is a complete, transitive relation on $\mathcal{G}$.
M_2 If $P \succ Q \succ R$, then $xP + \tilde{x}R \approx Q$ for some x.
M_3 $P \succ Q \succ R \succ S$ for some P, Q, R, S.
M_4 If $Q \approx R$, then $xP + \tilde{x}Q \approx xP + \tilde{x}R$ for all P, x.

Representation Theorem: *Given $M_1 - M_4$ there exist functions eu satisfying Eq. (8.2): These are unique up to a positive affine transformation.*

InTRApersonal Comparison of Preference Intensities: To compare i's preference intensity for P_1 over P_2 with that for P_3 over P_4, select suitable test-gambles P_{14}, P_{23} from $\mathcal{G}$, that is,

$$P_{14} = \frac{1}{2}P_1 + \frac{1}{2}P_4, \qquad P_{23} = \frac{1}{2}P_2 + \frac{1}{2}P_3, \tag{8.4}$$

and note their relative positions in i's preference ranking. It will turn out that $eu_i(P_1) - eu_i(P_2) \geq eu_i(P_3) - eu_i(P_4)$ iff $P_{14} \succeq_i P_{23}$, for by Eq. (8.2), the three conditions (8.5) are equivalent:

$$\frac{eu_i(P_1) - eu_i(P_2)}{eu_i(P_3) - eu_i(P_4)} \geq 1, \quad \frac{eu_i(\frac{1}{2}P_1 + \frac{1}{2}P_4)}{eu_i(\frac{1}{2}P_2 + \frac{1}{2}P_3)} \geq 1, \quad P_{14} \succeq P_{23} \tag{8.5}$$

In a single episode of group decision making, the group (e.g., perhaps, a legislature) will choose from a small set of pairwise incompatible options (perhaps, bills for combinations of taxation and public expenditure). The set $\mathcal{G}$ of all probability distributions over those options is the common field of the group preference ranking $\succeq_0$ and the individual preference rankings $\succeq_i$ of group options. In Harsanyi's postulates, the number 0 represents a group and the numbers $1, \ldots, I$ represent the individuals who make it up.

Harsanyi's postulates: For $i, j = 1, \ldots, I$ and $P, P_i, Q \in \mathcal{G}$:

H_1 All individuals' rankings $\succeq_i$ satisfy $M_1 - M_4$.
H_2 So does the group's ranking, $\succeq_0$.
H_3 *Functionality:* $P \approx_0 Q$ if $P \approx_i Q$ for all i.

[7] $\succeq$, $\succ$, and $\approx$ are the relations of weak preference, strong preference, and indifference.

H_4 *Uniqueness:* $\exists Q\ \forall i\ \exists P\ \forall j \neq i\ (P \succ_i Q$ but $P \approx_j Q)$.[8]
H_5 *Positivity:* $P \succeq_0 Q$ if $P \succeq_i Q$ for all i and $\succ_i$ for some i.

Harsanyi's Aggregation (= Representation) Theorem: *Postulates* $H_1 - H_5$ *imply the existence of eu's for the preferences* $\succeq_0, \succeq_i$ *that satisfy the condition* $eu_0 = \sum_i eu_i$. *These are unique up to a positive affine transformation.*

For an accessible explanation of the axioms and a proof of a somewhat stronger form of this theorem, see Weymark (1991), section 3.[9]

When is individual i's preference intensity for P_1 over P_2 greater than (or less than, or equal to) individual j's for P_3 over P_4? This is the form questions of interpersonal comparison of utilities take when individual and group preferences determine only ratios of differences of utilities as in Eq. (8.3). These may well be substantive questions, which people do sometimes manage to answer correctly by various devices appropriate to particular persons and their situation.[10] Answers to such questions guide the synthesis of group preferences out of individual ones.

But here we work backward, from a group preference ranking that all find acceptable as an evenhanded aggregation of their various preferences to the interpersonal comparison of individual utility differences that ranking presupposes. Whether or not the individuals have accurately answered the substantive questions, their group ranking can be analyzed to discover what are in effect common judgments, right or wrong, of form "r = the ratio of i's preference intensity for P_1 over P_2 to j's for P_3 over P_4."

The idea is adequately illustrated in the case of a two-person group. Suppose that, somehow or other, individuals 1 and 2 have come to regard a particular preference ranking $\succeq_0$, satisfying H_1–H_5 for the group constituted by the two of them, as an evenhanded aggregation of their individual preference rankings, $\succeq_1$ and $\succeq_2$. Then any function eu_0 representing $\succeq_0$ can be used to determine whether given functions eu_1, eu_2 representing the personal rankings are interval commensurate:

[8] Harsanyi (1955) does not state H_4 as an axiom, but presupposes it in the first sentence of the proof of his Theorem 5. In H_4, P depends on i, but Q does not.

[9] In his treatment, Weymark (1991, p. 272) permutes the first two quantifiers in H_4 to obtain a weaker axiom ("Independent Prospects") in which both P and Q depend on i and which still yields uniqueness.

[10] See Harsanyi (1955, 1990) and Weymark's (1991) counterarguments. See also Jeffrey (1992), chapter 10.

Interval Commensuration Revealed Retrospectively: If H_1–H_5 hold with $I = 2$, then by H_4 there are $P_1, P_2, Q \in \mathcal{G}$ satisfying (a) and (b).

$$\text{(a) } P_1 \succ_1 Q \approx_1 P_2, \qquad \text{(b) } P_2 \succ_2 Q \approx_2 P_1.$$

Representations eu_1, eu_2 of $\succeq_1, \succeq_2$ will be called "interval commensurate" iff some (and, so, every) representation eu_0 of $\succeq_0$ satisfies

$$\frac{eu_1(P_1) - eu_1(Q)}{eu_2(P_2) - eu_2(Q)} = \frac{eu_0(P_1) - eu_0(Q)}{eu_0(P_2) - eu_0(Q)}. \tag{8.6}$$

Given conditions (c) and (d), formula (8.6) follows from conditions (a) and (b):[11]

$$\text{(c) } eu_0(P) = eu_1(P) + eu_2(P),$$

$$\text{(d) } eu_0, eu_1, eu_2 \text{ represent } \succeq_0, \succeq_1, \succeq_2.$$

Differences of form $eu_j(P)$—$eu_j(Q)$ are not uniquely determined by the corresponding relation $\succ_j$, but ratios of such differences *are*, for example, as on the right-hand side of Eq. (8.6).[12] Then in view of Eq. (8.6) the ratio of differences for $j = 1, 2$ (i.e., a ratio of interval commensurate preference intensities) is fixed by certain group preference intensities and thus, in view of Marschak's representation theorem, by the group's preference ranking.[13]

We conclude this section with three remarks:

1. Of course questions of interpersonal comparison are idle if Harsanyi's aggregation theorem is vitiated by an ex ante impossibility theorem, as some would seem to think,[14] but it is not so. On the contrary, Harsanyi's method of utility aggregation is immune to ex ante impossibility theorems simply because, as we have observed, it is neither ex ante nor ex post.

[11] *Proof.* By (a), (b), (d) the denominator on the left of Eq. (8.6) is non-null. Now operate on the right: First apply (c) to the four eu_0 terms; by (a) and (b) we may now substitute $eu_1(Q)$ for $eu_1(P_2)$ and $eu_2(Q)$ for $eu_2(P_1)$; after canceling the $\pm eu_2(Q)$ terms in the numerator and the $\pm eu_1(Q)$ terms in the denominator, Eq. (8.6) becomes an identity.

[12] The social preference ranking determines eu_0 uniquely up to an affine transformation $eu_0 \mapsto a \cdot eu_0 + b$ with $a > 0$, and the value of the right-hand side of Eq. (8.6) is unaffected by any such transformation because we can drop $b - b$ from the numerator and the denominator, after which the a's in the numerator cancel those in the denominator.

[13] By confining this commensuration technique to consecutive pairs $(1, 2), \ldots (I - 1, I)$ of individuals, Harsanyi's aggregation result might be obtained with H_4 weakened to this: $\forall i = 1, \ldots, I - 1, [\exists P \exists Q (P \succ_i Q \text{ but } P \approx_{i+1} Q) \text{ and } \exists P \exists Q (P \succ_{i+1} Q \text{ but } P \approx_i Q)]$.

[14] Broome (1991, pp. 160, 201) *seems* to be saying that Harsanyi's scheme is vitiated in that way, but this impression is created by his broad use of the term "Harsanyi's theorem" not only for Harsany's own aggregation theorem, but for variants of it in which the vN–M framework is replaced by frameworks like those of Savage and Bolker–Jeffrey, in which personal probabilities figure alongside utilities.

2. The object of the vN–M and Marschak axiomatic treatments of preference was to counter the view that game theory's cardinal concept of utility was metaphysical nonsense. Since there were no such qualms about the long-run frequency view of cardinal *probability*, von Neumann and Morgenstern adopted that view in their exposition (1944, 1947, 1953, p. 19):

> Probability has often been visualized as a subjective concept, more or less in the nature of an estimation. Since we propose to use it in constructing an individual, numerical estimation of utility, the above view of probability would not serve our purpose. The simplest procedure is, therefore, to insist upon the alternative, perfectly well founded interpretation of probability as frequency in long runs. This gives directly the necessary numerical foothold.[2]
>
> [2] If one objects to the frequency interpretation of probability then the two concepts (probability and preference) can be axiomatized together. This too leads to a satisfactory numerical concept of utility which will be discussed on another occasion.

But what made Harsanyi adopt the vN–M framework was no commitment to a long-run frequency view of probability; rather, it was his view of probability as (in von Neumann and Morgenstern's words) "a subjective concept, more or less in the nature of an estimation." Harsanyi was that sort of subjectivist well before Savage showed how personal probabilities of events can be recovered from personal *eu*'s (i.e., ultimately, from personal preferences among gambles on those events). From the start, Harsanyi took it for granted that your expectations concerning random variables would be represented by probability–weighted means in which the probabilities are "subjective," representing your own uncertain judgments.[15] He could use the vN–M utility theory without the sorts of qualms mentioned in the unkept promise made in their footnote 2 – a promise that Savage later made good.[16] The vN–M theory provided Harsanyi with a random variable *u* that could be combined with personal probabilities, exogenous to that theory, to yield exogenous personal *eu*'s. It was Ramsey (1931) and Savage (1954) who provided decision theories with endogenous personal probabilities as well as utilities.

3. We *form* our preference ranking of acts under uncertainty by judging the probabilities and utilities of the possible outcomes of those acts as best we can. From this constructive point of view, it is our probability and

[15] In this sense of the term, Carnap (1945, 1950, 1962) was also a subjectivist. Like Carnap, Harsanyi took the legitimate source of the differences between different people's "subjective" probability judgments to be differences in the data on which those judgments are based.

[16] Savage (1954) points out that Ramsey (1931) had made the promise good decades earlier.

utility judgments that determine our *eu*'s, and our *eu*'s that determine our preferences. This way of forming preferences has been tuned up over the past three centuries and more. A high-tech version can be found in Raiffa's 1968 how-to-think book for MBAs. And a low-tech version had the place of honor at the end Arnauld's 1662 how-to- think book for the innumerate:

> To judge what one must do to obtain a good or avoid an evil, it is necessary to consider not only the good and the evil in itself, but also the probability that it happens or does not happen; and to view geometrically the proportion that all these things have together.

Representation theorems are analytical, not constructive: given a fully formed preference ranking that satisfies the axioms, they assure us of the existence of *eu* functions that represent the ranking, and of the uniqueness of those representations up to a positive linear transformation. Of course we do not have fully formed preference rankings over all the prospects that interest us. (If we did, we could simply read the solutions to our decision problems off them.) The problem in decision making is the constructive one of forming or discovering preferences we can live with. From the analytical point of view taken in representation theorems, it is true enough that an *eu* function is a mere representation of a given preference ranking. But from the point of view of decision makers, it is their preference rankings that merely represent their *eu* functions, which in turn merely reflect their probabilities and utilities.

8.3 Aggregation Ex Ante

We now turn to frameworks for preference in which actual personal probabilities play a role, in particular, the Savage framework in this section, and the Bolker–Jeffrey framework in Section 8.4. In the vN–M framework numerical probabilities of lottery outcomes are specified explicitly, and actual personal probabilities play no role. In the new frameworks, personal probabilities play a central role and are recoverable from the given preference ranking if it satisfies the relevant axioms. Here are thumbnail sketches of the two frameworks:

Savage: Preference is a relation between "acts." Acts are represented by functions f, each of which assigns to each possible "state of nature" s a definite "consequence" $f(s)$. If the act is betting \$10 on Bluebell to win, then we have

$$f(s) = \text{“be \$10 richer” if Bluebell wins in state } s,$$

$$f(s) = \text{“be \$10 poorer” if Bluebell does not win in state } s.$$

The expected utility $eu(f)$ of an act f is the mean value of $u(f(s))$ for all states of nature s, weighted with the individual's personal probability distribution P over the states of nature. Savage's representation theorem guarantees the existence of functions u and P, which together represent the preference ranking in the sense that act f is preferred to act g if and only if $eu(f)$ is greater than $eu(g)$.

Bolker–Jeffrey: Here preference is a relation between "events" A (i.e., between the same things to which probabilities are attributed), and utilities $u(s)$ are attributed to states of nature s. Performing an act is a matter of making some particular event true, for example, the event of betting \$10 on Bluebell to win. Given a utility function u and a probability function P, the "desirability" $des(A)$ of an event A is defined as the mean value of $u(s)$ for all states of nature s, weighted with the conditional probability distribution $P(-|A)$. According to Bolker's representation theorem, truth of event A is preferred to truth of event B if and only if the desirability of A is greater than that of B.[17]

Desirability can be defined as conditional expectation of utility,[18] $des(A) = E(u|A) = \int_A u\, dP(-|A)$. In the discrete case, where the set S of states of nature is finite or countably infinite, the integral becomes a sum:

$$des(A) = \sum_{s \in A} u(s) P(\{s\}|A). \tag{8.7}$$

Example: "Dessert?" Consider Alice's problem of deciding whether to say "Yes" or "No" in answer to this question. She is sure that dessert would turn out to be chocolate ice cream (c), vanilla ice cream (v), or pie (p), that is, Dessert $= \{c, v, p\}$, but she does not know which.

Data: For these possibilities, her probabilities conditionally on Dessert are $P_{\text{Alice}}(\{c\}|\{c, v, p\}) = P_{\text{Alice}}(\{v\}|\{c, v, p\}) = \frac{1}{8}$ and $P_{\text{Alice}}(\{p\}|\{c, v, p\}) =$

[17] For accessible overviews of the theory, see Bolker (1967), Jeffrey (1983), and Broome (1990). For important modifications of the theory, see Joyce (1992) and Bradley (1997).

[18] Bolker's (1965, 1966, 1967) representation theorem guarantees the existence of a function *des* representing preference between elements of a Boolean algebra—but on assumptions under which the algebra cannot be a field of sets (of "states"). Under those assumptions, the function *des* is not the conditional expectation of any function $u(s)$. But of course existence of such a representation when those assumptions hold does not imply nonexistence when they do not. Jeffrey (1992, chapter 15) recasts Bolker's theorem in a form applicable to Boolean algebras of sets of states—algebras on which $des(A)$ can be defined as $E(u|A)$ after all. (The gimmick is like the one Kolmogorov [1948, 1995] uses to transform fields of sets on which probability measures exist into Boolean algebras of the sort postulated in Bolker's theorem.)

$\frac{3}{4}$, and her utilities are $u_{\text{Alice}}(c) = 68$, $u_{\text{Alice}}(v) = -100$, $u_{\text{Alice}}(p) = 16$. For the remaining possibility, None (n), her utility is $u_{\text{Alice}}(n) = 0$.

Solution: As Alice sees it, the states of nature form the set $S = \{c, v, p, n\}$ and the event Dessert has desirability $des_{\text{Alice}}(\{c, v, p\}) =$

$$\textstyle\sum_{s\in\{c,v,p\}} u_{\text{Alice}}(s) P_{\text{Alice}}(\{s\}|\{c, v, p\}) = 68(\frac{1}{8}) - 100(\frac{1}{8}) + 16(\frac{3}{4}) = 8.$$

Then, since $des_{\text{Alice}}(\{n\}) = u_{\text{Alice}}(n) = 0 < 18$, Alice does want dessert: $\{c, v, p\} \succ_{\text{Alice}} \{n\}$. Similar calculations show that she prefers pie to ice cream: $des_{\text{Alice}}(\{p\}) = u_{\text{Alice}}(p) = 16 > des_{\text{Alice}}(\{c, v\}) = -16$. Note that until she makes her decision, Alice's probability for dessert will be strictly between 0 and 1, for example, P_{Alice}(Dessert) might be 1/2, 7/10, or whatever. But the actual value makes no difference to her decision because the probabilities of interest are all conditional on Dessert, and we suppose (see Jeffrey 1996) that those remain constant as the unconditional probability of Dessert varies.

Where Savage assigns probabilities to events independently of what act is being performed, Bolker and Jeffrey assign conditional probabilities to events given acts. (Because acts are not events for Savage, these conditional probabilities make no sense for him.) The Bolker–Jeffrey framework allows probabilities to be updated either by observation or by decision: the updated unconditional probability will be the prior conditional probability given the event observed or chosen. But in the Savage framework choice of an option cannot affect probabilities. Note, too, that Savage's treatment is problematical in cases where it is important to consider players' probabilities for other players' performing various acts, as in interactive decision theory (= game theory).

Two Dismal Possibility Theorems. Here we note two specifications of the generic ex ante possibility theorem indicated in Section 8.1. The species is Mongin's (1995) modification of the Savage framework, a modification in which an additional postulate assures σ-additivity of the probability measure.

Let $\succeq_i$, u_i, and P_i be individual i's preference relation, utility function, and probability function. Mongin adopts analogues of Harsanyi's Pareto conditions H_3 (functionality) and H_5 (positivity). To give these postulates material to work on he adds an assumption of diversity (linear independence) of the various individuals' probabilities or utilities. Either assumption implies the following condition, which is an analogue of H_4:

Independence: Each individual i has some preference $f \succ_i g$ where all others are indifferent: $f \succ_i g$, but $f \approx_j g$ if $j \neq i$.

Finally, Mongin postulates a minimal *Agreement* condition:

Agreement: There exist consequences c_1, c_0 such that all individuals i assign higher utility to the former: $u_i(c_1) > u_i(c_0)$.

Mongin uses the term *overall dictator* for an individual whose probabilities and preference intensities are the same as society's. Of course, such individuals need not really be dictators; for example, they might be immensely public-spirited citizens, or ones whose personal attitudes are somehow formed by the same causes as the group's; or the "dictator" might be chosen by lot or by vote; or the coincidence might be the result of blind chance. As Hylland and Zeckhauser (1979) point out, real dictatorship would be a property of the preference aggregation scheme, W; that is, the property of assigning a particular individual's preferences to society regardless of what probabilities and utilities the others may have. But anyway it would be a very restrictive possibility theorem that implied the existence of Mongin's "dictators."

Our Theorems Mgn_1 and Mgn_2 are weaker consequences of Mongin's main possibility results.[19] Before stating these theorems, we introduce some further terminology. *Positivity* is the analog of H_5 (i.e., the group prefers f to g if some member does and none prefer g to f), and "Functionality" is the analogue of H_3. In Mgn_2, we use the terms *diverse* and *clone* as follows:

Probability Clones: Individuals with identical probability functions.

Utility Clones: Individuals with affine equivalent utility functions.

Diversity: The individuals' probability functions are all distinct and none are weighted averages of the others.

Theorem Mgn$_1$: *In the modified Savage framework with functionality and positivity there will be an overall* dictator *if no individual probability or utility function is a linear combination of others.*

Theorem Mgn$_2$: *In the modified Savage framework, Positivity and Agreement together imply (1) and (2):*

1. *If the probability functions are diverse, all are utility clones.*
2. *If not all are probability clones, some are utility clones.*

[19] See Mongin's (1995) observation 1 on p. 341, and Proposition 7 on pp. 343–344.

Politics makes strange bedfellows. Results like Mgn_1 and Mgn_2 may seem less disturbing – only to be expected – in the light of the well-known fact that unanimity about the relative ranking of two options may be based on quite incompatible assessments of probability or utility. Raiffa (1968, p. 230) offers a simple, striking example, with two options (a_1, a_2), two states of nature (θ_1, θ_2), and a pair of experts, Alice and Bob, who are indifferent between the options for very different reasons: Alice assigns probabilities .2, .8 to θ_1, θ_2 and utilities 1, 0, .5, 1 to $a_1\theta_1$, $a_2\theta_1$, $a_1\theta_2$, $a_2\theta_2$, while Bob assigns probabilities .8, .2 and utilities .5, 1, 1, 0 to the same states and act-state pairs. These experts have the same expected utilities (.6 for a_1, .8 for a_2) but for precisely opposite reasons. As Raiffa argues, such examples cast doubt on the seemingly ineluctable functionality principle, H_3. This idea is pushed further in the next section, under "flipping."

8.4 Aggregation Ex Post

The strange bedfellows phenomenon may be seen as a warning against muddling judgments of fact and value, and as a call to take the ex post stance, in which members' *eu*'s are not directly aggregated, but are first analyzed into probabilities and utilities, which are aggregated separately into group probabilities and group utilities, and only then recombined into group expected utilities.

This stance, with its rationale, was forcefully enunciated by Raiffa (1968) thirty years ago in sections 12 and 13 of his classical text *Decision Analysis*, for example, on "The Problem of the Panel of Experts" (pp. 232–233):

> If I were solely responsible as the decision maker, I should want to probe the opinions of my experts to assess my own utility and probability structure. I should try to keep my assessments for utilities separate from my assessments for probabilities, and I should try to exploit such common agreements as independence.[20] Wherever possible, I should want to decompose issues to get at basic sources of agreement and disagreement. I should compromise at the primitive levels of disagreement and adopt points of common agreement as my own, so long as these common agreements were not compensating aggregates of disagreements. I should do so knowing full well that I might end up choosing an action which my experts would say is not as good as an available alternative. Throughout this discussion, of course, I am assuming that I do not have to worry about the viability of my organization, its morale, and so on.

[20] Convex combinations of independently and identically distributed (i.i.d.) distributions are not generally i.i.d., so averaging such distributions would not be a way of preserving common agreement on independence. (To preserve independence one could form the average of the individuals' i.i.d. distributions and use that as the one–shot probability of an i.i.d. group distribution.)

There, too, he reports a result of Zeckhauser's that would be published eleven years later (Hylland and Zeckhauser, 1979) in a somewhat different version:

> Richard Zeckhauser has proved a mathematical theorem that states this result:
>
> "No matter what procedure you use for combining the utility functions and for combining the probability functions, so long as you keep these separate and do not single out one individual to dictate the group utility and probability assignments, then you can concoct an example in which your experts agree on which act to choose but in which you are led to a different conclusion." (Raiffa, 1968, p. 230)

Raiffa (1968, pp. 233–237) explores the tension this theorem reveals between the following two conditions.

Pareto Optimality: The group prefers one prospect to another if some members do and none have the opposite preference.

Reification: "The group members should consider themselves as constituting a panel of experts who advise the organizational entity: they should imagine the existence of a higher decision-making unit, the organization incarnate, so to speak, and ask what *it* should do. Just as it made sense to give up Pareto optimality in the problem of the panel of experts, it likewise seems to make sense in the group decision problem." (Raiffa, 1968, pp. 233–234)

We now introduce a new problem for ex post aggregation:[21]

Flipping Theorem: *In ex post aggregation, utility and probability profiles for individuals exist relative to which group preference between some pair of options reverses endlessly as the analysis is refined, even though all individual preferences remain unchanged.*

Note how the flipping theorem relates to the result of Hylland and Zeckhauser. They use the ex ante Pareto condition in an ex post framework, that is, a framework in which probabilities and utilities are aggregated separately. We use the ex post Pareto condition (i.e., on utilities, not expected utilities) in an ex post framework. Thus flipping is a problem inherent in the ex post

[21] Here we illustrate the problem for a particular aggregation rule, that is, straightforward averaging of probabilities and summing of utilities. But the problem can arise for any ex post Pareto optimal aggregation rule.

approach: Unlike the Hylland–Zeckhauser (1979) result, it does not depend on the tension between ex ante standards and ex post aggregation.[22]

The flipping phenomenon is illustrated by the following example, which we formulate here in the Bolker–Jeffrey framework sketched in Section 8.3.[23] In the example, initial group desirabilities 12, 0 of two options (Dessert, None) change to −8, 0 on closer examination of the first option and change back to 12, 0 on still closer examination. The group desirabilities can flip because the individuals have opposed probabilities and differently opposed utilities, somewhat as in the "politics makes strange bedfellows" example, but here with the opposed tendencies overbalancing in opposite directions at each stage of refinement.

Dessert Makes Strange Bedfellows: Alice and Bob are being given a dinner for two in which they must make the same choice from the menu, course by course. Having agreed on all courses so far, they are trying to decide whether to have dessert, which the menu lists with no details. Suppose that in fact they both prefer the event Dessert to the event None, and that on personal desirability scales des_{Alice}, des_{Bob}, which they regard as interpersonally commensurate, their desirabilities for Dessert are 8 and 4 as shown in Figure 8.1 (level 0, above the line), and their desirabilities for None (not shown in Figure 8.1) are both 0. Suppose they are sure that dessert will turn out to be Ice cream or Pie, concerning which their respective commensurate desirabilities are −16, 16 for Alice, and 16, −32 for Bob, as shown above the line at level 1 of Figure 8.1. Suppose that $P_{\text{Alice}}(\text{Ice}|\text{Dessert}) = 1/4$, $P_{\text{Alice}}(\text{Pie}|\text{Dessert}) = 3/4$, and that the values for P_{Bob} are just the reverse. These conditional probabilities are represented by the areas of the respective compartments, on a scale where the whole square has area 1. Since $des(A) = E(u|A)$, the desirability of the union of two events that are judged to be incompatible is a weighted average of their separate desirabilities: *If* $P(A \cap B) = 0$, then

$$des(A \cup B) = des(A)P(A|A \cup B) + des(B)P(B|A \cup B) \qquad (8.8)$$

It is easy to verify that with Dessert = Ice ∪ Pie = $A \cup B$ this equation, applied to Alice's and Bob's level 1 desirabilities and probabilities, yields their level 0 desirabilities, 8 and 4. And similarly, if both are convinced that

[22] See Hylland and Zeckhauser (1979, pp. 1325–1326). Their axioms 2 and 3 stipulate ex post aggregation of individual probabilities p^k and utilities u^k. Their axiom 5 is a weak ex ante Pareto optimality condition: "If $E(a_m|p^k, u^k) > E(a_i|p^k, u^k)$ for all k, then a_i is not an element of the choice set."

[23] That is, the simplest framework for the purpose.

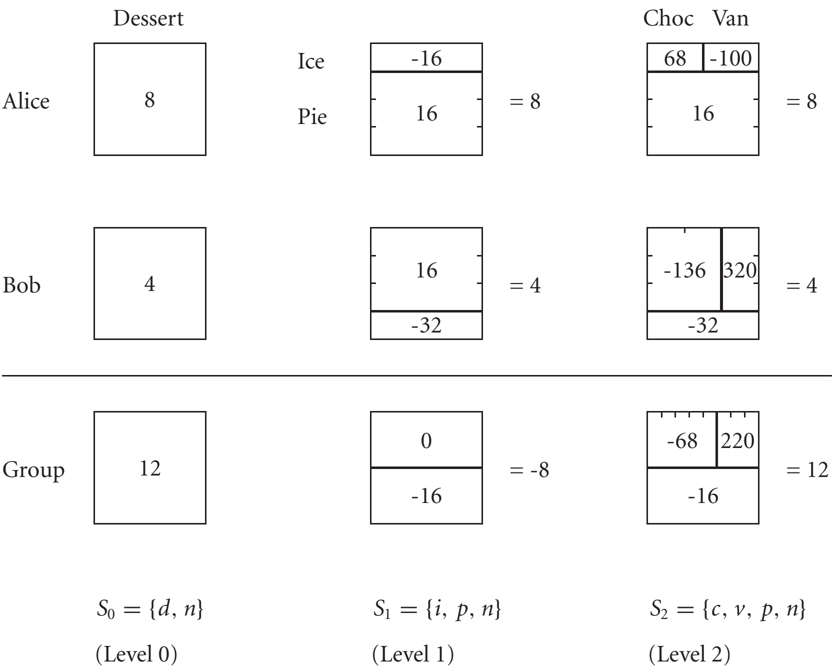


Figure 8.1. Flipping illustrated by refinements of the "Dessert" option.

Ice would turn out to be Choc(olate) or Van(illa), Eq. (8.8) delivers their ±16 level 1 desirabilities for Ice when their probabilities and desirabilities for Choc and Van are as shown at level 2. Then above the line, the three levels of analysis of Alice's attitudes depicted in Figure 8.1 are mutually consistent, as are the three levels of Bob's.

But ex post aggregation of Alice's and Bob's desirabilities by applying the following formula to the numbers shown in Figure 8.1 yields mutually inconsistent group desirabilities, for the results, shown below the line, exhibit the flipping phenomenon: group desirabilities for Dessert flip from 12 to -8 and back again as the aggregation process is applied to finer analyses of the individuals' probabilities and desirabilities.

$$des_{\text{Group}}(A) = des_{\text{Alice}}(A) + des_{\text{Bob}}(A). \tag{8.9}$$

It would be straightforward to devise probabilities and utilities for a further stage (say, with Pie = Apple ∪ Banana) at which group desirability flips back from 12 at stage 2 to -8 at a new stage 3; and one can give an algorithm for continuing the refinements of consistent individual probabilities and utilities so as to carry the 12, −8, 12, −8, ... flipping process as far as you like, even, endlessly.

The flipping problem has another aspect, that is, inconsistency of group probabilities and desirabilities with formula (8.8) when group probabilities conditionally on an act-event D (e.g., the event that we have dessert) are obtained by averaging:

$$P_{\text{Group}}(A|D) = \frac{1}{2}P_{\text{Alice}}(A|D) + \frac{1}{2}P_{\text{Bob}}(A|D). \tag{8.10}$$

Thus, the desirability of Ice at level 1, obtained via Eq. (8.9) as the simple sum of Alice's and Bob's level 1 desirabilities for Ice, is inconsistent with the value obtained via Eq. (8.8) as the probability-weighted average of the group's level 2 desirabilities for Choc and Van:

$$des_{\text{Group}}(\text{Ice}) = -16 + 16 = 0 \text{ from (8.9)},$$

$$des_{\text{Group}}(\text{Ice}) = \frac{5}{8}(-68) + \frac{3}{8}(220) = 40 \text{ from (8.8)}.$$

But is formula (8.9) a correct description of ex post aggregation? By definition, ex post aggregation adds *utilities*, not desirabilities, so that in genuine ex post aggregation, formula (8.9) would be replaced by the corresponding formula for utilities:

$$u_{\text{Group}}(s) = u_{\text{Alice}}(s) + u_{\text{Bob}}(s). \tag{8.11}$$

Can the effect of applying formula (8.11) be the same as that of applying formula (8.9) to the desirabilities of the smallest compartments in Figure 8.1? The answer is "Yes" if we represent the refinement process as applying primarily to the set S of states of nature, and only derivatively to the events, the subsets of S. Thus, at level 0 there are just two states of nature, the state d in which Alice and Bob have dessert, and the state n in which they have none: at level 0 the set of states of nature is $S_0 = \{d, n\}$ as indicated in Figure 8.1. The set S_1 of states at level 1 is obtained by replacing d by two states: a state i in which the waiter brings ice cream, and a state p in which he brings pie. And similarly S_2 comes from S_1 by replacing i by c (he brings chocolate ice cream) and v (he brings vanilla).

Here we have three Boolean algebras $\mathcal{A}_k$ of subsets of S_k, with $k = 0, 1, 2$. The algebra $\mathcal{A}_k$ contains $2^{(2^{k+1})}$ events, for example, $\mathcal{A}_0 = \{\emptyset, \{d\}, \{n\}, S_0\}$. In these, Dessert is represented by three different sets: by $\{d\}$ at level 0, by $\{i, p\}$ at level 1, and by $\{c, v, p\}$ at level 2. We shall say that these three are "associated" with each other, in order to indicate that they are all representations of what is informally seen as one and the same event, Dessert. In general, any $A \in \mathcal{A}_k$ for $k = 0, 1$ is associated with an $A' \in \mathcal{A}_{k+1}$ defined as follows, where $\{s\} = S_k - S_{k+1}$ and $\{s', s''\} = S_{k+1} - S_k$:[24]

$$A' = (A - \{s\}) \cup \{s', s''\} \text{ if } s \in A, \text{ else } A' = A. \tag{8.12}$$

As an ideal beyond human powers of attainment, one could think of continuing this process of refinement endlessly, specifying not only the ways in which Dessert and None might turn out but also possibilities about other things one might care about, for example, the weather tomorrow (and tomorrow, and tomorrow, etc.), various people's states of health, and births, deaths, wars, football scores—whatever. The *ultimate* states of nature are the maximal consistent sets of such specifications. From this idealized point of view, the elements of S_k for finite k will be pseudostates, events (sets of ultimate states) masquerading as states.

Where s ranges over ultimate states, aggregation via Eqs. (8.10) and (8.11) is immune to the flipping phenomenon illustrated by Figure 8.1, for example, because the putative utilities $u_{\text{Alice}}(p) = 16,\ u_{\text{Bob}}(p) = -32$ at level 1 must really be seen as desirabilities $des_{\text{Alice}}(\text{Pie}) = 16,\ \text{des}_{\text{Bob}}(\text{Pie}) = -32$ of an event Pie; and formula (8.11) is no warrant for summing desirabilities. But application of formula (8.11) to utilities of ultimate states is beyond human powers: this way out "in principle" leaves ex post aggregation impossible in practice. One way or the other, ex post aggregation looks like a pipe dream.

If the ex post approch is ruled out in this way, the ex ante approach has its own severe difficulties. In particular, the ex ante possibility theorems rule out any version of liberalism that satisfies the following two conditions.

1. Unanimous individual preferences are preserved as group preferences.
2. Diversity is tolerated as part of political reality, or even cherished, as in Mill's *On Liberty*. (By excluding all linear independence of probability measures and of utility functions, the ex ante possibility theorems exclude such diversity.) Liberalism that mets these two conditions violates Bayesian rationality of individuals or the group: It requires irrational people or an irrational society.

In closing, we recall that flipping does not arise in Harsanyi's aggregation scheme, for the vN–M or Marschak framework attributes no judgmental probabilities to groups or to individuals.[25] From a certain individualistic

[24] That is, s is the element of S_k that is split into two elements s', s'' to produce S_{k+1}.

[25] It does arise in other schemes that are neither ex ante nor ex post, for example, that of Levi (1997, chapter 9), and the pseudo–ex post scheme illustrated in Figure 8.1.

point of view, this opportunity to deny that groups have beliefs (i.e., judgmental probabilities) is most welcome. On that view, we may perhaps speak of groups as agents and even as having aggregate preferences, but on that view, groups are not the sorts of things to which beliefs are to be attributed, and so groups are not to be thought of as rational or irrational.

References

Arnauld, A. 1662. *La logique, ou l'art de penser*. Paris. Trans. (1964), *The Art of Thinking*, Bobbs–Merrill, Indianapolis.

Arrow, K. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy* 58, 328–346; reprinted in Arrow and Scitovsky (1969).

Arrow, K. 1951, 1963. *Social Choice and Individual Values.* Wiley, New York.

Arrow, K., and Scitovsky, T., eds. 1969. *Readings in Welfare Economics.* Allen and Unwin, London.

Bergson, A. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–334; reprinted in Arrow and Scitovsky (1969).

Bergson, A. 1948. Socialist economics. In *A Survey of Contemporary Welfare Economics*, ed. H. S. Ellis. Blakiston, Philadelphia, pp. 412–448.

Bolker, E. 1965. Functions resembling quotients of measures. Ph.D. diss., Harvard University.

Bolker, E. 1966. Functions resembling quotients of measures. *Transactions of the American Mathematical Society* 124, 293–312.

Bolker, E. 1967. A simultaneous axiomatization of subjective probability and utility. *Philosophy of Science* 34, 333–340.

Bradley, R. 1997. The representation of beliefs and desires within decision theory. Ph.d. diss., University of Chicago.

Broome, J. 1987. Utilitarianism and expected utility. *Journal of Philosophy* 84, 402–422.

Broome, J. 1990. Bolker–Jeffrey expected utility theory and axiomatic utilitarianism. *Review of Economic Studies* 57, 477–503.

Broome, J. 1991. *Weighing Goods.* Basil Blackwell, Oxford.

Carnap, R. 1945. On inductive logic. *Philosophy of Science* 12, 72–97.

Carnap, R. 1950, 1962. *Logical Foundations of Probability.* University of Chicago Press, Chicago.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321; reprinted in Arrow and Scitovsky (1969).

Harsanyi, J. 1990. Interpersonal utility comparisons. In *Utility and Probability*, ed. J. Eatwell, M. Milgate, and P. Newman. Norton, New York.

Hylland, A., and Zeckhauser R. 1979. The impossibility of Bayesian group decision making with separate aggregation of beliefs and values. *Econometrica* 47, 1321–1336.

Jeffrey, R. 1965, 1983. *The Logic of Decision*, 1st ed. McGraw-Hill, New York; 2nd ed. University of Chicago Press, Chicago.

Jeffrey, R. 1992. *Probability and the Art of Judgment.* Cambridge University Press, Cambridge.

Jeffrey, R. 1996. Decision kinematics. In *The Rational Foundations of Economic Behaviour*, ed. K. J. Arrow, E. Colombatto, and M. Perlman. Macmillan, London; St. Martin's, New York.

Joyce, J. M. 1992. *The Foundations of Causal Decision Theory*. Ph.d. diss., University of Michigan.

Kolmogorov, A. N. 1948, 1995. Algèbres de Boole métriques complètes. In *IV Zjadz Matemayików Poslkich*, Warsaw (1948), pp. 21–30. Trans. Complete metric Boolean algebras, *Philosophical Studies* 77 (1995), 57–66.

Levi, I. 1997. *The Covenant of Reason.* Cambridge University Press, Cambridge.

Marschak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18, 111–141.

Mongin, P. 1995. Consistent Bayesian aggregation. *Journal of Economic Theory* 66, 313–351.

Raiffa, H. 1968. *Decision Analysis.* Addison-Wesley, Reading, MA.

Ramsey, F. P. 1931. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*, ed. R. B. Braithwaite. Kegan Paul, London. Reprinted in Ramsey (1990).

Ramsey, F. P. 1990. *Philosophical Papers.* Ed. D. H. Mellor. Cambridge University Press, Cambridge.

Savage, L. J. 1954. *The Foundations of Statistics.* Wiley, New York.

Seidenfeld, T., Kadane, J. B., and Schervish, M. J. 1989. On the shared preferences of two Bayesian decision makers. *Journal of Philosophy* 86, 225–244.

Sen, A. 1970. *Collective Choice and Social Welfare.* Holden-Day, San Francisco.

von Neumann, J., and Morgenstern, O. (1944, 1947, 1953). *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well–Being*, ed. J. Elster and J. Roemer. Cambridge University Press, Cambridge, pp. 255–320.

PART THREE

GOODNESS AND WELL-BEING

9

Can There Be a Preference-Based Utilitarianism?

John Broome

9.1 Introduction

John Harsanyi has made several fundamental contributions to utilitarian thinking; they are so well known that I do not need to set them out here. It was natural for him, as an economist, to present his utilitarian arguments in terms of preferences. His great influence has been a major factor in diverting the mainstream of utilitarian thinking toward a preference-based – I shall call it *preferencist* – version of utilitarianism. Preferencism is the view that good – what is good for a person and what is good overall – is determined entirely by people's preferences. However, Harsanyi himself brings into his arguments elements that are not preferencist, and I think that was inevitable. Preferences may partly determine good, but other things must enter too.

To an extent, this is obvious. If good is determined by preferences, we have to ask what determines *how* it is determined by preferences. If good is a function of preferences, what determines the functional form? Perhaps the functional form might itself be determined by preferences, but then what determines the way that happens? At some level, something other than preferences must come into the determination. In this chapter, I shall investigate what extra besides preferences is required to produce a coherent version of utilitarianism. How preferencist can utilitarianism be? It does no great harm to preferencism if nonpreferencist considerations of some sort have to be brought in from elsewhere. But it would be seriously damaging if we had to import substantive claims that make good depend on something other than preferences. Claims like these would actually conflict with preferencism.

Many of us believe preferencism is false anyway. It is often argued that we have other moral aims besides satisfying preferences. Perhaps, indeed,

I am grateful to John Skorupski, the editors of this volume, and a referee for useful comments.

satisfying preferences should not in itself be a moral aim at all. I am sure many of these arguments are sound, but I shall not use them in this chapter. They are unlikely to convince a preferencist utilitarian, because utilitarians, in general, and preferencist utilitarians in particular, are usually reformist. If we have other moral aims besides satisfying preferences, they may well think we should change our moral aims. For the same reason, I shall not rely on our intuitive grasp of what is good, or of what is good for a person. In any case, I doubt we have an intuitive grasp that is adequate for the purposes of utilitarianism. Utilitarianism requires good to be quantitative in a particular sense I shall specify more exactly later. It is not enough for utilitarianism that things should be ordered by their goodness, so we have concepts of better and worse. We also need a concept of how much better one thing is than another. I doubt we have a clear intuitive concept of good that is quantitative in this sense. This is something that may be up for definition; a preferencist utilitarian might plausibly claim to be defining a quantitative concept of good. So instead of relying on intuition, I am going to argue on formal grounds. This will be an internal investigation of preferencist utilitarianism, testing its internal coherence. It will be asking whether preferencist utilitarianism is possible, not merely whether it is true.

Whatever the results, they will not put the value of Harsanyi's work in doubt. Harsanyi's formal arguments are very original and very important. But I think they should be cut free from their preferencist assumptions. They are more successful when reinterpreted in nonpreferencist terms. Most of my book *Weighing Goods* is an attempt to give them a more secure interpretation. That is a sign of the value I attach to them. I think we should let the preferencism go, and keep the formal arguments.

9.2 Uncertainty

Utilitarianism contains a theory of good and a theory of right. It is characteristic of the utilitarian theory of right that rightness is derived from goodness; how one should act is determined entirely by the goodness of things. The theory of good tells us how good things are, and the theory of right tells us how to act on the basis of how good things are. This chapter is about good and not right. But I need to say a little about the utilitarian theory of right by way of introduction.

For simplicity, I shall mention only the act-utilitarian version. The simplest act utilitarianism says that, when choosing between acts, you should choose the one that will produce the best results.[1] However, this principle is in practice useless in our uncertain world. We never know certainly what results will be produced by any of our acts. So, at the time we have to act, we can never know which act we ought to do according to this principle. In order to know how to act, we need a practical way of dealing with uncertainty.

Uncertainty can be handled within either a theory of right or a theory of good. Within the theory of right, utilitarians sometimes offer this principle: when choosing between acts, one should choose the one that gives the greatest expectation of good.[2] Daniel Bernoulli appears to have assumed this,[3] and it is a version of what I call *Bernoulli's hypothesis*. It is implausible, at least on the face of it, because it implies one should be neutral about risk to good. The act that produces the greatest expectation of good may be more risky than other options: The variance in the amount of good it leads to may be higher than for other options. If so, perhaps one should choose a safer act that gives a lower expectation of good. We should not take Bernoulli's hypothesis for granted, then. But once we give it up, it is not easy to produce a sufficiently general principle within the theory of right to handle uncertainty convincingly.

For that reason, I think uncertainty is better handled within the theory of good.[4] As a principle of right, I think utilitarians should say that, when choosing between acts, one should choose the one that will lead to the best *prospect*. Then, within their theory of good, they should have an account of the goodness of prospects. A prospect is a portfolio of possible *outcomes*, each of which might come about. The goodness of a prospect will depend on the goodness of its possible outcomes. Bernoulli's hypothesis implies specifically that it is the expected goodness of its possible outcomes. But there is room within the theory of good for a more general account of the goodness of prospects.

I wish to define outcomes in a way that excludes all uncertainty; uncertainty belongs to prospects only. This means that outcomes will have to be complete histories for the world. The description of a history will be an infinitely long conjunction. In practice, then, we shall never know what the outcome of an act has been till history has come to an end. I shall call outcomes *histories*, as a reminder of what they are. We can think of a history as a degenerate prospect: the prospect in which this history certainly occurs.

[1] This is G. E. Moore's version. See particularly his *Ethics*, pp. 99–101.

[2] See, for instance, Derek Parfit, *Reasons and Persons*, p. 25.

[3] See his "Specimen theoriae novae de mensura sortis."

[4] This argument is more fully spelt out in my *Weighing Goods*, section 6.1.

9.3 Additivity

To keep things simple, I am going to ignore problems that involve changes in the world's population. Given an unchanging population, one central feature of the utilitarian theory of good is that good is added across people. Utilitarians are committed to at least this:

Additive Principle for Histories: One history is better than another if and only if the total of people's good is greater in the first than in the second.

Since utilitarians need to determine when one *prospect* is better than another, they will certainly need more than this. But in this chapter I shall not need to call on any stronger additive principle.

The additive principle is about aggregating together the good of different people. Next, utilitarianism needs a theory of what determines the good of the people individually. Preferencism is such a theory; I shall come to it soon. But first I must mention an attempt to derive additivity itself from preferencism. The additive principle is part of the function through which, according to preferencist utilitarians, preferences determine overall good. Unless it can be derived from preferencism, it is a nonpreferencist element within the utilitarian story. So we need to check whether the derivation can really be done.

Harsanyi tried to derive additivity from preferencism in his article "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." His argument is founded on a mathematical proof. The conclusion of the proof is certainly a sort of additivity, though it is open to question whether it is precisely the additivity of good that is set out in the additive principle. However, I am not going to pursue this question, because there is a more definitive way to refute the argument. The premises of the proof are mutually inconsistent, so they cannot all be true. Therefore, the conclusion is unsound.

There are three premises. First the Pareto principle:

Pareto Principle for Prospects: If everyone is indifferent between two prospects, these prospects are equally good. If someone prefers one prospect to another and no one prefers the other to the one, then the one is better than the other.

This principle expresses preferencism in a pure form: It says that good depends on people's preferences, and that is all it says. Harsanyi's second premise is that the relation "better than," which appears in the Pareto principle, conforms to expected utility theory. (That is to say, this relation satisfies the axioms of expected utility theory. Expected utility theory is normally formulated as a theory of preferences, but as a formal theory, it can be applied to other two-place relations besides preferences, including the relation of betterness.) The third is that each individual's preferences also conform to expected utility theory.

These premises are mutually inconsistent because of an empirical fact that Harsanyi ignores: People do not all agree about the probabilities of every event. Some events, such as a coin's falling heads on a particular occasion, may have objective probabilities. Harsanyi's proof of his theorem assumes that all events are like that, and furthermore that everyone knows what their objective probabilities are. This was implicit when he adopted von Neumann and Morgenstern's version of expected utility theory[5] in proving the theorem; this version assumes all probabilities are objective. But in real life, many events have no objective probability; for instance, there is no objective probability that Scotland will leave the United Kingdom. Even rational, well-informed people may assign different probabilities to events like these. It turns out that when people disagree about probabilities, Harsanyi's three premises cannot all be true.[6]

At least one of them has to go, therefore. Which should it be? Perhaps more than one. But for reasons I shall not go into here,[7] the Pareto principle definitely has to be abandoned. This is not in itself much of a blow to preferencism, because this weaker version of the Pareto principle is not compromised by the objection I have given:

Pareto Principle for Histories: If everyone is indifferent between two histories, these histories are equally good. If someone prefers one history to another and no one prefers the other to the one, then the one is better than the other.

This forms the basis of the so-called "ex post" school of welfare economics,[8] and it is a solidly preferencist principle.

A bigger loss to preferencism is that the additive principle will have to come from elsewhere. It cannot itself be derived from preferencism as Harsanyi hoped. If a preferencist is to be utilitarian, then the aggregative principle of utilitarianism will have to come from some other source besides preferencism. This need not be a deep blow to preferencism, for two reasons. First,

[5] See von Neumann and Morgenstern, *Theory of Games and Economic Behavior*, chapter 1.

[6] This fact has been formally proved many times. See, most recently, Philippe Mongin, "Consistent Bayesian Aggregation."

[7] See *Weighing Goods*, chapter 7.

[8] See the discussion in Hild, Jeffrey, and Risse, "Preference Aggregation after Harsanyi."

additivity may be derivable by Harsanyi's own methods, if they are suitably reinterpreted. My *Weighing Goods* develops this idea.[9] The reinterpretation could preserve important elements of preferencism, such as the Pareto principle for histories. Second, preferencism could anyhow live happily with an independently derived additive principle. The additive principle is about aggregating the good of different people, whereas preferencism is most fundamentally about the good of individual people. So the two may be coexist independently.

9.4 Preferencism as an Account of Individual Good

From now on, therefore, I shall concentrate on preferencism as an account of individual good. It is one of several competing accounts that exist within the body of utilitarian thinking. It says:

Preferencist Biconditional: One history is better for a person than another if and only if the person prefers the one to the other.

Preferencism also says that the determination in this biconditional goes from right to left. The biconditional could be true in an entirely unpreferencist way. A person's good could be determined in some way independently of her preferences, and then the person could form her preferences by always preferring histories that are better for her to histories that are worse. In that case, the biconditional would be true, but preferences would be determined by good. If a person's good is to be determined by her preferences, as preferencism requires, her preferences must themselves be independent of her good.

For one thing, this means we have to be careful about the concept of preference we adopt. One concept is the dispositional one: To prefer *A* to *B* is to be disposed to choose *A* rather than *B* when you have a choice between them. This is consistent with preferencism. But the existence of another concept is revealed by this fact: I prefer to get up early rather than waste time lying in bed on Saturday mornings, but sometimes I fail to do so. Evidently, I am sometimes not disposed to get up early, but nevertheless I prefer it. I do not prefer it in the dispositional sense, but in some other sense. In fact, I prefer it in the sense that I think it would be better for me. Thinking better is one concept of preference, but it does not suit a preferencist, because a preferencist needs preference to be independent of good. The preferencist must stick to preference as a disposition.

[9] See chapter 10 particularly.

9.5 Ideal Preferencism

The version of preferencism expressed in the preferencist biconditional is too pure for almost everyone. People's preferences are often hasty, badly thought out, ill informed, inconsistent, and in various other ways defective. Even hard-line preferencists find it implausible that a person's good should be determined by such defective preferences. Most preferencists rely on preferences that are idealized in one way or another: well informed, settled in a cool hour, made mutually consistent, and so on. This gives us:

One history is better for a person than another if and only if in ideal conditions the person would prefer the one to the other.

The notion of "ideal conditions" then needs to be spelled out. However, this improved claim also seems implausible, even before spelling it out. What a person would prefer in ideal conditions might perhaps be good for her in those conditions. But what would be good for her in those conditions might be different from what is good for her in her actual unideal conditions. If you were in a cool hour, a quiet cup of coffee might be good for you, whereas as things are you need a stiff drink. To fix this problem, we have to imagine the person, in her ideal conditions, forming preferences on behalf of herself in her actual unideal conditions. We get:

Ideal Preferencist Biconditional: One history is better for a person than another if and only if the person would in ideal conditions prefer the one to the other on behalf of herself as she is.

Let us stick with this form of the biconditional. By good fortune, it cuts through another difficulty that afflicts the original preferencist biconditional. People often have altruistic preferences: They are disposed to make choices on behalf of someone else rather than themselves. These preferences evidently do not determine what is good for themselves. But now we are picking out only the preferences they have on behalf of themselves, so we are ignoring altruistic preferences.

Once again, the determination has to go from right to left. This requirement is now not so easy to secure.[10] Ideal conditions are likely to include the condition that the person thinks about her preference. But preferencists cannot allow her to think about it in a way that presumes a notion of her good. She must not ask herself which histories would be better for her than

[10] This objection is developed by James Griffin, *Well-Being*, p. 17.

which, and determine her preferences on that basis. Instead her thinking must presumably proceed something like this. She must represent the alternative histories to herself as accurately as she can, and then just allow herself to end up preferring one or the other. This is not the most plausible model of thinking, but it is the one the preferencist must rely on.

For brevity, from now on the only preferences I shall mention are those a person would have in ideal conditions on behalf of herself as she is. I shall call these *ideal preferences.* Even when I simply say "preference," it is to be understood this way.

9.6 A Quantitative Concept of Good

The preferencist biconditional is not enough for utilitarian purposes. For each person, it determines what is better for her than what; it orders things according to their goodness for her. But a utilitarian needs more than an order; she needs a quantitative concept of good. Otherwise, the additive principle could not be applied; we could not make sense of the *total* of people's good. We must have a concept of quantities, or *degrees*, of good for a person. To cut a long story short, these degrees must be co-cardinal. This means that ratios of differences of good must be determinate both for a single person and between people. (In general, it is not enough for differences of good simply to be ordered as greater or less; their ratios must be determinate.) How can this be achieved on a preferencist basis?

Evidently, we must have a concept of degree or strength or intensity of preference that is also measured on a co-cardinal scale. That is to say, first, the degree to which a person prefers one history to another must be comparable to the degree to which she prefers a third history to a fourth. Second, this degree must also be comparable to the degree to which another person prefers one history to another. The comparability must be ratio-comparability, which means we can attach meaning to statements like "this preference is twice as strong as this one." How can this much comparability of degrees of preference be achieved?

It cannot be taken for granted. Many authors treat preferencism as the view that one should maximize the amount of satisfaction of people's preferences.[11] But this takes for granted a quantitative notion of preferences, which we are not entitled to without work. If preferencism is to progress beyond the preferencist biconditional, work has to be done.

[11] For instance, Brian Barry, "Rationality and Want-Satisfaction."

The question is conceptual. We must ask, What *concept* of degree of preference do we have, or can we construct, that satisfies the requirements. Having done that, we may then be up against the epistemological question of how we can find out what the degree of a particular preference is. The epistemological question may turn out easy or difficult, depending on what the appropriate concept of degree of preference turns out to be. But in any case, it is not the question I have to answer now. I am concerned with the ethical question of what makes histories good or bad. Given an answer to that, there will then be the subsidiary epistemological question of how we find out which histories are good or bad. I am not concerned with that.

Many authors have assumed that the only question is the epistemological one. R. M. Hare makes this assumption explicit.[12] He does not deal with the conceptual question, because he takes a particular concept of degree of preference for granted. He does not say explicitly what it is, but it is revealed by his argument. He says, "What I am going to discuss is the interpersonal comparison of degrees or strengths of preferences," but immediately beforehand he has said that the problem concerns "our knowledge of other people's experiences."[13] Evidently, then, he takes a degree of preference to be an experience. But degrees of preference conceived as experiences are plainly inadequate for our purposes. I dare say we have experiences associated with the degrees of some of our preferences; occasionally I experience strong longings and more occasionally weaker ones. But a huge multitude of preferences is needed to construct a measure of my good, and most of them give me no experiences whatsoever. I have a preference for being paid £120,000 annually rather than £119,950, and a preference for being paid £119,950 rather than £119,900, but I do not have time to experience these preferences. Just because we have so many preferences, most of them must be what Hume called "calm passions," "which, tho' they be real passions, produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation."[14] Hare sometimes confuses the degree of a preference between one option and another with the difference in the experiences that will result if one or the other option comes about. But that is to abandon preferencism for hedonism. Hare's work illustrates how important it is to get clear about our quantitative concept of preference before coming to epistemology.

[12] Hare, Moral Thinking, p. 117.

[13] Both quotations from *Moral Thinking*, p. 117.

[14] David Hume, *A Treatise of Human Nature*, book 2, part 3, section 3.

So, the question is conceptual: What ratio-comparable concept of degree of preference do we have or can we construct? Once we have a suitable concept, the preferencist utilitarian can use it to give us a ratio-comparable concept of degree of goodness. She can adopt the following principle, which is complicated to formulate but obviously what she requires:

Preferencist Principle: Let A, B, C, and D be histories. Let $g_i(A)$ and $g_i(B)$ be the goodnesses of A and B for a person i. Let $g_j(C)$ and $g_j(D)$ be the goodnesses of C and D for a person j. Then the ratio $\{g_i(A) - g_i(B)\}/\{g_j(C) - g_j(D)\}$ is equal to the ratio of i's degree of preference for A over B to j's degree of preference for C over D.

To establish degrees of good for a single person, we need only this extract from the preferencist principle:

Intrapersonal Preferencist Principle: Let A, B, C, and D be histories. Let $g(A)$, $g(B)$, $g(C)$, and $g(D)$ be their respective goodnesses for a person. Then the ratio $\{g(A) - g(B)\}/\{g(C) - g(D)\}$ is equal to the ratio of the person's degree of preference for A over B to her degree of preference for C over D.

9.7 The Expectational Concept

I have ruled out Hare's experience concept of degrees of preference. Expected utility theory supplies a better candidate concept, which provides comparability for a single person. Expected utility theory suggests that the degrees of a person's preferences about histories can be given by the person's preferences about uncertain prospects. The idea is this. Suppose the person prefers history A to B and history B to C. But suppose she is indifferent between B for sure and a gamble giving her either A or C at odds of one to two (that is to say, a gamble giving a 1/3 probability to A and a 2/3 probability to C). In effect, she is willing to accept one chance of making a gain from B to A in exchange for two chances of making a loss from B to C. Since she is willing to accept this gamble, the suggestion is that we should take her degree of preference for A over B to be twice her degree of preference for B over C. Developing this idea generally, expected utility theory supplies a way of constructing a complete scale of degrees of preference. It assigns a value called a *utility* to each history. The difference between the utility of one history and the utility of another is the degree to which the first is preferred to the second.

This certainly supplies a workable concept of degree of preference. I shall call it the *expectational* concept. There are alternatives. Any increasing transform – the square, for instance – of utilities measured this way provides a rival concept of degree. But there is something to be said for the expectational concept as opposed to these others. The use of probabilities provides a natural analogue of a pair of scales for measuring the strength – analogous to weight – of preferences. In the example, two chances of the loss from *B* to *C* balance the scales against one chance of the gain from *B* to *A*, so we naturally take the preference for the gain to be twice as strong as the preference against the loss. The rival concepts are less natural. Compare our concept of physical weight. Any increasing transform of weight could supply a rival concept of weight, but it would be less natural than our present concept. We use our concept because it has the natural and convenient feature that two objects each weighing one pound balance in a scale against one object weighing two pounds.

The expectational concept of degree is the most natural, but it is not forced on us by preferences alone. Preferences by themselves do not determine a concept of degree. The expectational concept is derived from preferences together with an idea of naturalness. So in adopting it, we are once more adding something to preferencism. How significant is this addition? That depends on the effect it has on our idea of good. If we adopt this concept of degrees of preference, the intrapersonal preferencist principle draws from it a corresponding concept of degrees of goodness for a person. I shall call it the expectational concept of good. Is it acceptable? Several authors have objected that it is not, or at least not necessarily. Indeed, this might be called the standard objection to Harsanyi's argument.[15]

I explained that many other concepts of degrees of preference are available. Each can pass over into an alternative concept of degrees of goodness. According to the standard objection, there is no reason to prefer one concept to another. This objection can be reinforced with another. If we adopt the expectational concept of good, it follows that, when faced with a choice between prospects, the person always (in ideal conditions) prefers the one with the greatest expectation of her good. This is Bernoulli's hypothesis again, in a different form. I have already said that Bernoulli's hypothesis is not very plausible on the face of it, because it implies risk neutrality about good. So this is a further objection.

[15] See, for instance, John Roemer, "Harsanyi's Impartial Observer Is *Not* a Utilitarian", Amartya Sen, "Welfare Inequalities and Rawlsian Axiomatics"; and John Weymark, "A Reconsideration of the Harsanyi–Sen Debate on Utilitarianism".

I am not convinced by the standard objection. We have a reason to prefer the expectational concept of degree of preference to others: It is more natural. This reason carries over to expectational degrees of goodness. The preferencist may reasonably say she is constructing a quantitative concept of good for a person, and this is the one she is going to construct. If we had a clear prior concept of degrees of good for a person, which was different from the expectational one, we could use it against the expectational concept. But the preferencist may say we do not. I agree with her about that. I believe our concept of degrees of good is not immediately intuitive, and needs to be constructed in some way. If we are to go any way with the preferencist, I do not think we can deny her this construction.

The objection to Bernoulli's hypothesis also rests on a presumed prior quantitative concept of good. Since I doubt we have one, I doubt the objection succeeds. A preferencist may plausibly say that Bernoulli's hypothesis is true because our quantitative concept of a person's good is constructed in such a way that the person is risk neutral about it.

Adopting the preferencist's concept is not merely a technical matter. It has concrete consequences within utilitarianism, because it helps to determine how we ought to act: We ought to maximize the total of people's good conceived this way. So the preferencist's idea of naturalness has moral consequences. This is exactly what she intends. She believes that people's preferences, together with the most natural concept of degrees of preference, determine how we should act. If we had an alternative intuitive concept, which gave us an alternative intuition about how we should act, we could use it against her. But we do not.

9.8 Interpersonal Comparability

In sum, I think the preferencist utilitarian can survive the standard objection. Her real problem is over comparisons between people. Can she produce a concept of degree of preference that is comparable between people to the required extent?

I shall assume from now on that we have already adopted the expectational concept of degrees of preference for each person. This is cardinal; it has all the intrapersonal comparability that is required. Only one thing more is required to give full co-cardinality: Each person's degree must be made ratio-comparable with each other person's. In effect, we have to pick a unit of degree of preference for each person. Since degrees of preference are measured by utility differences, we have to make utility differences comparable between people.

The leading contender for a preferencist way of making degrees comparable is the idea of *extended preferences.* We are assuming people have preferences between histories. For instance, I prefer a history where I teach philosophy to one where I teach economics. People may also have preferences between alternatives of the form: having the characteristics of a particular person and living in a particular history. For instance, I prefer having my characteristics and living in a history where I teach philosophy to having the characteristics of an economist and living in a history where I teach economics. Harsanyi calls preferences like these *extended preferences.* He calls the objects of these preferences *extended alternatives.* Each is a pair: a set of personal characteristics together with a history. I shall call it a *life.*

Suppose I have preferences over all extended alternatives. Then my preferences will rank all the possible lives of each person; they will compare the lives of different people. Suppose furthermore that I have preferences over uncertain prospects made up of extended alternatives. Then these will determine degrees of preference in the way I have described. Because everyone's life is included within my preferences, these degrees will be comparable between different people's lives. Here are interpersonally comparable degrees of preference, in a sense.

However, they are *my* preferences only: my preferences between different sorts of lives. Other people will have different extended preferences. Because of this, extended preferences cannot give us an interpersonal scale on grounds of preference alone. We would have to choose some particular person's preferences to go on, and that could scarcely be done on a preferencist basis. Indeed, presumably it could not be done on any good basis at all. But Harsanyi and others think they have a way of overcoming this problem.[16] They claim that, once we understand the idea of extended preferences properly, we shall see that everyone has the same extended preferences as everyone else. Extended preferences are universal. Consequently, there is a firm preferencist basis for making interpersonal comparisons of degrees of preference.

I am sure this is wrong. There is no reason why people should all have the same extended preferences, and many reasons why they should not. One reason why not is that people have different values. Their values will help to determine their preferences between different lives, so these preferences will differ. For instance, I value philosophy more highly than economics,

[16] This argument appears most clearly set out in John Harsanyi, *Rational Behavior and Bargaining Equilibrium*, pp. 58–59. See also Kenneth Arrow, "Extended Sympathy and the Possibility of Social Choice."

so I prefer working as a philosopher and having the characteristics of a philosopher to working as an economist and having the characteristics of an economist. I imagine many economists might have the opposite preference.

To be sure, when comparing my life with an economist's, I must do it properly. In deciding whether I prefer the life and characteristics of an economist, I am supposed to take account of everything that goes with them, including having the values of an economist. I must recognize that if I had the characteristics of an economist, I would value the life of an economist. But it is my extended preferences we are talking about, not the economist's extended preferences. As it happens, I prefer not to have the values of an economist. That is one reason I prefer not to be an economist.

We should also make sure we are dealing with ideal preferences. But when two people's extended preferences disagree, neither's need be less than ideal. Each person's preferences may be fully considered and so on. At least a preferencist must think that. For a preferencist, the standard of idealness for ideal preferences cannot be so stringent as to demand that different people's values coincide. Is there a true answer to the question of whether an economist's life is better or worse than a philosopher's? Suppose there is not. In that case, even if we were in such ideal conditions that we knew everything that is true, our values need not coincide. Alternatively, suppose there is a true answer. Then perhaps in ideal conditions our preferences would coincide because they would conform to the truth. But in that case preferencism would be false. Our ideal preferences would be determined by the truth of which life was better, whereas preferencism requires the determination to be the other way round.

So it seems the extended preferences of different people need not coincide. Yet, Harsanyi offers two arguments intended to show they must coincide. One is explicit; the other implicit. The explicit argument starts off from the correct observation that if people have different extended preferences, there is a causal explanation of why they do. Of course there must be some causal explanation of why I value philosophy, and why economists value economics (if they do). Suppose it is the star signs we were born under. Harsanyi claims that if we include this cause among the objects of our preferences, then our preferences will all be the same. But this is false. Perhaps we care about star signs and perhaps we do not, but at any rate, there is no reason why our preferences about them, or about anything else, should coincide. Harsanyi was led to his conclusion by a technical mistake, which I have said enough about elsewhere.[17]

[17] See my "A Cause of Preference Is Not an Object of Preference" and "Extended Preferences."

It is true that if we were all in the same causal situation, we would all have the same preferences. But we are not. Perhaps we could pick out some privileged causal situation, and base our interpersonal comparisons on the extended preferences we would have in that situation. Harsanyi sometimes seems to have in mind for this role a sort of causally empty situation, where we have been acted on by no causes apart from our bare human nature. He suggests we should use the preferences we would have in this causally empty situation. This is his second, implicit, argument for the claim that extended preferences are universal. But it is surely a fantasy to suppose we could have preferences determined by bare human nature.[18]

When he comes to a concrete case, Harsanyi has a quite different way of proceeding. He says,

> For example, if I want to compare the utility that I would derive from a new car with the utility that a friend would derive from a new sailboat, then I must ask myself what utility I would derive from a sailboat if I had taken up sailing for a regular hobby as my friend has done, and if I could suddenly acquire my friend's expert sailing skill, and so forth.[19]

Harsanyi evidently proposes to estimate how well off he would be if he had acquired a new sailboat and all his friend's sailing skills. He seems to be planning to form his extended preferences on the basis of an estimate of the benefits of leading a life like his friend's. This implies that the benefits of this life are determined in advance of Harsanyi's preferences. It is an antipreferencist view. It presupposes an idea of people's good that is independent of preferences. This is why I said in Section 9.1 that Harsanyi's own theory contains nonpreferencist elements.

9.9 Evolutionary Equilibrium

Ken Binmore offers a new theory developing the idea of extended preferences.[20] He argues that causal processes of social evolution determine our extended preferences. In the long run,[21] he argues, extended preferences will converge. This provides a potential new basis for preferencism. The difficulty with using extended preferences to provide interpersonal comparisons is that people's extended preferences differ. But Binmore supplies an argument to say they will not differ in the long run. No doubt we shall always

[18] For a discussion, see M. Kaneko, "On Interpersonal Utility Comparisons."

[19] Harsanyi, *Rational Behavior and Bargaining Equilibrium*, p. 59.

[20] Binmore, "Naturalizing Harsanyi and Rawls."

[21] Technically this is Binmore's *medium term*.

find some disagreements in our actual extended preferences: I suggested mine differ from an economist's. But Binmore would think these are minor deviations that exist only because social evolution has not had time to iron them out. From a broad viewpoint, adopting a long timescale, extended preferences converge.

Let me describe Binmore's view in a little more detail. His argument is set in a special strategic situation, where people regularly negotiate with each other behind an imagined veil of ignorance. People negotiate in pairs, to distribute goods between each other. Behind the imagined veil, neither person is supposed to know whose position she will occupy once the negotiation is completed; it might be her own position with her own characteristics or the other person's position with the other person's characteristics. In these conditions, the two settle on a distribution on the basis of their extended preferences. These preferences are formed by social evolution. This means that people tend to copy the attitudes of people they see doing well in their negotiations. Binmore argues that this process will drive us all to the same extended preferences in the long run. To be more precise, we will all make the same interpersonal comparisons of degrees of preference.

A more specific outcome of Binmore's argument is surprising. It turns out that in the long run we will assign high degrees of preference to people who have a lot of bargaining power. Let us suppose everyone prefers a day of sunshine to a day of rain. We will assign a higher degree to this preference when it belongs to a powerful person than we do when it belongs to a less powerful person. We shall suppose powerful people have more intense preferences. If we feed this conclusion into the preferencist principle, we shall conclude that powerful people tend to get more benefit from good things than less powerful people do. Consequently, if goods are distributed on a utilitarian basis, the lion's share will go to the powerful. Naturally, these same people will also get the lion's share if the goods are distributed by a free-for-all. In the long run, utilitarianism will reproduce what would have been the result of a free-for-all. Binmore derives this conclusion by mathematics, but he does not offer an intuitive explanation of why it happens.

What does all this do for preferencism? At first, it seems to give it support. Preferencism was laboring under the difficulty that preferences did not seem to provide a basis for interpersonal comparisons of degrees of preference. Extended preferences were supposed to do the job, but different people have different extended preferences, and there are no preferencist grounds for choosing between them. Now Binmore suggests these differences are unimportant. They are temporary only. Our extended preferences will converge in the long run because evolutionary processes will make sure they do. So we can perhaps ignore the differences. Moreover, the preferences we are converging on are determined entirely by blind causal forces. They contain no taint of a nonpreferencist theory of good. All this is good for preferencism.

But actually this very ethical neutrality prevents Binmore's argument from supporting preferencism. Binmore calls his theory "naturalistic." I believe he means to say it is a natural history of ethical beliefs. He thinks people's extended preferences are a natural feature of people, determined by natural, causal processes, and he aims to give an explanation of these processes. Because they determine extended preferences, these natural processes determine degrees of preference that are comparable between people. They will lead people to make interpersonal comparisons of degrees of good, corresponding to the degrees of their preferences. That is to say, these evolutionary processes will cause people to have certain beliefs about how good or bad things are for people. In the end, they will also lead people to have particular beliefs about how they and others ought to act. Let us suppose they will lead them to utilitarian beliefs, with interpersonal comparisons determined in the way described by Binmore.

A successful natural history of ethics might explain why people will believe they ought to maximize the total of people's good. It may also show that their notion of good will derive from preferences. But a preferencist utilitarian needs something quite different. She needs a demonstration that people ought to maximize the total of people's good, where people's good is determined by their preferences. A natural history of ethics gives no support to these claims whatsoever. It may tell us what people will believe, but it does so in a way that gives no grounds for their beliefs. This type of naturalism passes ethics by. It is irrelevant to preferencism, since preferencism is an ethical theory.

9.10 Conclusion

I conclude that preferencist utilitarianism fails. Preferencism cannot generate a concept of good solid enough to make sense of interpersonal comparisons of good. Interpersonal comparisons can only be achieved by means of a different, nonpreferencist theory of good.

References

Arrow, K. J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review, Papers and Proceedings* 67, 219–225; reprinted in his *Collected Papers,* Vol. 1: *Social Choice and Justice*, Harvard University Press, Cambridge, MA, 1983, pp. 147–161.

Barry, B. 2008. Rationality and want-satisfaction. *In Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 281–299.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, vol. 5; trans. Louise Sommer, Exposition of a new theory on the measurement of risk. *Econometrica* 22 (1954), 23–36.

Binmore, K. 2008. Naturalizing Harsanyi and Rawls. *In Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 303–333.

Broome, J. 1993. A cause of preference is not an object of preference. *Social Choice and Welfare* 10 (1993), 57–68.

Broome, J. 1998. Extended preferences. In *Preferences*, ed. C. Fehige and U. Wessels, de Gruyter, Berlin, pp. 279–296.

Broome, J. 1991. *Weighing Goods*. Blackwell, Oxford.

Griffin, J. 1986. *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford University Press, Oxford.

Hare, R. M. 1982. *Moral Thinking: Its Levels, Method and Point*. Oxford University Press, Oxford.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321; reprinted in his *Essays on Ethics, Social Behavior, and Scientific Explanation*, D. Reidel, Dordrecht, 1976, pp. 6–23.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge.

Hild, M., Jeffrey, R., and Risse, M. 2008. Preference aggregation after Harsanyi. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 198–217.

Hume, D. 1978. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch. Oxford University Press, Oxford.

Kaneko, M. 1984. On interpersonal utility comparisons. *Social Choice and Welfare* 1, 165–175.

Mongin, P. 1995. Consistent Bayesian aggregation. *Journal of Economic Theory* 66, 313–351.

Moore, G. E. 1966. *Ethics*, 2nd ed. Oxford University Press, Oxford.

Parfit, D. 1984. *Reasons and Persons*. Oxford University Press, Oxford.

Roemer, J. Harsanyi's impartial observer is *not* a utilitarian. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 129–135.

Sen, A. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7, 243–262.

von Neumann, J. and Morgenstern, O. 1947.*Theory of Games and Economic Behavior*, 2nd ed. Princeton University Press, Princeton, NJ.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, 255–320.

10

Harsanyi, Rawls, and the Search for a Common Currency of Advantage

Robert Sugden

The aim of welfare economics is to evaluate alternative allocations of resources, or alternative economic institutions, in terms of their impact on people's well-being or interests. One classic line of approach is to look for what I shall call an *interpersonal common currency of advantage* – a measure which integrates all aspects of a person's well-being or interests or opportunities (or, on some more modest accounts, all "economic" aspects of one of these) and which permits interpersonal comparisons. In this chapter, I review some attempts to find such a common currency, focusing particularly on the work of John Harsanyi and John Rawls. I argue that Harsanyi's ambitious and economically sophisticated attempt fails. In contrast, Rawls's general strategy is sound, but he offers only a rough sketch of how it might be translated into welfare economics. I make some suggestions about how Rawls's approach might be extended to generate a common currency of advantage. The basic idea is to construct a money metric of economic opportunity.

10.1 The Common Currency of Pleasure

I have said that this chapter is about the search for a common currency of advantage. *Advantage* is a shorthand expression I shall use to represent any conception of what makes a life good for a person, or of what provides a person with the opportunity for a good life, and which is deemed to be relevant for some problem of social choice. I realize that *advantage* is a

This chapter develops some ideas from a paper that I presented at the Caen Conference in honour of John Harsanyi and John Rawls in June 1996. I am grateful to the participants at that conference, particularly Richard Arneson, Brian Barry, Hervé Moulin, Philip Pettit, and my anonymous referee, for valuable comments.

clumsy term, but I cannot think of a better one that is sufficiently general to encompass both the actual *attainment* of desirable states of affairs (which is often called *well-being*) and the *opportunity to attain* such states. At this stage, I wish to leave open the question of whether value should be assigned to attainments or to opportunities.

By a *currency* of advantage, I mean the following. There is some binary relation of the form "gives at least as much *X* as," which is complete and transitive in some relevant domain of social choice and which can be represented by a real-valued function. If increases in the value of this function are asserted to correspond with increases in advantage, then *X* is a currency of advantage. In speaking of "some relevant domain of social choice," I am being deliberately imprecise. The domain may be narrow or broad; for example, it might include all aspects of life or only those that are narrowly economic. It may be specific to a single individual, or it may range across individuals. It may be a domain of attainments or a domain of opportunities. I am also leaving open the question of whether a currency provides *more than* an ordinal measurement of advantage. (On the different ways in which advantage – or utility – might be measurable, see Bossert and Weymark, 2004.) A currency is *common* to a set of domains if each of those domains is a subset of the domain of the currency. For example, market value is a currency that is common to all bundles of tradable goods. In this chapter, I am concerned with currencies of advantage that are common to many aspects of life and to many individuals.

The roots of modern welfare economics can be found in utilitarianism. For a utilitarian in the classical, hedonistic tradition of Bentham, there is a common currency of advantage: pleasure. According to hedonistic utilitarianism, the proper object of government is the good of society; the good of society is the aggregate of the individual goods of each of its members; and the good of an individual is the net aggregate of pleasure minus pain experienced over his or her life, pain being understood as the negative of pleasure. Thus, pleasure is taken to be a common currency both *intrapersonally* and *interpersonally*. Intrapersonally, it is the common currency in which different experiences of the same person can be compared; interpersonally, it is the common currency for comparing the experiences of different persons. Although few economists now subscribe to hedonistic utilitarianism, much of the formal structure of welfare economics has been carried over from the utilitarian past. To understand modern welfare economics, it is necessary to recognize its hedonistic origins.

If the classical utilitarian project were to succeed, we would need to be able to identify a single, interpersonal scale of mental experience that corresponded reasonably closely with our prescientific understanding of pleasure. No one has ever come close to finding such a scale or even to finding a promising method of looking for one. Neuroscience tells us that the human brain, and the system of mental states that it induces, is a vastly complex spontaneous order of interrelated circuits and modules, with no single control centre. Given what we now know, the idea that the mind has a single accounting system for experience, in the same way that a firm might have a single accounting system for revenue, is simply incredible.

Well before the development of neuroscience, nineteenth-century economists and philosophers realized that a common currency of pleasure was unlikely to be found. It also came to be recognized that people desire things other than pleasure – indeed, that they desire some things that are not even experiences. At this point, the utilitarian tradition divided.

One line of inquiry has focused on the question of what makes a person's life good for him. Hedonistic utilitarians had asserted that pleasure was the only ultimate source of goodness. Later utilitarians have looked for a richer conception of individual well-being. James Griffin's (1986, 1991) account of well-being is representative of modern work in this tradition. For Griffin, there is a list of "prudential values" that are "valuable in any life" (1991, p. 64). What makes a life valuable to the person who lives it is its containing a good mix of these prudential values, something like the way a cake is good if it contains a good mix of the right sorts of ingredients (1986, p. 36). Since statements about well-being are *judgments* about what is valuable, we may be able to impose some structure on such statements by appealing to logical properties that are possessed by all judgments of value. Following this strategy, Broome (1991) shows how certain axioms of coherence among judgments of value imply that, for any given set of judgments, there is an implicit common currency of goodness.

In this chapter, I am more concerned with a different branch of the utilitarian tradition – the branch that led to modern welfare economics. This branch began from the efforts of the neoclassical economists of the late nineteenth century to find adequate foundations for their theory of choice. Early versions of neoclassical theory had assumed an intrapersonal common currency of pleasure and started from the hypothesis that a rational individual seeks to maximize pleasure (Edgeworth, 1881; Jevons, 1871/1970; Marshall, 1890/1920; Walras, 1889/1954). Recognizing the problems created by this assumption, neoclassical economists were attracted by an alternative strategy, which seemed to avoid the need for any psychological assumptions at all. The clearest early statement of this strategy is Pareto's (1906/1972), although Fisher (1892/1925) has some claim to be regarded as its pioneer. In

Pareto's words, the aim of this strategy was to develop a form of economics whose structure was like that of "rational mechanics," "deduc[ing] its results from experience, without bringing in any metaphysical entity" (1906/1972, p. 113).

The modern economic theory of choice is in direct line of descent from Pareto's work. This theory has a common *intra*personal currency of advantage: preference satisfaction. This currency is still usually called *utility*, but with utility functions being interpreted merely as representations of preference orderings and preferences being interpreted as those dispositions that are revealed in a person's choices. However, there is an intrapersonal currency of preference satisfaction only if preferences are assumed to be complete and transitive. In effect, we have to assume that each person acts as if maximizing *something*. What grounds do we have for making that assumption, if we have abandoned the utilitarian claim that a rational person prefers whatever maximizes his pleasure, and if we refuse to speculate about any other "metaphysical entities" that might be maximized?

Two different answers can be found in the literature of economics. One answer is implicit in the logic of Samuelson's (1947, pp. 90–124) revealed preference approach: The hypothesis that preferences are complete and transitive is (it is claimed) consistent with most observations of people's actual choices. This is just a fact about behaviour for which no explanation need be sought. The other answer is that of Savage: the preferences of a *rational* person are complete and transitive because completeness and transitivity are necessary properties of rationality – necessary in something like the sense that the rules of logic are necessary truths (Savage, 1954, p. 6). These two answers are logically independent, but they are often combined by adding the hypothesis that, as a matter of fact, people tend to be rational, at least in the situations in which economists study them. Then the alleged truth that completeness and transitivity are necessary properties of rational preferences provides part of the explanation for the alleged fact that people tend to behave as if they had complete and transitive preferences.

Neither Samuelson's answer nor Savage's is completely convincing, for reasons that I have set out elsewhere (Sugden, 1991; see also Anand, 1993). For the purposes of this chapter, however, I shall set these reservations to one side and accept the supposition that for each person preference satisfaction is an intrapersonal common currency. My concern is with two subsequent questions. Can preference-satisfaction be regarded as an intrapersonal currency *of well-being*? And if it can, is there an *inter*personal common currency of preference-satisfaction that can play the role in modern welfare economics that pleasure plays in hedonistic utilitarianism?

10.2 Harsanyi, Imaginative Empathy, and Rational Preference

Over many years, Harsanyi (1953, 1977, 1982) has argued that each of us can gain access to an interpersonal common currency of well-being by "imaginative empathy." In setting out Harsanyi's ideas, I shall refer mainly to his 1982 article. The critique that I shall develop owes much to the previous work of Scanlon (1991), Broome (1991, 1993), Griffin (1991), and Hausman and McPherson (1996, pp. 71–83).

Harsanyi offers an analysis of what he calls "moral value judgements." Suppose there are two alternative sets of rules, *A* and *B* (say, those of socialism and of capitalism), either of which could be used to govern our society. Under each of those sets of rules, each individual has a "social position." To arrive at a moral value judgment between these rules:

> All we have to do is to ask ourselves the question, "Would I prefer to live in a society conforming to standard *A* or in a society conforming to standard *B*? – assuming I would not know in advance what my actual social position would be in either society but rather would have to assume to have an equal chance of ending up in any one of the possible positions." (1982, p. 46)

Thus, Harsanyi is assuming the existence of preferences over the set of all (r, i) pairs, where r is a rule and i is one of the "social positions" associated with that rule; (r, i) stands for "having position i in a society governed by rule r." Such preferences are often called *extended preferences.* Since Harsanyi assumes that extended preferences are complete and transitive, he is assuming a common currency in which all (r, i) pairs can be evaluated; and since those comparisons include comparisons between social positions occupied by different people, this currency is interpersonal.

The difficulty is to find a satisfactory interpretation for propositions of the form "I prefer having position i under rule r to having position j under rule s," when at least one of those two positions is not occupied by oneself. If we are to assume that extended preferences over (r, i) pairs are complete, we need to be sure that all the relevant comparisons are meaningful. Further, they must be meaningful in the right way. Harsanyi is trying to derive *moral* value judgments. The sense of morality here, I take it, is some kind of *neutrality* as between individuals. Thus, extended preferences must be neutral. That is, in a proposition of the form "I prefer (r, i) to (s, j)," the "I" should be doing as little work as possible: I should not be using my personal values to evaluate what it would be like to be in someone else's position, when that other person does not share my values.

Harsanyi tries to convince us that we can make all these comparisons. In an ideal world of perfect information and perfect rationality, we would all have the *same* extended preferences, so that "I prefer (r, i) to (s, j)" would be reducible to an impersonal "(r, i) is preferable to (s, j)." In everyday life, Harsanyi says, we make "interpersonal utility comparisons" all the time. How do we do this? Harsanyi offers the following explanation:

> Simple reflection will show that the basic intellectual operation in such interpersonal comparisons is imaginative empathy. We imagine ourselves to be in the shoes of another person, and ask ourselves the question, "If I were now really in *his* position, and had *his* taste, *his* education, *his* social background, *his* cultural values, and *his* psychological make-up, then what would now be *my* preferences between various alternatives, and how much satisfaction or dissatisfaction would *I* derive from any given alternative?"...
>
> In other words, any interpersonal utility comparison is based on what I will call the *similarity postulate*, to be defined as the assumption that, once proper allowances have been made for the empirically given differences in taste, education, etc., between me and another person, then it is reasonable to assume that our basic psychological reactions to any given alternative will otherwise be much the same. (1982, p. 50)

Harsanyi defends the similarity postulate as an a priori principle of parsimony which corresponds with good scientific practice: we should not postulate unobservable differences if we can explain our observations without them (1982, pp. 50–52).

I have no quarrel with the similarity postulate, but it does not do all the work that Harsanyi wants it to do. Consider one of Harsanyi's examples. He compares the utility he derives from a new car with the utility a friend derives from a new sailboat; he makes this comparison, he says, by asking himself what utility he would derive from a new sailboat if he had taken up sailing as a regular hobby as his friend has done (1977, p. 59). But how exactly does this comparison work? By imaginative empathy, Harsanyi presumably simulates in his own mind the nearest equivalent he can find to his friend's experience of the new sailboat. Appealing to the similarity postulate, he then assumes that the friend's actual psychological reaction to the sailboat is the same as his own simulated reaction. So far, so good. Harsanyi now compares that simulated reaction with his own actual reaction to the new car. The problem, I suggest, is to understand what *this* comparison is.

In Harsanyi's theory, it is a *preference*: it is the preference of whoever makes the comparison (in the story of the sailboat and the car, the preferences of Harsanyi himself). If a preference can be understood as a disposition toward choice, then the comparison seems to amount to the following question: Which of the two psychological reactions would Harsanyi choose to experience? However, Harsanyi needs to claim that each of us would arrive at the *same* answer if, having imaginatively simulated the two psychological reactions – Harsanyi's reaction to the car and his friend's reaction to the sailboat – we each asked the question: Which would *I* choose to experience?

To see why this degree of agreement is required, look again at the passage which begins, "Simple reflection will show." If I really were in your shoes, with your tastes, education, social background, cultural values, and psychological makeup, then my preferences between any given pair of alternatives would clearly be the same as yours: if I were like you in every respect, I would have the same preferences as you. But that is a truth of logic; it can be said without any appeal to imaginative empathy. Harsanyi wants to say in addition that I can arrive at your preferences by simulating your experiences in my mind and then comparing those simulated experiences. In other words, your preferences over your actual experiences are the same as my preferences over my simulations of those experiences.

Harsanyi's references to "psychological reactions," and his use of the words "satisfaction" and "utility," strongly suggest that he has in mind something like the classical utilitarian idea of a common currency of pleasure. If there were a common currency of mental experience (and if we all had access to it), then the comparisons Harsanyi requires us to make would be straightforward. But, as I have said, no such common currency has been found.

In fact, Harsanyi denies that he is a hedonistic utilitarian, dismissing hedonistic psychology as "completely outdated." He categorizes himself as a *preference utilitarian.* Preference utilitarianism, he says, defines each person's utility function in terms of his personal preferences rather than in terms of pleasure and pain; the implication is that preference is a primitive concept (1982, p. 54). In justification of preference utilitarianism, Harsanyi says that this is

> the only form of utilitarianism consistent with the important philosophical principle of *preference autonomy.* By this I mean the principle that, in deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own wants and his own preferences. (1982, p. 55)

He qualifies the principle of preference autonomy by distinguishing between *manifest preferences* and *true preferences* and by interpreting preference autonomy in terms of true preferences:

> [A person's] manifest preferences are his actual preferences as manifested by his observed behaviour, including preferences possibly based on erroneous factual beliefs, or on careless logical analysis, or on strong emotions that at the moment greatly hinder rational choice. In contrast, a person's true preferences are the preferences he *would* have if he had all the relevant factual information, always reasoned with

> the greatest possible care, and were in a state of mind most conducive to rational choice. (1982, p. 55)

The principle of preference autonomy asserts that for each person the satisfaction of her true preferences is an intrapersonal currency of well-being. However, in Harsanyi's system, judgments about well-being are located in the mind of whoever is making "moral value judgments" – let us say, in the mind of the *judge.* Thus, whether person *i*'s well-being is judged to be greater in one situation than in another is a property of the judge's extended preferences, and those extended preferences are the judge's personal preferences over his simulations of *i*'s experiences. Again, we see that it is crucial for Harsanyi that this process of simulation, when carried out by any judge under the right conditions, should reproduce the true preferences of the person in question. Otherwise, the principle of preference autonomy would not hold.

Why, then, does Harsanyi believe that, under ideal conditions, imaginative empathy allows each of us to reproduce any other person's true preferences? Only one answer is consistent with the rest of Harsanyi's argument. This answer is in the same spirit as the *common prior* assumption in game theory – the principle that rational individuals who share the same information must hold the same subjective beliefs – which is due to Harsanyi (1968). It is to claim that every fully specified decision problem has a unique rational solution (rational indifference being allowed to count as a possible solution). I shall call this the *common preference principle.* According to this principle, if a decision maker has all the relevant factual information about a particular problem, if he reasons with the greatest possible care, and if he is not distracted by strong emotions, he will arrive at that problem's unique rational solution. If the problem is to decide what a particular person should do, then that person's tastes, education, social background, cultural values, and psychological makeup are all part of the relevant factual information. Given all this information, a unique rational solution exists, which is accessible to every rational individual. Thus, imaginative empathy allows us to reproduce other people's true preferences.

Extending this argument, suppose we follow Harsanyi in treating a hypothetical choice between the social positions of different people – for example, a choice between (r, i) and (r, j) when i and j are different positions – as a valid decision problem. Then the common preference principle implies that rational and fully informed individuals will have the same extended preferences.

The difficulty with this line of argument is that we are given no reasons for accepting the common preference principle. The similarity postulate does not provide such a reason because we are not dealing with cases between which there are no observable differences. It is an empirical question whether all individuals solve decision problems in the same way when they are fully informed, do not make logical errors, and are not swayed by strong emotions. If people can reach different conclusions under such circumstances, then there is an observable difference to be explained, and to explain it, we would have to reject the common preference principle.

The most obvious way of justifying the common preference principle would be to appeal to some objective common currency of well-being to which (it would have to be claimed) every fully rational person has access. Then it would be natural to claim that any person's true preferences coincide with the ranking of alternatives in terms of that person's well-being and that any rational and fully informed judge can replicate those preferences by making use of the same currency of well-being. My hunch is that Harsanyi is unconsciously assuming that the preferences of all rational and fully informed individuals are grounded in some unique interpersonal common currency of well-being, which is not preference. (That is also Broome's [1993, p. 67] diagnosis of what he too sees as problems in Harsanyi's argument.) But if we need to find such a common currency to legitimate Harsanyi's method of imaginative empathy, we are back to square one: the currency we need has yet to be found. Conversely, if we do not even look for that currency, the common preference principle can be no more than an act of faith.

10.3 Rawls and Primary Goods

One of the earliest critiques of Harsanyi's theory of imaginative empathy can be found in Rawls's (1971) discussion of interpersonal comparisons. In response to what he sees as fatal objections to Harsanyi's approach, Rawls proposes a common currency of "primary goods." Rawls's objections are worth quoting at length:

> Let us distinguish between evaluating objective situations and evaluating aspects of the person: abilities, traits of character, and system of aims. Now from our point of view it is often easy enough to appraise another individual's situation, as specified say by his social position, wealth, and the like. . . . We put ourselves in his shoes, complete with our character and preferences (not his), and take account of how our plans would be affected. We can go much further. We can assess the worth to us of being in another's place with at least some of his traits and aims. Knowing our plan of life, we can decide whether it would be rational for us to have those traits and aims, and therefore advisable for us to develop and encourage them if we can . . . . [But] what we cannot do is to evaluate another person's total circumstances, his objective situation plus his character and system of ends, without any reference

> to the details of our own conception of the good. If we are to judge these things from our own standpoint at all, we must know what our plan of life is. (1971, p. 174)

What Rawls is saying, I think, is that there can be no disembodied preferences: each of us has preferences only from his own “standpoint.” A person’s standpoint is partly defined by his own “conception of the good” or “rational plan of life.” (For Rawls, these two concepts are roughly interchangeable. He endorses a theory of “goodness as rationality” in which a person’s good is determined by what for that person is “the most rational plan of life” [1971, p. 395].) Rational preferences are contingent on conceptions of the good: I prefer things to the extent that they advance my conception of the good. Thus, I cannot have preferences among different conceptions of the good. To put the same point another way: There is no neutral standpoint from which I can evaluate alternative systems of values.

To say that there can be no disembodied preferences is not to deny the *meaningfulness* of judgments of the form “person i in situation x is better off than person j in situation y”: it is to deny their *neutrality*. To make such a judgment, one must already have a conception of the good and that need not be the same as i’s or j’s. Of course, the force of Rawls’s argument depends on the presupposition that there is no one conception of the good whose correctness can be established publicly; were there such a conception, judgments made from within it would be neutral in the sense that Harsanyi needs. The more one believes in the existence of values that are (and can be agreed to be) valuable in any life – as, for example, Griffin (1986) does – the less one will be swayed by Rawls’s argument. Here I can only record that I side with Rawls (see Sugden, 1989).

Rawls’s argument does not cut against hedonistic utilitarianism. For the hedonistic utilitarian, the amount of pleasure in a person’s life is a matter of fact: it is an element of that person’s “objective situation.” The utilitarian principle that pleasure is the only good is a particular conception of the good. From the standpoint of that conception of the good, we can evaluate alternative objective situations. Similarly, Harsanyi’s approach would be immune to Rawls’s argument if what Harsanyi calls *social positions* were defined objectively and if his judges – the people who form moral value judgements – were assumed to share a common conception of the good. I think Harsanyi *does* intend that social positions are interpreted objectively, with such factors as a person’s tastes, education, and psychological makeup being treated as among the objective facts of a position. But Harsanyi does not propose any substantive theory of the good analogous with hedonism. If I am right that he endorses the principle of common preference, Harsanyi is effectively assuming that all rational individuals, by virtue of their rationality, subscribe to a common conception of the good, but he does not tell us what that conception is. That is the missing common currency.

Rawls is looking for a theory of justice that will provide the basis for a stable political system, construed as a voluntary association of free persons. Thus, it is essential that the theory should command general agreement. For Rawls, it is axiomatic that we cannot hope to achieve general agreement on any common conception of the good. (This is a particular theme of his later work [e.g. Rawls, 1985].) For his theory of justice, he needs a common currency of advantage that does not presuppose any particular conception of the good. Rawls proposes the concept of *primary goods* as a way of satisfying these requirements.

Primary goods are goods that normally have a use whatever a person's rational plan of life and thus that every rational person can be presumed to want (1971, p. 62). A person's *expectations* are defined by an index of primary goods; it is assumed that everyone prefers more rather than less primary goods, as measured by this index. Rawls does not look behind primary goods to try to measure the value of the satisfactions individuals achieve from them: that would require a substantive theory of the good, and Rawls does not want to invoke any such theory. The requirements of justice are satisfied if the distribution of expectations is fair. Rawls describes the strategy of focusing on primary goods as a "simplifying device" and as representing an agreement on "the most feasible way to establish a publicly recognized objective measure" of people's situations (1971, p. 95). The idea seems to be that the index of primary goods is a simple and workable model of individuals' judgements about what serves their interests and that this model is neutral with respect to alternative conceptions of the good.

In the language of preferences, we might say that primary goods are things that normally have value whatever a person's preferences. In other words, the more primary goods a person has, the better able she is to satisfy her preferences, whatever those preferences may happen to be. Rawls's veil of ignorance prevents his "contracting parties" from using knowledge about the actual preferences they have as individuals; but they are allowed to use knowledge about the psychology and sociology of preferences in general (Rawls, 1971, pp. 136–142). We might think of this psychological and sociological knowledge as delimiting a range of possible preferences. Using only this knowledge, Rawls thinks, each contracting party can recognize that primary goods are valuable to her. They are valuable to her because they are generally useful for satisfying the range of preferences that she might possibly have.

Rawls's aim, I take it, is to formulate his theory so that each individual's claims of justice are independent of her personal preferences. (I use the word *personal* because impersonal facts about preferences in general are relevant to the claims that everyone can make.) Thus, Rawls's theory does not evaluate a person's actual consumption, as utilitarian theories do. Instead, what is evaluated is the person's set of opportunities, as represented by her endowment of primary goods. And by using primary goods themselves as the currency of advantage, Rawls does not allow personal preferences to enter into the evaluation of opportunities.

Rawls's intention to exclude personal preferences from his theory is made clear in his discussion of expensive tastes. He asks us to imagine "two persons, one satisfied with a diet of milk, bread and beans, while the other is distraught without expensive wines and exotic dishes." Rawls asks whether the second person can appeal to his expensive tastes as grounds for claims of justice and concludes that the answer is "No": it is, he says, reasonable "to hold such persons responsible for their preferences" (Rawls, 1982, pp. 168–169). The implication seems to be that the demands of justice are satisfied if the two persons in the example have equal opportunities to satisfy the sorts of preferences that people in general can be expected to have. Which opportunities, from those that are open to her, the individual then chooses to take up is her responsibility and is outside the scope of justice.

Of course, Rawls's example of expensive wines and exotic dishes is morally loaded and rather trivial. Later contributors to the theory of justice have discussed examples that raise more significant and more difficult questions about the relationship between personal preferences and justice. What about a person – say, an elderly woman who has spent all her life in a poverty-stricken village in Bangladesh – who has learned not to desire anything more than a bare subsistence? Does the fact that her preferences can be satisfied at low cost give her less claim on resources than a person who has had a less deprived life? What about a young man whose upbringing in a crime-ridden part of a decaying city has induced a taste for expensive illegal drugs and a deep aversion to the discipline of employment? Do these preferences give him more claim on welfare benefits than someone who has been brought up to want to support himself by work? What about a person who, because of physical handicap, has an expensive taste for a single-level house with wheelchair access to all rooms? What about a person who, after developing his musical talents over many years of hard practice, has an expensive taste for a high-quality violin? There has been much discussion about where to draw the line between those preferences for which each individual should be held responsible and "objective" differences between individuals that can properly constitute the grounds for differential claims of justice (Arneson, 1989; Cohen, 1989; Dworkin, 1981a, 1981b; Roemer, 1996; Sen, 1992; Van Parijs, 1995). For the present, what matters is Rawls's position: that claims of justice may not be grounded on personal preferences.

Rawls's list of primary goods comprises rights and liberties, powers and opportunities, income and wealth, and self-respect (1971, p. 62). Income and wealth are the goods in this list that are most directly related to the conventional domain of economics, and I shall focus on these. The analysis that I shall develop may be applicable to a wider domain, but there are problems enough in dealing with income and wealth. Rawls usually treats these two goods as if they were different ways of talking about the same thing: often he uses the words "income and wealth" as a composite formula. Since wealth is just the capitalized value of a flow of income, little is lost by considering income alone. That is what I shall do.

Surprisingly, Rawls does not seem to recognise any problem in measuring income. He treats income as if it were a physical object like rice, with a natural unit of quantity; just as rice can be used to promote life plans that involve being well fed, so income can be used to promote a wide range of life plans. Rawls's thought, I take it, is that all rational life plans require some goods that can be bought in markets. Although specific goods may be valuable only in relation to specific plans, purchasing power – the ability to buy marketed goods in general – is valuable for *all* rational life plans, and thus is a primary good. Rawls's concept of "income," then, should be understood as purchasing power.

From the perspective of an economist, Rawls's treatment of income hides some deep problems. Income is not a physical object, it is a summary statistic. Income measures are constructed by means of various theoretical and conventional procedures. As Gibbard (1979) points out, any measure of real income requires some set of relative price weights. As long as we wish to compare the situations of different individuals who face the same array of market prices, there is no problem in using market value as the metric of purchasing power. But if we wish to make comparisons across economic regimes, as Rawls needs to do if he is to assess the justice of alternative economic institutions, we cannot avoid the index number problem. That is, the ranking of any pair of price-income combinations in terms of purchasing power may depend on which price weights are used. (There is a corresponding problem if, as Rawls [1971, pp. 93–95] intends, we try to construct an index of "expectations" that aggregates across primary goods; see Plott [1978].)

Unfortunately, many of the standard economic methods for dealing with the index-number problem work by making use of personal preferences. For example, an argument familiar to most beginning students of economics shows that if, when the prices of some period t have been used as weights, the volume of a person's consumption is calculated to be lower in some other period t' than in t, then that person is unambiguously worse off in period t'. Or if a chain index of consumption through time is constructed by continuously updating the price base, and if a person's consumption so calculated increases continuously, then that person is becoming better-off through time. Results like this are useful to economists who wish to use market-generated data to measure preference-satisfaction but as Barry (1996) has noticed, they do not help us to find a measure of purchasing power that is independent of personal preferences.

A further problem with conventional income measures is that they depend on the assumption that each good can be bought, in whatever quantity the consumer chooses, at a given price. This assumption breaks down when consumers face quantity constraints (as, for example, in most systems of socialized or insurance-based medical care). It also breaks down in the case of public goods (which might be seen as an extreme case of quantity constraint: the individual has no choice about how much to consume). For an economist, the natural response to such problems is to turn to welfare economics. For example, the methods of cost-benefit analysis can be used to assign money values to a person's consumption of public goods. But such methods are designed to measure the satisfaction of personal preferences, while (to repeat) Rawls needs a measure of economic opportunity that is independent of personal preferences.

I shall now suggest an approach that may take us some way toward the measure of economic opportunity that Rawls's project needs. The guiding idea is to try to measure the range of opportunity offered by any *opportunity set,* that is, by any set of mutually exclusive and exhaustive options. To keep to the spirit of Rawls's analysis, that measure should be independent of the personal preferences of whoever faces the opportunity set. If that can be achieved, the measure of opportunity will be an interpersonal common currency of advantage in the same way that a Rawlsian primary good is. Just as a primary good is valuable because it has a use in a wide range of life plans, so a set of opportunities should be deemed to have value to the extent that it can be used to satisfy a wide range of possible preferences. Just as Rawls maintains neutrality between rival conceptions of the good, so the measure of opportunity should be neutral with respect to the different preferences that people might be expected to have.

My approach draws on ideas from two separate bodies of literature. When discussing Rawls's theory of justice, commentators often contrast the primary goods approach used by Rawls with the preference utilitarianism (or welfarism) of welfare economics. Such commentaries tend to overlook a long-standing tradition in welfare economics which, although focusing on personal preferences, effectively uses income as a common currency of advantage. The tradition I have in mind centres on the concept of economic efficiency and on the principle of the compensation test; its practical application is in cost-benefit analysis. In this tradition, money is often said to be a "measuring rod" for preferences. The idea is that the preferences of different individuals are made commensurable by way of money equivalences. There are many different variants of this general theoretical strategy. For present purposes, the *money metric* approach is most useful (McKenzie, 1983).

I shall combine the idea of the money metric with some ideas from one strand of the growing literature on measuring "freedom of choice" or "effective freedom." Many writers have proposed criteria for ranking opportunity sets, given a set of preference orderings over options. I shall refer to the orderings in such a set as *potential preferences*, while leaving open the interpretation to be given to "potential." This approach was first used to measure *flexibility*, that is, the instrumental value that a wide range of opportunity has to an expected-utility-maximizing individual who is uncertain about her future preferences (Arrow, 1995; Koopmans, 1964; Kreps, 1979). In a theory of flexibility, a potential preference ordering is an ordering which, ex ante, is admitted as possible; ex post, one of the set of potential preference orderings is realized. The theory of flexibility has some similarities with Harsanyi's analysis of imaginative empathy; the main difference is that the cross-preference comparisons in the theory of flexibility are intrapersonal rather than interpersonal. Versions of the theory that assume intrapersonal extended preferences seem to be open to the same objections as Harsanyi's theory of interpersonal extended preferences (see Section 10.2). More recently, measures of opportunity have been proposed based on an interpretation of the set of potential preference orderings as the set of those preferences that are "reasonable," or those that the relevant person might, counterfactually speaking, have had (Foster, 1993; Jones and Sugden, 1982; Pattanaik and Xu, 1998). In this chapter, I develop a Rawlsian interpretation of potential preferences.

10.4 Money Metrics

To explain my proposal, it is useful to proceed in stages. I begin with the money metric, as familiarly used in welfare economics. Suppose that there are private consumption goods $1, \ldots, g$ and public goods $1, \ldots, h$.

For the present, I focus on a given individual, with given (and well-behaved) preferences. A consumption bundle can be written as $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} := (x_1, \ldots, x_g)$ and $\mathbf{y} := (y_1, \ldots, y_h)$. The individual's preferences are represented by a utility function $u_i(\mathbf{x}, \mathbf{y})$; the subscript i refers to the individual in question. Let $c_i(u, \mathbf{p}, \mathbf{y})$ be the individual's *expenditure function*, where $\mathbf{p} := (p_1, \ldots, p_g)$ is a vector of private-good prices. That is, $c_i(u, \mathbf{p}, \mathbf{y})$ is the minimum expenditure on private goods that will allow her to achieve the utility level u, given that the vector of private-good prices is $\mathbf{p}$ and the vector of public goods is $\mathbf{y}$.

To construct a money metric, it is necessary to start with a vector of *reference prices*, $\mathbf{p}^* := (p_1^*, \ldots, p_g^*)$. In addition, we need a vector of *reference levels* of consumption of public goods, $\mathbf{y}^* := (y_1^*, \ldots, y_h^*)$. The conventional money metric is a function that assigns a money value to each consumption bundle. It is defined by

$$M_i(\mathbf{x}, \mathbf{y}) := c_i(u[\mathbf{x}, \mathbf{y}], \mathbf{p}^*, \mathbf{y}^*). \tag{10.1}$$

In effect, this function assigns a money value to each of the individual's indifference curves. Thus, the money metric is a particular representation of her preferences, that is, a particular utility function. This analysis can be applied to each of any number of individuals, and the resulting metric is in the same dimension – money – for all of them. Thus, we have arrived at an interpersonal common currency of personal preference. To avoid confusion with the money metrics that I shall consider later, I shall call $M_i(\mathbf{x}, \mathbf{y})$ a *personal money metric of consumption.*

It is straightforward to extend this approach so that money-metric values can be given to opportunity sets. Still focusing on a given individual i, consider any opportunity set S made up of $(\mathbf{x}, \mathbf{y})$ bundles. The *personal money metric of opportunity* is defined by:

$$\psi_i(S) := \max_{(\mathbf{x}, \mathbf{y}) \in S} M_i(\mathbf{x}, \mathbf{y}). \tag{10.2}$$

In other words, $\psi_i(S)$ is the indirect utility of S to i when i's utility is measured by the money metric $M_i(\mathbf{x}, \mathbf{y})$; it is a measure of the value of an opportunity set to a particular person, on the assumption that opportunity sets have only instrumental value as means to the satisfaction of preferences.

In using these money metrics, no claims are being made about the interpersonal comparability of mental states. These metrics are not based on any assumptions about mental states, apart from the assumption that each individual has his or her own (well-behaved) preferences over consumption bundles. Nor is there any attempt to compare what Rawls calls people's "total circumstances." For example, suppose that Joe is consuming the bundle of goods $(\mathbf{x}_1, \mathbf{y})$ and that Jane is consuming $(\mathbf{x}_2, \mathbf{y})$; and suppose that the money metric happens to assign the same value m' to each person's consumption. That does not mean that Joe and Jane are experiencing mental states that are in some way equivalent nor that it is equally good to be in Joe's circumstances as to be in Jane's.

So what *does* it mean? Let $B(m, \mathbf{p}, \mathbf{y})$ denote the opportunity set for an individual whose public-good consumption is $\mathbf{y}$ and who is free to spend up to m on private goods, which can be bought at the prices $\mathbf{p}$. (That is, $B(m, \mathbf{p}, \mathbf{y})$ contains all those bundles $(\mathbf{x}, \mathbf{y})$ such that $\mathbf{p} \cdot \mathbf{x} \leq m$.) We can infer that Joe is indifferent between the particular consumption bundle $(\mathbf{x}_1, \mathbf{y})$ and the opportunity to choose the bundle he most prefers from the set $B(m', \mathbf{p}^*, \mathbf{y}^*)$. Similarly, Jane is indifferent between $(\mathbf{x}_2, \mathbf{y})$ and the opportunity to choose the bundle she most prefers from $B(m', \mathbf{p}^*, \mathbf{y}^*)$. Thus, we can point to a particular opportunity set which in terms of each person's preferences is equivalent to his or her actual consumption bundle. By selecting particular reference vectors $\mathbf{p}^*$ and $\mathbf{y}^*$ of prices and public goods, we are selecting a particular family of *standard* opportunity sets – that is, the family of all $B(m, \mathbf{p}^*, \mathbf{y}^*)$ with $\mathbf{p}^*$ and $\mathbf{y}^*$ fixed – to use as the basis for such comparisons. In this way, a money metric expresses equivalences, from the point of view of the relevant consumer, between particular consumption bundles (or particular opportunity sets) and members of the family of standard opportunity sets.

Equivalences of this kind are significant if we can interpret standard opportunity sets in something like the way that Rawls interprets income. Whatever a person's plan of life, we may say, a standard opportunity set gives her the power to buy things that have use within that plan; and the larger the standard opportunity set, the more such power she is given. Thus, standard opportunity sets are things that everyone can be presumed to want; and everyone can be presumed to want larger such sets in preference to smaller ones. In this sense, standard opportunity sets seem to provide the right sort of metric to be used in a Rawlsian theory of justice.

An example may help to explain this argument. Suppose that for some person, the money-metric value of his consumption bundle is much higher than the average for the society in which he lives. However, that bundle is made up of goods most people would not want. Suppose the person is Ranulph, whose favourite – and very expensive – leisure pursuit is manhauling a sledge across Antarctica. Say that the money-metric value of annual consumption is £10,000 for an average person in the society, and £500,000 for Ranulph. It is difficult for us to know from looking at Ranulph's actual consumption bundle how desirable that bundle is to him. After all, it is far from desirable to us (or, at any rate, to me). We learn much more when we discover that he regards his actual consumption as equally preferable as a purchasing power of £500,000 per year, since *that* is something we all want. What is going on here is not imaginative empathy: we are not simulating the mental experiences that Ranulph enjoys as he drags his sledge across the icefields. Nor are we assessing the value of a life of Antarctic travel in terms of some universal theory of what makes for a good life. We are simply noting that Ranulph values Antarctic sledging very highly relative to opportunities that everyone can be presumed to want.

To all this, it may be objected that standard opportunity sets are defined in terms of particular reference levels of prices and public-good consumption. Unless we have a theory that picks out and justifies a unique set of reference levels, the money metric is arbitrary. That arbitrariness might be construed as a lack of neutrality between conceptions of the good. (For example, Ranulph might complain that, by using current market prices as our reference prices,we have built in a bias against his preferred way of life. It is only because the goods that Antarctic sledging requires are so high priced that the money-metric value of his consumption bundle is so large. Why not use reference prices that make Antarctic sledging cheap and, say, watching network television expensive?)

The reference levels for a money metric *are* to some degree arbitrary. That is the index-number problem again. However, the need for weights is an inescapable part of any income-based approach. Income is a one-dimensional index of purchasing power in a world in which there are many goods. To construct any such index, we must have some set of weights for the different goods that might be purchased. Perhaps the most we can ask is that the reference levels for our index are *salient*; that is, they have some power of suggestion that makes them appear obvious or natural. That thought is not completely foreign to Rawls's enterprise. Recall that Rawls explains his strategy of focusing on primary goods as a simple, workable method on which people who needed a public conception of justice might agree (see Section 10.3). In an ongoing society with a market economy, current market prices are, I suggest, highly salient; it is not implausible to suppose that there would be general agreement that if a one-dimensional public measure of consumption bundles is needed, it should be based on current market prices. Admittedly, this kind of local salience is not enough to underpin the grand comparisons that Rawls wants to make between societies with different "basic structures," but it may be sufficient for many of the more limited problems that welfare economists are concerned with.

Although it is true that reference prices may indirectly favour some conceptions of the good over others, the money-metric approach goes a long way toward the ideal of neutrality between such conceptions. Once we have settled on a family of standard opportunity sets, equivalences between particular consumption bundles and standard opportunity sets are established by using the preferences of the consumer in question. In this respect, all preferences are treated equally: there is no attempt to assess whether a person's preferences are in accord with some overarching account of the good life. Given equal willingness to pay, we might say, pushpin is as good as poetry.

10.5 An Impersonal Money-Metric of Opportunity

I have argued that the money-metric approach can achieve something of what Rawls intends to achieve through his use of the ill-defined primary good – income. However, the conventional money-metric approach is based on personal preferences. A personal money-metric value of a consumption bundle tells us which of a family of standard opportunity sets is just as desirable as that bundle, *as evaluated by the preferences of a particular person.* In contrast, it is central to Rawls's primary-goods approach that the measurement of a person's opportunities is independent of that person's preferences: opportunities are the currency of justice, and claims of justice are not to be grounded on personal preferences. We need a money metric that is independent of personal preferences.

The conventional money-metric approach, I have said, measures the desirability of each consumption bundle (or each opportunity set) to a particular person, given her preferences. The standard of comparison is provided by opportunity sets presumed to be desirable to everyone, but the comparisons themselves are in terms of personal preferences. Now suppose we ask instead how desirable each consumption bundle is, not to any particular person, but *to people in general.* Recall that Rawls's definition of primary goods depends on some such concept of "desirability to people in general," or *impersonal desirability.* Where there is no danger of confusion, I shall simply use the unqualified term *desirability.*

Although economists do not typically use the concept of (impersonal) desirability, I think it is meaningful. In the sense in which I use the word, desirability is a *secondary property* of objects: an object is desirable to the extent that it has a general tendency to induce sentiments of desire in people. (This sense of "desirable" is value-neutral: there is no implication that a desirable object is a *worthy* object of desire.) Compare the property of yellowness. We say that an object is yellow if it has a general tendency to induce the sensation of "seeing it as yellow"; we do not require that this sensation is *always* induced. In normal light, for people with normal vision, a daffodil induces that sensation, but even in the dark, a daffodil is a yellow flower. Similarly, an E-class Mercedes is a desirable car. Not everyone desires an E-class Mercedes, but it is certainly an object with a general tendency to induce strong sentiments of desire.

The personal money metric of consumption measures the desirability of a consumption bundle to a particular person while maintaining neutrality with respect to different conceptions of the good. Recall that we infer that Antarctic sledging is desirable to Ranulph because he regards it as equivalent to other opportunities everyone can be presumed to want. Analogously, we can infer that an E-class Mercedes is impersonally desirable because people in general tend to regard it as equivalent to opportunities everyone can be presumed to want.

If we are to generate a money metric of impersonal desirability, we need a *reference distribution* of preferences to represent "preferences in general." To keep the notation simple, I shall assume that the set of conceivable options is finite, which implies that there is a finite number of logically possible orderings of options. I denote this set of possible orderings by $\mathcal{R} := \{R_1, \ldots, R_n\}$. Then the reference distribution is a function π^* from $\mathcal{R}$ to the set of non-negative real numbers such that $\Sigma_i \, \pi^*(R_i) = 1$.

A natural money-metric index of the desirability of $(\mathbf{x}, \mathbf{y})$, relative to the reference distribution of preferences, is given by

$$M(\mathbf{x}, \mathbf{y}) := \Sigma_i \, \pi^*(R_i) M_i(\mathbf{x}, \mathbf{y}). \tag{10.3}$$

I shall call this the *impersonal money metric of consumption.* The corresponding *impersonal money metric of opportunity* is given by

$$\psi(S) := \Sigma_i \, \pi^*(R_i) \psi_i(S). \tag{10.4}$$

I propose that we use an impersonal money metric of opportunity as the common currency of advantage.

In general, an impersonal money metric does not accord with individuals' preferences. Thus, if opportunity is defined by Eq. (10.4), we cannot say, as Rawls seems to want to say of primary goods, that increases in opportunity are *always* valuable, irrespective of a person's preferences (or rational plan of life). But if options are multidimensional and personal preferences vary, that is an inevitable consequence of the Rawlsian strategy of looking for a common currency of opportunity that is independent of personal preferences. Recall that Rawls runs into exactly the same problem when he aggregates all primary goods into a composite index (see Section 10.3).

A lot turns on the choice of the reference distribution of preferences. As with the choice of reference prices, it can be objected that this is where judgments of value are smuggled into the metric. We must ask what a reference distribution of preferences is.

Given that we are taking a broadly Rawlsian approach, what we are looking for is some conception of the range of preferences that a person might be expected to have, those expectations being formed behind some kind of veil of ignorance. That veil of ignorance should exclude knowledge of the person's conception of the good; thus, expectations should range over preferences associated with different such conceptions. However, the veil should not exclude sociological and psychological knowledge about the general properties of human preferences. Nor, I suggest, should it exclude knowledge about general associations between preferences and (at least) such broad "objective" variables as age, sex, and health status. Thus, in forming expectations about the preferences of a person, we are allowed to condition those expectations on such objective variables: our expectations about the preferences of a healthy 20-year-old woman need not be the same as our expectations about the preferences of a sick 80-year-old man. Which variables we may and may not condition such expectations on is, of course, a deep question for any theory of justice in the Rawlsian mould. Essentially, what is being decided here is which preferences a person should be held responsible for (recall the discussion of expensive tastes in Section 10.3). Any variation in preferences that is not accounted for by the variables on which expectations are conditioned is being treated as a matter of individual responsibility. I have nothing to add to the debate on expensive tastes and responsibility: I merely note that the money-metric approach is compatible with a range of different positions that might be taken on these questions.

Just as current market prices are salient when we are looking for agreement on a vector of reference prices, so (I suggest) the current frequency distribution of actual preferences is salient when we are looking for a reference distribution of preferences. If, for example, we ask what preferences 80-year-old men can be expected to have, it is surely natural to look at the actual preferences of 80-year-old men. I realize that this suggestion is not fully in the spirit of Rawls's *A Theory of Justice* because it makes no use of the "rational" part of Rawls's notion of a "rational plan of life." Instead of ranging over the plans of life that persons might rationally form, my proposal ranges over the desires that persons might in fact have, rationally or not. In this respect, my proposal is Humean while Rawls is more Kantian. Recall, however, that Rawls deliberately avoids building contestable claims about the rationality of particular plans of life into the central structure of his theory. By focusing on primary goods, he needs to claim only that all rational plans of life, whatever they may be, are promoted by those goods. Thus, Rawls's theory makes little use of the concept of rationality as applied to the evaluation of preferences or plans of life. My suggestion, then, is that we construct reference distributions of preferences by defining *reference classes* of individuals, such that any preference differences within a reference class are deemed to be matters of personal responsibility. Then the frequency distribution of preferences in any given reference class is taken to be the reference distribution of preferences for each individual member of the class.

The upshot of all this is that my common currency of advantage is based on preferences, but in an impersonal way. In assessing any individual's opportunities, we refer to the overall distribution of preferences in the population of people who are objectively like her, but not to her personal preferences. Thus, no individual can be said to be advantaged or disadvantaged by virtue of her preferences.

10.6 Conclusion

This chapter grappled with one of the central problems of welfare economics – the problem of finding a common currency of advantage. I have argued in favour of Rawls's general strategy of focusing on primary goods. Working within that strategy, I have developed a measure of the desirability of opportunity sets, which reflects the ability of those sets to satisfy the range of preferences that people can be expected to have. Like all income-based measures, my measure of opportunity has the limitation that it requires a vector of reference prices to be taken as given. However, this seems to be the price we have to pay for a measure of opportunity that is sensitive to information about preferences in general but independent of personal preferences.

References

Anand, P. 1993. *Foundations of Rational Choice Under Risk.* Clarendon Press, Oxford.

Arneson, R. 1989. Equality of opportunity for welfare. *Philosophical Studies* 56, 77–93.

Arrow, K. J. 1995. A note on freedom and flexibility. In *Choice, Welfare, and Development: A Festschrift in Honour of Amartya K. Sen,* ed. K. Basu, P. Pattanaik, and K. Suzumura. Clarendon Press, Oxford, pp. 7–16.

Barry, B. 1996. Survey article: Real freedom and basic income. *Journal of Political Philosophy* 4, 242–277.

Bossert, W., and Weymark, J. A. 2004. Utility in social choice. In *Handbook of Utility Theory,* Vol. 2, *Extensions*, ed. S. Barberà, P. Hammond, and C. Seidl. Kluwer, Boston, pp. 1099–1177.

Broome, J. 1991. *Weighing Goods*. Blackwell, Oxford.

Broome, J. 1993. A cause of preference is not an object of preference. *Social Choice and Welfare* 10, 57–68.

Cohen, G. A. 1989. On the currency of egalitarian justice. *Ethics* 99, 906–944.

Dworkin, R. 1981a. What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs* 10, 185–246.

Dworkin, R. 1981b. What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs* 10, 283–345.

Edgeworth, F. Y. 1881. *Mathematical Psychics*. Kegan Paul, London.

Fisher, I. 1892/1925. *Mathematical Investigations in the Theory of Value and Prices*. Yale University Press, New Haven, CT. First published 1892.

Foster, J. 1993. Notes on effective freedom. Mimeo, Vanderbilt University.

Gibbard, A. 1979. Disparate goods and Rawls's difference principle: A social choice theoretic treatment. *Theory and Decision* 11, 267–288.

Griffin, J. 1986. *Well-Being: Its Meaning, Measurement and Moral Importance*. Clarendon Press, Oxford.

Griffin, J. 1991. Against the taste model. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, pp. 45–69.

Harsanyi, J. C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61, 434–435.

Harsanyi, J. C. 1968. Games with incomplete information played by "Bayesian" players. *Management Science* 14, 159–182, 320–334.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge.

Harsanyi, J. C. 1982. Morality and the theory of rational behaviour. In *Utilitarianism and Beyond*, ed. A. K. Sen and B. Williams. Cambridge University Press, Cambridge, pp. 39–62.

Hausman, D. M., and McPherson, M. S. 1996. *Economic Analysis and Moral Philosophy*. Cambridge University Press, Cambridge.

Jevons, W. S. 1871/1970. *The Theory of Political Economy*. Penguin, Harmondsworth. First published 1871.

Jones, P., and Sugden, R. 1982. Evaluating choice. *International Review of Law and Economics* 2, 47–65.

Koopmans, T. C. 1964. On the flexibility of future preferences. In *Human Judgments and Optimality*, ed. M. W. Shelley and J. L. Bryans. Wiley, New York, pp. 243–254.

Kreps, D. M. 1979. A representation theorem for "preference for flexibility." *Econometrica* 47, 565–577.

Marshall, A. 1890/1920. *Principles of Economics*. Macmillan, London. First published 1890.

McKenzie, G. W. 1983. *Measuring Economic Welfare: New Methods*. Cambridge University Press, Cambridge.

Pattanaik, P., and Xu, Y. 1998. On preference and freedom. *Theory and Decision* 44, 173–198.

Pareto, V. 1906/1972. *Manual of Political Economy*. Trans. A. S. Schweir. Macmillan, London. First published in Italian in 1906.

Plott, C. 1978. Rawls's theory of justice: An impossibility result. In *Decision Theory and Social Ethics*, ed. H. W. Gottinger and W. Leinfellner. Reidel, Dordrecht, pp. 201–214.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA.

Rawls, J. 1982. Social unity and primary goods. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge University Press, Cambridge, pp. 159–185.

Rawls, J. 1985. Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs* 14, 223–251.

Roemer, J. E. 1996. *Theories of Distributive Justice*. Harvard University Press, Cambridge, MA.

Samuelson, P. A. 1947. *Foundations of Economic Analysis*. Harvard University Press, Cambridge, MA.

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

Scanlon, T. M. 1991. The moral basis of interpersonal comparisons. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, pp. 17–44.

Sen, A. 1992. *Inequality Reexamined*. Harvard University Press, Cambridge, MA.

Sugden, R. 1989. Review of *Well-Being: Its Meaning, Measurement and Moral Importance* by James Griffin. *Economics and Philosophy* 5, 103–108.

Sugden, R. 1991. Rational choice: A survey of contributions from economics and philosophy. *Economic Journal* 101, 751–785.

Van Parijs, P. 1995. *Real Freedom for All: What (If Anything) Is Wrong with Capitalism*? Clarendon Press, Oxford.

Walras, L. 1889/1954. *Elements of Pure Economics*. Trans. by W. Jaffé. Allen and Unwin, London. First published in French 1889.

11

Utilitarianism versus Fairness in Welfare Economics

Marc Fleurbaey and François Maniquet

11.1 Introduction

Utilitarianism has been opposed to theories of fairness (especially Rawls's theory) in many respects. We want to focus here on a particular division that has been seldom discussed, although it is reflected in the structure of welfare economics. Welfare economics is indeed currently separated into two very different branches. One branch deals with social welfare functions and devotes a substantial energy to the study of utilitarianism. The other studies fair allocation in economic models and, formally, its main focus is on allocation rules. The difference is the following. A social welfare function associates each member in a class of possible contexts with a ranking of all possible alternatives, whereas an allocation rule only associates each member in the class with a selection of "best" alternatives. As it has long since been noted in the theory of social choice, an allocation rule is a kind of ranking, albeit simple (any two selected allocations as well as any two nonselected allocations being deemed socially indifferent), and a ranking immediately leads to an allocation rule (which selects the best alternative in every context).

Actually, there is a second important difference between these two branches. The arguments of the social welfare functions studied by the former are interpersonally comparable utilities (usually comparable in levels, differences, or ratios), whereas the whole body of literature representing the latter is purely ordinal, making use of no other welfare information than the preferences of the agents over simple alternatives.[1]

[1] Excellent recent summaries of the two branches are available in Mongin and d'Aspremont (1998) for the former and Thomson (2008) and Moulin and Thomson (1997) for the latter.

We would like to thank a reviewer and the editors of this volume for many valuable comments and suggestions on an earlier version.

These two branches are not separated because of a theoretical opposition on some deep issue. On the contrary, each school would be happy to share some feature with the other one. The fairness theorists regret that they are unable to compare unfair allocations, whereas some specialists in the other branch would be happy to get rid of the burden of interpersonal comparisons of utility, although most of them think that they cannot avoid such comparisons. We think that these dual drawbacks deprive each of these two branches of most of its practical relevance for policy issues. The policy maker is obviously extremely reluctant to engage in interpersonal comparisons of such impalpable objects as utilities, which makes social welfare functions look like wonderful machines that just lack the appropriate fuel. We shall argue below that such reluctance is ethically sound. However, first best efficient allocation rules are irrelevant to the analysis of piecemeal reforms in an imperfect world. The only way in which policy recommendations have been able to appear in this branch is through sophisticated implementation games that are often considered as requiring too much information and intelligence from the agents. Such games, and the corresponding allocation rules, resemble wonderful machines that only go on roads that do not exist.

What we want to do is simply to study purely ordinal social orderings. If we could find nice objects of this sort, then we would be able to compare any pair of allocations without relying on utilities and interpersonal comparisons of utilities. Knowledge of the agents' preferences would be enough. We do not, however, want to suggest that such tools could be directly applied, and in particular that revelation problems would be easily solved. But we will argue that ordinal social orderings would be extremely useful tools in some contexts.

The definition of purely ordinal social orderings is of course far from being new because it is the goal of Arrovian social choice theory. The almost exclusively negative results that were obtained in this approach merely reinforced the feeling among theorists that interpersonal comparisons of utility were the price to pay for reasonable social orderings. We think that Arrow's approach was too demanding in two respects.

First, it considered abstract spaces of alternatives and pretended to solve all aggregation problems at the same time without taking into account specific features of economic allocations. On the contrary, we argue that the economic context (private or public goods, distribution or production, goods or bads, etc.) and the particular features of the alternatives may significantly alter the social judgment, and rightly so, since there are ethical principles that are context-dependent, at least in their precise application.[2] For

[2] See Moulin (1990).

instance, depending on the returns to scale or the rivalry of consumption, the stand-alone principle either says that an agent should not be better off, or worse off, than if she were alone in the economy. Another example is provided by the no veto power requirement. It is legitimately imposed on voting procedures for a large number of people. In the distribution problem of a single commodity whose consumption may lead to satiation, however, applying no veto power when all agents are identical may lead us to give their preferred consumption level to all but one agent, and her worst consumption level to the last one, thereby violating the basic horizontal equity requirement.

The second excessive requisite in Arrow's approach was the axiom of Independence of Irrelevant Alternatives. This axiom requires the social ranking of two alternatives to depend only on how the agents rank these two alternatives. We will argue below that even if this axiom has nice justifications, there is no reason to impose it at all cost, and we will show how much it must be weakened to open the way to ordinal social orderings. In conclusion, turning to economic environments and weakening the axiom of Independence is the way in which we propose to avoid the Arrovian impossibilities.[3]

Ordinal social orderings have recently been studied by Maniquet (1994) and more specifically by Bossert, Fleurbaey, and Van de gaer (1999), in a quite general framework where the agents' characteristics may be anything. We focus here on the basic case where agents differ only in their preferences.

The first section in this chapter presents arguments in favor of ordinal social orderings. In Subsection 11.2.1, we argue in favor of ordinalism and against the idea that interpersonal comparisons of utility are unavoidable. In Subsection 11.2.2, we discuss the relative merits of social orderings and allocation rules. The main topic is tackled in Section 11.3, where examples of ordinal social orderings are studied in the restricted framework of division economies. In a first subsection, we introduce properties which can be directly used to select among plausible social ordering functions. A social ordering function associates a social ordering to each economy in a class of admissible economies. In particular, we propose an alternative to Arrow's Independence of Irrelevant Alternatives axiom that can be combined quite easily with other axioms. In a second and last subsection, we examine how to derive social ordering functions from allocation rules. Our main result is

[3] The theory of social choice in economic environments (reviewed in Donaldson and Weymark, 1988; and Le Breton, 1997) comes close to our project but has mostly retained the axiom of Independence and the negative flavor of results that ensue.

the proof that the three major solutions to the fair division problem (that is, the Fixed Numeraire Egalitarian Equivalent rule, the Pazner–Schmeidler Egalitarian Equivalent rule, and the Equal Income Walrasian rule) can be rationalized by social ordering functions satisfying basic desirable properties (that is, Weak Pareto, Pareto Indifference, and Anonymity) and the independence axiom we introduce in the preceding subsection. Concluding comments are given in Section 11.4.

11.2 Justifying Social Orderings

11.2.1 Ordinalism versus Interpersonal Comparisons of Utility

In the layman's (possibly a policy maker's or at least a voter's) reluctance to try and make interpersonal comparisons of utility in distributive issues beyond the family circle, one can decipher a mixture of two objections: One cannot make such comparisons and one should not. The former is familiar to economists, the latter has been developed by philosophers and in particular by Rawls.

That consumers' demand behavior reveals ordinal preferences only is well known and implies that interpersonal comparisons require additional information that may not be easily collected. This is the main reason why economists have problems with interpersonal utility comparisons, and we think that this provides sufficient justification for our ordinal approach. But even if utility information could be gathered costlessly, there are ethical arguments against using it. In Rawls's (1971, 1982), Dworkin's (1981) and Van Parijs's (1995) theories of justice, individuals should assume responsibility for their ends, preferences, and satisfaction levels. The idea is that there would be some contradiction in treating individuals as morally autonomous in the formation of their life plans and, at the same time, redistributing resources so as to guarantee that they reach some utility level, given that utility functions are viewed as part of one's life plans. The general problem of distributive justice is to allocate resources to individuals who will use them as they wish, not to allocate utility. Also, the local problem of distributive justice, which we address in Section 11.3, that is, the division of unproduced commodities, is to allocate goods in an equal way. In either case, a pure utility deficit does not give a valid claim against others.

This argument is quite convincing because it essentially makes explicit what is currently enacted in liberal societies, where respect for the individual's private sphere entails such a division of labor between social institutions and individual initiative.[4] The implications of this argument are far-reaching because they support the idea that the allocation of resources should be independent of individuals' preferences and life plans, but one must be careful. It would certainly be excessive to propose making the allocation of goods independent of preferences, and the philosophers' theories do not have such an implication. In Rawls's theory, only primary goods (which are essentially all-purpose goods) should be distributed independently of preferences, and in Dworkin's proposal as well as Van Parijs's proposal, the distribution of income only is discussed, so that in all theories, the market mechanism determines the final allocation of goods, in a way that is obviously dependent on preferences.

Here we follow the usual practice in economic theory, which is to assume nothing about institutions from the outset. We look for orderings of allocations, without assuming the market to have a special role. These arguments, in our opinion, justify that we rely on ordinal noncomparable information only. Moreover, we will retain the idea that the allocation of goods (or the social ordering) should be as independent of preferences as possible but still compatible with the Pareto principle. The idea that the Pareto principle represents the minimal requirement about how much preferences must matter is standard and quite natural because going against unanimous preference seems to be the strongest way in which individuals' preferences can be disregarded. As a result, the preference independence idea will take the form of independence axioms with respect to *changes* in preferences.

An often raised objection to our approach must be met before proceeding. This objection could be called the *planner's ethical preferences revelation principle*. It is, indeed, often contended that, willy-nilly, interpersonal comparisons of utility are *always* made in allocation decisions. We strongly oppose this claim. The rule that a cake must be cut in equal pieces, for instance, the ordinal allocation rules of the fair allocation literature (e.g., the Equal Income Walrasian allocation rule), and the ordinal social orderings that we study below do not involve any interpersonal comparisons of utility or welfare. This planner's ethical preferences revelation principle originates in the fact that the same decision rule can usually be rationalized by several ethical principles. For instance, the market mechanism in an Arrow-Debreu world

[4] This approach has, however, been challenged by other philosophers (Arneson, 1989; Cohen, 1989) who claim that people should be held responsible only for what is under their control, so that a pure utility deficit that does not derive from individual choice should be deemed a valid claim for a resource transfer. For a criticism, see Fleurbaey (1995).

may be chosen either by a libertarian planner or by a utilitarian planner using appropriate utility functions. It seems clear to us that this does not mean that the libertarian planner makes interpersonal comparisons of utility. It is true that decision rules based on ordinal considerations can be, ex post, rationalized as a utilitarian or any other welfarist rule for an appropriate cardinalization of preferences. But this fact hardly proves that ordinalists are necessarily engaged in (even implicit) utility comparisons.

11.2.2 Social Rankings versus Allocation Rules

A few notations and assumptions will be useful. An economy e is defined by a population $N = \{1, \ldots, n\}$, a profile of characteristics $\theta_N = (\theta_1, \ldots, \theta_n)$, and a set Z of feasible allocations:

$$e = (\theta_N, Z).$$

A typical allocation in Z is denoted $z_N = (z_1, \ldots, z_n)$, where z_i is agent i's bundle. In our applications in this chapter, the only characteristics that describe the agents are their self-centered preferences: $\theta_N = R_N = (R_1, \ldots, R_n)$. For each agent i $(i \in N)$, R_i is a complete ordering (with strict preference P_i and indifference I_i) over some consumption set X such that $Z \subset X^n$. We also assume that X is always a subset of a finite-dimensional Euclidean space.

An *allocation rule* is a correspondence S that selects in each economy of some domain $\mathcal{D}^S$ a subset of feasible allocations:

$$S\colon e \mapsto S(e) \subset Z, \forall\, e \in \mathcal{D}^S.$$

A *social ordering function* is a function R that defines for each economy of some domain $\mathcal{E}^R$ a complete ordering over its feasible allocations:

$$\mathsf{R}\colon e \mapsto \mathsf{R}(e) \text{ complete ordering over } Z.$$

Let **R** denote the class of all admissible social ordering functions. For a given social ordering $\mathsf{R}(e)$, the related strict preference relation will be denoted $\mathsf{P}(e)$ and the indifference relation $\mathsf{I}(e)$.

In this subsection, we take for granted that the planner may face information, incentive, or observation constraints in such a way that the set of attainable allocations turns out to be a strict subset of Z. A first best problem is defined as a problem wherein the planner has the opportunity to make the economy reach any allocation in Z. If this is not the case, then the problem is called a *second best problem.* Let $\mathcal{Y} \subseteq 2^Z$ denote the family of plausible feasible sets in which the planner may have to search for an optimal policy.

In this second best setup, we would like to recall some reasons why social orderings should be preferred to allocation rules.

First, it should be clear that, provided Z is compact, a social ordering function always gives rise to an allocation rule (defined as the selection of allocations socially preferred to all the other allocations). Therefore, if the problem to solve turns out to be a first best problem, then a social ordering function is as useful as an allocation rule.

On the contrary, an allocation rule may prove insufficient to solve all the plausible second best problems in a satisfactory way. There are two reasons for that. The first reason is that some problem (R_N, Y) may not be in the domain of the allocation rule. In other words, the selection may be empty. The second reason is that the selections operated by an allocation rule may be inconsistent in the sense that an allocation selected for a given problem $Y \in \mathcal{Y}$ is not necessarily selected for smaller problems $Y' \subset Y$ containing this allocation.

These two problems, however, can be overcome by imposing nonemptiness and contraction consistency requirements on the allocation rules. But it is well known that these two requirements together are almost equivalent to requiring that the allocation rule be consistent with some social ordering function (provided $\mathcal{Y}$ is sufficiently rich), in the sense that, for any problem, it selects the allocations that are considered as socially better by some complete ranking of the allocations in Z.[5]

Therefore, as a social ordering is a natural tool in the second best context, our opinion is that the search for social ordering functions should be the focus of welfare economics, even if allocation rules may be sufficient in the first best context when the available information enables the planner to select any allocation.

11.3 Constructing Social Orderings

Now, we consider that the construction of either a social ordering or an allocation rule should be made on the basis of the properties they satisfy. In other words, it should be axiomatic.[6] In this section, we investigate how to construct social ordering functions in simple division economies. We restrict ourselves to division economies for two reasons. First, we simply need to illustrate the approach, and simple examples are therefore sufficient. Second,

[5] We refer here to Arrow's (1959) proof that condition C4 on a choice function implies that the associated binary relation of social preference is transitive.

[6] There are examples in the literature, however, of nonaxiomatic constructions of allocation rules or social rankings of allocations (e.g., Harsanyi, 1977). For a powerful statement that rules and rankings should be derived axiomatically, see Thomson (2001).

it is a major feature of the axiomatic approach that its conclusions are context dependent; that is, the possibilities to combine axioms vary from one model to another. Consequently, we have to begin our inquiry in a specific model, and the model of division economies is among the best-known models. At the end of this section, we briefly comment on whether the results we obtain can be adapted to other models.

Let the set of feasible allocations be defined with respect to a total amount $\Omega \in \mathbb{R}^l_+$ of goods and be denoted $Z(\Omega) \in \mathbb{R}^{ln}_+$; that is, $z_N = (z_1, \ldots, z_n) \in Z(\Omega) \Leftrightarrow \sum_{i=1}^n z_i \leq \Omega$. Each agent is now equipped with continuous, strictly monotonic, and convex preferences over her consumption set $\mathbb{R}^l_+$. Let $\mathcal{R}$ denote the set of all admissible preferences. An *economy* e can be described by a list $(R_1, \ldots, R_n, \Omega) \in \mathcal{R}^n \times \mathbb{R}^l_+$. Let $\mathcal{E}$ denote the class of admissible economies. An allocation $z_N = (z_1, \ldots, z_n)$ is *Pareto efficient for the economy* $e = (R_N, \Omega) \in \mathcal{E}$ if for all $z'_N = (z'_1, \ldots, z'_n) \in Z(\Omega)$, $[z'_i \; R_i \; z_i, \forall \, i \in N] \Rightarrow [z'_i \; I_i \; z_i, \forall \, i \in N]$. Let $P(e) \subset Z(\Omega)$ denote the set of Pareto efficient allocations for e.

Allocation rules and social ordering functions are defined as in the previous section. Their domains are denoted $\mathcal{E}^S$ and $\mathcal{E}^R$, respectively.

There are two ways of constructing ordinal social orderings, that is, the direct inquiry and the rationalization of given allocation rules. Since we focus in this chapter on the relationship between the social welfare and the fair allocation approaches, we would like to emphasize the latter way of constructing social ordering functions. It will prove useful, however, to begin with a few remarks on the direct inquiry of social ordering functions.

11.3.1 Direct Inquiry

The first way of constructing social orderings is by defining axioms directly bearing on these orderings and trying to combine them. In this subsection, we will give some examples of such axioms, all inspired by the Arrovian axioms.

Let us first define Weak Pareto, Pareto Indifference, and Anonymity, which we consider as essential in our inquiry. Weak Pareto requires that an allocation be deemed socially preferred to another as soon as all the agents strictly prefer the bundle they get in the former to the bundle they get in the latter.

Weak Pareto:

$$\forall \, e = (R_N, \Omega) \in \mathcal{E}^R, \forall z_N, z'_N \in Z(\Omega),$$
$$[\forall i \in N, z_i \; P_i \; z'_i \Rightarrow z_N \, \mathsf{P}(e) \, z'_N].$$

Pareto Indifference requires that two allocations considered equivalent by all agents be also considered equivalent by society.

Pareto Indifference:

$$\forall e = (R_N, \Omega) \in \mathcal{E}^R, \forall z_N, z'_N \in Z(\Omega),$$
$$[\forall\, i \in N,\, z_i\; I_i\; z'_i \Rightarrow z_N\, \mathsf{I}(e)\, z'_N].$$

Anonymity requires that the name of the agents do not influence the social ranking. The definition of this property requires the following notation. Let $\pi : N \to N$ denote a permutation of elements of N, and let Π denote the set of all permutations from N to itself. For any n-dimensional list a_N, $\pi(a_N)$ denotes the list obtained by permuting elements of a_N according to π.

Anonymity:

$$\forall e = (R_N, \Omega) \in \mathcal{E}^R, \forall \pi \in \Pi, \forall z_N, z'_N \in Z(\Omega),$$
$$z_N\, \mathsf{R}(e)\, z'_N \Leftrightarrow \pi(z_N)\mathsf{R}((\pi(R_N), \Omega))\pi(z'_N).$$

Our three next examples are derived from Arrow's Independence of Irrelevant Alternatives axiom, in favor of which we would like to argue now. Arrow's Independence of Irrelevant Alternatives states that the social ranking of two alternatives should be independent of changes in individual preferences, provided these changes do not affect the way in which individuals rank these two allocations (or, more precisely, the bundles they get in these two allocations). This axiom is often justified on grounds of simplicity, but we think that it has an immediate value with respect to the responsibility idea discussed in Section 11.2.[7] If one wants to have the social ordering as independent as possible (under a constraint of compatibility with the Pareto conditions) from individual preferences, such an axiom goes a long way in that direction.

Actually, it goes too far, as the bulk of the social choice literature has uncovered. Therefore, one must look for independence axioms that can be combined with at least the three requirements previously listed. Here are two such axioms. The axiom of Extended Interval Independence requires that the ranking of two allocations remains unaffected by changes in preferences having the double property that the socially preferred allocation increases in

[7] Arrow's Independence of Irrelevant Alternatives is also justified on grounds of robustness to changes in the choice set, if one thinks of the derived choice rule. See Arrow (1963).

the individual ranking of all agents, and the other allocation decreases in the individual ranking of all agents (we have assumed that agents are interested only in their own consumption, so that by saying that an agent prefers an allocation to another we mean that she prefers the bundle she is assigned in the first allocation to the one she is assigned in the second allocation). Formally,

Extended Interval Independence:

$$\forall e = (R_N, \Omega),\, e' = (R'_N, \Omega) \in \mathcal{E}^R,\, \forall z_N, z'_N \in Z(\Omega),$$
$$[\forall\, i \in N, \forall z \in \mathbb{R}^l_+,\, z_i\ R_i\ z \Rightarrow z_i\ R'_i\ z \text{ and } z\ R_i\ z'_i \Rightarrow z\ R'_i\ z'_i] \Rightarrow$$
$$[z_N\, \mathsf{R}(e)\, z'_N \Rightarrow z_N\, \mathsf{R}(e')\, z'_N].$$

Examples in the next subsection will show that, in contrast with Arrow's Independence of Irrelevant Alternatives, Extended Interval Independence is easily combined with Weak Pareto, Pareto Indifference, and Anonymity.[8] Let us observe that this axiom is *not* logically weaker than Arrow's axiom since a reversal in an agent's ranking of the two allocations is allowed.[9] It seems to us that Extended Interval Independence is one of the most reasonable axioms that can express the idea that social preferences should be as independent from individual preferences as possible.

Our second axiom, called Unchanged Contour Independence, is weaker than Extended Interval Independence, and also weaker than Arrow's Independence of Irrelevant Alternatives. Unchanged Contour Independence requires that the ranking of two allocations be independent of changes in preferences having the property of leaving unaffected the contours (that is, the upper contour set, the indifference curve, and the lower contour set) of each agent at the two allocations. Formally,

[8] An idea that is related to that kind of independence with respect to preferences is to rank distributions of opportunity sets (see Kranich (1996), Herrero (1996), Herrero, Iturbe, and Nieto (1998), Bossert, Fleurbaey, and Van de gaer (1999) for examples of such rankings). In our framework, one might want to rank allocations on the basis of the distributions of opportunity sets that might have led to them. In economic domains, one may think in particular of ranking distributions of (not necessarily parallel) budget sets. But given the fact that individual indirect preferences over budget sets contain as much information as direct preferences over bundles (see e.g. Blackorby, Primont, and Russell (1978)), the Pareto conditions obviously require making the social ranking of distributions of budget sets depend on individual preferences, and then there is little independence from individual preferences to gain along these lines.

[9] We thank M. C. Sanchez for having pointed this fact out to us.

Unchanged Contour Independence:

$$\forall e = (R_N, \Omega), e' = (R'_N, \Omega) \in \mathcal{E}^R, \forall z_N, z'_N \in Z(\Omega),$$
$$[\forall\, i \in N, \forall z \in \mathbb{R}^l_+, z_i\, I_i\, z \Rightarrow z_i\, I'_i\, z \text{ and } z'_i\, I_i\, z \Rightarrow z'_i\, I'_i\, z] \Rightarrow$$
$$[z_N\, \mathsf{R}(e)\, z'_N \Rightarrow z_N\, \mathsf{R}(e')\, z'_N].$$

The last axiom says that a social ordering should only depend on the agents' preferences, and not on the set of allocations which are feasible. Actually, such a requirement is implicit in the Arrovian framework when the set of alternatives is viewed as the largest possible set of conceivable social states, whereas the social choice may take place among elements of a strict subset of this large set.

Independence of the Feasible Set:

$$\forall e = (R_N, \Omega), e' = (R_N, \Omega') \in \mathcal{E}, \forall z_N, z'_N \in Z(\Omega) \cap Z(\Omega'),$$
$$z_N\, \mathsf{R}(e)\, z'_N \Leftrightarrow z_N\, \mathsf{R}(e')\, z'_N.$$

We have restricted ourselves to axioms inspired by the Arrovian social choice theory. We believe though that economic domains give us the possibility to express more specific ethical principles. We do not develop this line of research here. But to conclude, we simply say that the analysis of social ordering functions in economic domains by direct axiomatic inquiry seems to us an urgent task.

11.3.2 Rationalizing Allocation Rules

In this subsection, we study how social ordering functions can be constructed when desirable allocation rules have been identified, and one would like to extend them into fine-grained social orderings. In other words, given an allocation rule S defined on $\mathcal{E}^S$, we look for a social ordering function $\mathsf{R}^S \in \mathbf{R}$ that rationalizes it in the sense that the allocations in $S(e)$ must always be top ranked by the social ordering $\mathsf{R}^S(e)$.

Rationalization of S:

$$\forall\, e \in \mathcal{E}^S,$$
$$S(e) = \{z_N \in Z(\Omega) \mid \forall\, z'_N \in Z(\Omega), z_N\, \mathsf{R}^S(e)\, z'_N\}.$$

Clearly, an allocation rule can always be rationalized by a two class social ordering function where selected allocations are socially indifferent to each other and form the first class, and nonselected allocations are also socially indifferent to each other and form the second class. But the resulting social orderings typically fail to satisfy even the three basic properties discussed at the beginning of the previous subsection (Weak Pareto, Pareto Indifference, and Anonymity).

Theorem 11.1 proves that the three major solutions to the fair division problem (see Moulin, 1990) can be rationalized by social ordering functions satisfying the three basic properties as well as Extended Interval Independence, and, in one of the three cases, Independence of the Feasible Set.[10] These allocation rules are the Fixed Numeraire Egalitarian Equivalent rule, the Pazner-Schmeidler Egalitarian Equivalent rule, and the Equal Income Walrasian rule.

The Fixed Numeraire z^* Egalitarian equivalent rule selects all the Pareto efficient allocations that have the property that each agent is indifferent between her bundle and a reference bundle defined as a multiple of the numeraire z^*.

The Fixed Numeraire z^* Egalitarian Equivalent rule E_{z^*}:

$$\forall\, e = (R_N, \Omega) \in \mathcal{E},\, z_N = (z_1, \ldots, z_n) \in P(e),$$

$$z_N \in E_{z^*}(e) \Leftrightarrow \exists\, \lambda \in \mathbb{R}_+ \text{ s.t. } z_i\ I_i\ \lambda z^*, \forall\, i \in N.$$

The second allocation rule is the Pazner-Schmeidler Egalitarian Equivalent rule E_Ω. Rather than taking a fixed numeraire z^*, it takes the reference bundle proportional to Ω.

The Pazner-Schmeidler Egalitarian Equivalent rule E_Ω:

$$\forall\, e = (R_N, \Omega) \in \mathcal{E},\, z_N = (z_1, \ldots, z_n) \in P(e),$$

$$z_N \in E_\Omega(e) \Leftrightarrow \exists\, \lambda \in \mathbb{R}_+ \text{ s.t. } z_i\ I_i\ \lambda\Omega, \forall\, i \in N.$$

The third rule is the Equal Income Walrasian rule W.

[10] We disregard the Strong Pareto condition in this chapter only for simplicity. Additional results relative to this condition are readily obtained, and are left here to the reader.

Equal Income Walrasian rule W

$$\forall\, e = (R_N, \Omega) \in \mathcal{E},\, z_N \in P(e),$$

$$[z_N \in W(e)] \Leftrightarrow [\exists\, p \in \mathbb{R}^l_+, \forall\, i \in N, \forall\, x \in \mathbb{R}^l_+,\ px \leq p\frac{\omega}{n} \Rightarrow z_i\ R_i\ x].$$

Theorem 11.1:

1. *The Fixed Numeraire z^* Egalitarian Equivalent rule can be rationalized by a social ordering function satisfying Weak Pareto, Pareto Indifference, Anonymity, Extended Interval Independence, and Independence of the Feasible Set.*
2. *The Pazner-Schmeidler Egalitarian Equivalent rule E_Ω and the Equal Income Walrasian rule W can be rationalized by a social ordering function satisfying Weak Pareto, Pareto Indifference, Anonymity, and Extended Interval Independence.*
3. *However, there is no social ordering function satisfying Weak Pareto and Independence of the Feasible Set that rationalizes either E_Ω or W.*

We view this theorem as the success evidence of the approach. It tells us not only that our alternative to Arrow's Independence of Irrelevant Alternatives can be combined with Weak Pareto, Pareto Indifference, and Anonymity, but it tells us also that in addition to satisfying these properties, the social ordering functions we obtain may rationalize famous equitable allocation rules. We also have to emphasize the negative content of point 3: A social ordering function rationalizing an equitable allocation rule is quite likely to depend on the resources of the economy.[11]

We begin the proof of the theorem here by defining social ordering functions that rationalize the three solutions and that satisfy the properties stated in Theorem 11.1. The proof is completed in the appendix.

A social ordering function $\mathsf{R}^{E_{z^*}}$ rationalizing E_{z^*} can be constructed by cardinalizing preferences as follows: $u_i^{z^*}(z_i) = \lambda \Leftrightarrow z_i\ I_i\ \lambda z^*$, for all $i \in N$. Then, we define $\mathsf{R}^{E_{z^*}}(e)$ by applying the maximin criterion to the vector $(u_i^{z^*}(z_i); i \in N)$.

[11] Eisenberg (1961) and Milleron (1970) have proved that the impossibility of rationalizing W by a social ordering function satisfying Independence of the Feasible Set disappears on the subdomain of homothetic preferences. The social ordering they propose is representable, interestingly, by the product of values taken by homogeneous (i.e., least concave) representations of the agents' preferences. We do not know of any result of this sort for the E_Ω rule.

The Social Ordering Function $\mathsf{R}^{E_{z^*}}$:

$$\forall\, e = (R_N, \Omega) \in \mathcal{E},\; z_N = (z_1, \ldots, z_n),\; z'_N = (z'_1, \ldots, z'_n) \in \mathbb{R}^{ln}_+,$$
$$z_N\, \mathsf{R}^{E_{z^*}}(e)\, z'_N \Leftrightarrow \min_{i \in N}\{u_i^{z^*}(z_i)\} \leq \min_{i \in N}\{u_i^{z^*}(z'_i)\}.$$

A social ordering function R^{E_Ω} rationalizing E_Ω can be constructed by cardinalizing preferences R_i as follows: $u_i^{\Omega}(z_i) = \lambda \Leftrightarrow z_i\; I_i\; \lambda\Omega$, for all $i \in N$. Then, we define $\mathsf{R}^{E_\Omega}(e)$ by applying the maximin criterion to the vector $(u_i^{\Omega}(z_i); i \in N)$.

The Social Ordering Function R^{E_Ω}:

$$\forall\, e = (R_N, \Omega) \in \mathcal{E},\; z_N = (z_1, \ldots, z_n),\; z'_N = (z'_1, \ldots, z'_n) \in \mathbb{R}^{ln}_+,$$
$$z_N\, \mathsf{R}^{E_\Omega}(e)\, z'_N \Leftrightarrow \min_{i \in N}\{u_i^{\Omega}(z_i)\} \leq \min_{i \in N}\{u_i^{\Omega}(z'_i)\}.$$

A simple social ordering function that rationalizes W can be constructed as follows. For $z_i \in \mathbb{R}^l_+$ and $R_i \in \mathcal{R}$, let $U(z_i, R_i)$ denote the upper contour set of z_i for R_i. For $A \subset \mathbb{R}^l_+$, let $\mathrm{co}(A)$ denote the convex hull of A. Take any economy e in $\mathcal{E}$. First, we give to each allocation a value depending on the intersection of the Ω-ray and the lower frontier of the convex hull of all agents, upper contour sets at their bundles. An allocation is preferred when its value is greater.

The Social Ordering Function R^{W}:

$$\forall\, e = (R_N, \Omega) \in \mathcal{E},\; z_N = (z_1, \ldots, z_n),\; z'_N = (z'_1, \ldots, z'_n) \in \mathbb{R}^{ln}_+,$$
$$z_N\, \mathsf{R}^{W}(e)\, z'_N \Leftrightarrow$$
$$\min\{\lambda | \lambda\omega \in \mathrm{co}(\cup_{i \in N} U(z_i, R_i))\} \leq \min\{\lambda | \lambda\omega \in \mathrm{co}(\cup_{i \in N} U(z'_i, R_i))\}.$$

Theorem 11.1 does not allow us, however, to identify the class of allocation rules that can be rationalized by social ordering functions satisfying the three basic properties as well as Extended Interval Independence or Independence of the Feasible Set. We still do not have general results regarding this question.

We find interesting, however, to complete this section by defining two ways of constructing social ordering functions from allocation rules. First, we introduce the appropriate concept. We propose to call a *social ordering functional* a function associating to each allocation rule in some admissible domain a social ordering function that rationalizes it. The two social ordering functionals are both based on a real-valued function computing what could be called the value of an allocation, depending on the parameters of the economy. Then, the social ordering function is derived from the principle that an allocation is socially preferred if its value is greater.

Let us consider a first way of computing the value of an allocation. Let S denote the allocation rule that one tries to rationalize. The value of an allocation z_N for an economy $e = (R_N, \Omega)$ is given by the highest share λ satisfying the property that an S-optimal allocation z'_N exists in the economy $(R_N, \lambda\Omega)$ to which all agents weakly prefer z_N. Formally, $v\colon \mathbb{R}_+^{ln} \times \mathcal{E}^S \to \mathbb{R}_+$ is defined by: for all $e = (R_N, \Omega) \in \mathcal{E}^S$,

$$v(z_N, R_N, \Omega) = \sup\{\lambda | \exists z'_N \in S(R_N, \lambda\Omega) \text{ s.t. } z_i \; R_i \; z'_i, \forall\, i \in N\}.$$

We construct R^S as follows: $z_N \,\mathsf{R}^S(e)\, z'_N \Leftrightarrow v(z_N, e) \geq v(z'_N, e)$.

The second value function is similar to the first one, except that the supremum is no longer defined with respect to economies $(R_N, \lambda\Omega)$ but to economies $(R'_N, \lambda\Omega)$, where the choice of R'_N is simply restricted by the condition that agent i's indifference curve through z_i be the same at R_i as at R'_i (we do not define this value function formally, to avoid long mathematical notations).

These value functions are clearly inspired by the Debreu (1951) coefficient of resource utilization. Actually, Debreu's coefficient corresponds to the function we obtain by applying either social ordering functional to the Pareto rule. However, the careful reader will have noted that by applying our first functional to E_{z^*} or to E_Ω, we come to the corresponding social ordering functions $\mathsf{R}^{E_{z^*}}$ or R^{E_Ω}, whereas the second functional, applied to W, gives us R^W. We think that social ordering functionals like these ones could prove extremely useful in the general research for appealing social ordering functions. Finally, let us note that any social ordering function obtained by applying the second functional satisfies Extended Interval Independence.

Before closing this subsection, we note that the results presented here in the context of division economies can be adapted to private or public good production economies. For instance, in the provision of one public good model, the solution discussed by Moulin (1987) can be rationalized in a similar way as E_{z^*}, whereas the Lindhal rule can be rationalized in a similar way as W. In conclusion, even in the absence of a clear general result, it does not seem too difficult to get allocation rules rationalizable by social ordering functions satisfying Weak Pareto, Pareto Independence, Anonymity, and Extended Interval Independence.

11.4 Concluding Comments

This chapter is exploratory and provides more questions than solutions. However, we hope that the kind of social orderings proposed here can contribute to bridging the gap between the social welfare approach and the fair allocation approach. We hope, above all, that such social ordering function will come closer to policy applications than traditional social welfare functions or first best allocation rules. The application of such orderings to second best problems in public economics might yield new policy recommendations that would rely only on ordinal noncomparable information.

Many questions remain open. First, the list of axioms that can bear on social orderings is far from closed. Second, the list of social ordering functionals must also be enlarged. Third, the characterization of allocation rules that can be rationalized by social ordering functions satisfying appealing properties is still an open problem. The axioms defined in this chapter and Theorem 11.1, however, already show that the analysis of social ordering functions is not limited at uncovering impossibility results.

References

Arneson, R. J. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56, 77–93.

Arrow, K. J. 1959. Rational choice functions and orderings. *Economica* 26, 121–127.

Arrow, K. J. 1963. *Social Choice and Individual Values*, 2nd ed. Wiley, New York.

Blackorby, C., Primont, D., and Russell, R. R. 1978. *Duality, Separability, and Functional Structure*. North-Holland, New York.

Bossert, W., Fleurbaey, M., and Van de gaer, D. 1999. Responsibility, talent, and compensation: A second-best analysis. *Review of Economic Design* 4, 35–55.

Cohen, G. A. 1989. On the currency of egalitarian justice. *Ethics* 99, 906–944.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19, 273–292.

Donaldson, D., and Weymark, J. A. 1988. Social choice in economic environments. *Journal of Economic Theory* 46, 291–308.

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.

Eisenberg, E. 1961. Aggregation of utility functions. *Management Science* 7, 337–350.

Fleurbaey, M. 1995. Equal opportunity or equal social outcome? *Economics and Philosophy* 11, 25–55.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge.

Herrero, C. 1996. Capabilities and utilities. *Economic Design* 2, 69–88.

Herrero, C., Iturbe, I., and Nieto, J. 1998. Ranking opportunity profiles on the basis of the common opportunities. *Mathematical Social Sciences* 35, 273–289.

Kranich, L. 1996. Equitable opportunities: An axiomatic approach. *Journal of Economic Theory* 71, 131–147.

Le Breton, M. 1997. Arrovian social choice on economic domains. In *Social Choice Re-examined*, Vol. 1, ed. K. J. Arrow, A. K. Sen, and K. Suzumura. Macmillan, London, pp. 72–96.

Maniquet, F. 1994. On equity and implementation in economic environments, Ph.D. diss. University of Namur.

Milleron, J. C. 1970. Distribution of income, social welfare functions and the criterion of consumer surplus. *European Economic Review* 2, 45–77.

Mongin, P., and d'Aspremont, C. 1998. Utility theory and ethics. In *Handbook of Utility Theory*, Vol. 1, ed. S. Barbera, P. J. Hammond, and C. Seidl. Kluwer, Boston, pp. 371–481.

Moulin, H. 1987. Egalitarian-equivalent cost sharing of a public good. *Econometrica* 55, 149–170.

Moulin, H. 1990. Fair division under joint ownership: Recent results and open problems. *Social Choice and Welfare* 7, 149–170.

Moulin, H., and Thomson, W. 1997. Axiomatic analysis of resource allocation problems. In *Social Choice Re-examined*, Vol. 1, ed. K. J. Arrow, A. K. Sen, and K. Suzumura. Macmillan, London, pp. 101–120.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA.

Rawls, J. 1982. Social unity and primary goods. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge University Press, Cambridge, pp. 159–185.

Thomson, W. 2001. On the axiomatic method and its recent applications to game theory and resource allocation. *Social Choice and Welfare* 18, 327–387.

Thomson, W. 2008. Fair allocation rules. In *Handbook of Social Choice and Welfare*, Vol. 2, ed. K. J. Arrow, A. K. Sen, and K. Suzumura, North-Holland, Amsterdam, forthcoming.

Van Parijs, P. 1995. *Real Freedom for All*. Oxford University Press, Oxford.

APPENDIX

Proof of Theorem 1: The straightforward proof that R^{E_Ω}, $\mathsf{R}^{E_{z^*}}$, and R^W all satisfy Weak Pareto, Pareto Indifference, and Anonymity is left to the reader.

Let us consider Extended Interval Independence. Let i and z_i be fixed. Let $u_i^{z^*}, u_i'^{z^*} \colon \mathbb{R}_+^L \to \mathbb{R}_+$ representing R_i and $R_i' \in \mathcal{R}$, respectively, be defined as in the construction of $\mathsf{R}^{E_{z^*}}$. If for all $z \in \mathbb{R}_+^l$, $z_i\ R_i\ z \Rightarrow z_i\ R_i'\ z$, then $u_i^{z^*}(z_i) \leq u_i'^{z^*}(z_i)$, whereas if for all $z \in \mathbb{R}_+^l$, $z\ R_i\ z_i \Rightarrow z\ R_i'\ z_i$, then $u_i^{z^*}(z_i) \geq u_i'^{z^*}(z_i)$. The claim that $\mathsf{R}^{E_{z^*}}$ satisfies Extended Interval Independence follows directly from this fact.

By replacing $u_i^{z^*}$ by u_i^{Ω}, we can prove that R^{E_Ω} satisfies Extended Interval Independence.

Let $R_N, R_N' \in \mathcal{R}^n$ be given with the property that for all $i \in N$, for all $z \in \mathbb{R}_+^l$, $z_i\ R_i\ z \Rightarrow z_i\ R_i'\ z$. Then we have $\mathrm{co}(\cup_{i \in N} U(z_i, R_i)) \supseteq \mathrm{co}(\cup_{i \in N} U(z_i, R_i'))$, and the proof that R^W satisfies Extended Interval Independence is straightforward.

Let us now consider Independence of the Feasible Set. First, it is clear that the $u_i^{z^*}$ functions do not depend on Ω at all and, consequently, $\mathsf{R}^{E_{z^*}}$ satisfies Independence of the Feasible Set.

Finally, take economies with $n = l = 2$, and preferences represented by $U_1(x, y) = 10\sqrt{x} + y$, $U_2(x, y) = x + 10\sqrt{y}$. Consider W first. Quasi-linearity and strict convexity of the preferences implies that the Walrasian equilibrium allocation is always unique. Let γ denote the ratio of 1's income over 2's income. Then for e with $\Omega = (25, 50)$, $W(e) = z_N^1 = ((19.16, 17.39), (5.84, 32.61))$ with $U_1 = 61.17$, $U_2 = 62.94$, and if $\gamma = 1.095$, one has the allocation $z_N^2 = ((19.85, 18.52), (5.15, 31.48))$ with $U_1 = 63.08$, $U_2 = 61.26$. And we have symmetric values for e' with $\Omega' = (50, 25)$ (allocations z_N^3 and z_N^4). We see that $z_N^1 \mathsf{P}(e) z_N^2$ and $z_N^3 \mathsf{P}(e') z_N^4$ by rationalization of W and Independence of the Feasible Set. Let e'' be the economy with $\Omega'' = (50, 50)$. By Independence of the Feasible Set, $\mathsf{R}(e'')$ coincides with $\mathsf{R}(e)$ in e, and with $\mathsf{R}(e')$ in e'. Besides, $z_N^2 \mathsf{P}(e'') z_N^3$ and $z_N^4 \mathsf{P}(e'') z_N^1$ by Weak Pareto. Therefore, $z_N^1 \mathsf{P}(e'') z_N^2 \mathsf{P}(e'') z_N^3 \mathsf{P}(e'') z_N^4 \mathsf{P}(e'') z_N^1$.

Now consider E_Ω. We have $E_\Omega(e) = z_N^1 = ((19.11, 17.30), (5.89, 32.70))$ with $U_1 = 61.01$, $U_2 = 63.08$, and $z_N^2 = ((19.89, 18.57), (5.11, 31.43)) \in P(e)$ with $U_1 = 63.16$ and $U_2 = 61.18$. The rest of the example is constructed as in the preceding paragraph. Q.E.D.

12

Rationality and Want-Satisfaction

Brian Barry

12.1 Introduction

The proposition that I wish to defend in this chapter is that the satisfaction of wants is in general good. More precisely, I argue that this proposition is not vulnerable to two objections against it that are quite commonly made and widely regarded as decisive. Both objections turn on the way in which wants are formed. According to one, the satisfaction of wants is compromised by the phenomenon of adaptive preference formation: We tend to want what we can get and not what we cannot. According to the other, the proposition that want-satisfaction has value commits us to the counterintuitive claim that everybody must, on pain of irrationality, work to create in themselves (and in others for whom they are responsible) wants that are easily satisfied. I take the canonical statements of these two objections to have been made by Jon Elster and John Rawls, respectively, and it is their versions that I shall discuss.[1]

Between them, these two objections suggest that there are serious problems with the idea that want-satisfaction is valuable. If we let nature take its course, changes in our preferences will occur "behind our backs," as Elster puts it, in such a way as to bring them in line with what is available, and this (he supposes) casts doubt on the value of getting what we want. But if we intervene self-consciously in the formation of our preferences, Rawls is telling us that we must deliberately try to do even more effectively exactly

[1] The essay by Jon Elster is "Sour Grapes." Page references are to pp. 109–140 in his *Sour Grapes: Studies in the Subversion of Rationality* (Cambridge: Cambridge University Press, 1983). The essay was originally published in Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (London: Cambridge University Press, 1982), pp. 219–238, when it bore the subtitle "Utilitarianism and the Genesis of Wants." The essay by John Rawls is "Social Unity and Primary Goods," in Sen and Williams, eds., pp. 159–185.

what (according to Elster) tends to happen anyway. More precisely, if want-satisfaction is our aim, we can either try to make what we want available or change our wants so that we want what is available already. These are equally valid alternatives, as far as want-satisfaction is concerned, and the choice between them should be made by deciding which is likely to be easier to effect. If, however, we think that the two strategies are not interchangeable, we must reject the notion that want-satisfaction is good because we cannot otherwise account for our differentiating them.

As one would expect in the case of two such distinguished scholars, what Elster and Rawls have to say is not simply mistaken. Indeed, Elster's central example, the fable of the fox and the grapes, is an example of irrational preference formation, but not for the reason that Elster says it is; and Rawls's criticism of the version of the want-satisfaction thesis that he attacks is entirely successful. But what follows from this is simply that we must be very careful about the way in which we state the thesis to be defended. What we are then left with is the proposition that it is better for the actual wants that people have to be satisfied than for them not to be satisfied, and this in general includes (*contra* Elster) those that have arisen by adaptive preference-formation.

12.2 Rawls's Objection

According to Rawls, then, utilitarianism "starts by regarding persons in terms of their capacities for satisfaction. It then interprets the problem of justice as how to allocate the means of satisfaction so as to produce the greatest sum of well-being."[2] I do not think that "satisfaction" and "well-being" are necessarily equivalent, and I do not see why utilitarianism should be equated with the notion that the good consists in want-satisfaction. (I shall take this up later.) It is, however, reasonable to think that utilitarians (and others, indeed) will accept that the satisfaction of the wants people actually have is in general a good thing. The point made by Rawls, though, is that utilitarians must be "ready to consider any new convictions and aims, and even to abandon attachments and loyalties, when doing this promises a life of greater overall satisfaction."[3] I deny that this follows.

An analogy is the following. Suppose that someone is a supporter of a certain football team: He roots for it, he wants it to win, he is pleased when

[2] Rawls, "Social Unity and Primary Goods," p. 181.

[3] Ibid.

it does and downcast when it doesn't, and so on. We might then imagine a philosopher saying: "Ah, what you want is that the team you support should win. But, given that that is your aim, it is quite irrational of you to support the collection of losers that you do support. You should switch to supporting some team with good long-run prospects of winning." The response would obviously be that this was to misunderstand the situation. What is entailed in being a fan of some team is that you identify with the fortunes of that team, so of course you want it to do well. To say that you want the team you support to win is correct, but only against the background assumption that there is some given team that you support. It would simply be an error to suppose that the statement could be detached from the background conditions under which it makes sense and used to generate a prescription about how to select a team to support.

Typically, a person becomes a supporter of some team on the basis of history. Often people will retain through life a loyalty to the local team from the place in which they grew up. And even if they do change their allegiance later on, this is normally explainable by later history. The whole idea of *choosing* a team to support tends subtly to undermine the value of being a supporter, and calculatingly choosing a team in order to maximize the chance of supporting a winner surely destroys it. We might just as well say that a good general wants the army he is leading to win and then deduce that an extremely good general will always be prepared to change sides if the side he is on looks like it's being beaten.

Suppose that someone continually changes the team he supports – perhaps even in the middle of a game – so as to support the winning side. We might say that such a person is not merely a fair-weather fan, he is no fan at all. Similarly, someone whose preferences were so labile that they changed continually so as always to be satisfied might well be said to be not a real person at all. It is rational in a quite straightforward way to want to satisfy the preferences I have. But there is nothing in that to suggest that I would be equally rational to want to become some entirely different kind of person who would be more satisfied with the things that would be available to that person than I am now with the things available to me.

Does utilitarianism entail that it would be an improvement to kill somebody as long as he could be replaced by somebody else with a higher level of want-satisfaction? Utilitarians rightly resist this conclusion. But in that case they can surely deny consistently that it must be an improvement for somebody to turn into another person with quite different tastes and aspirations, even if this new person's wants are more fully satisfied. This point is illustrated by a film I once made the mistake of watching on a transatlantic flight.[4] (It is, I imagine, the kind of film that is never seen anywhere else.) In it, George Burns (the Devil) makes a pact, on the usual terms, with an unsuccessful songwriter to make him a success. But instead of simply making his songs hits, the Devil puts the songwriter into the body of a hugely successful rock star whose soul he has just repossessed. Since somebody else now has the songwriter's body, lives in his house with his wife, and carries on writing songs, there are obvious logical problems in saying that the contracting party is now the rock star rather than that things have stayed exactly the same as they were before, with a successful rock star and an unsuccessful writer. (This is fudged by combining the songwriter's memories with the rock star's performing abilities.) But the point is in any case that the man who made the contract (if we grant that he is the rock star) has not got what he bargained for. What he wanted was that he should stay the same, and that all that would change would be that he would be a success at what he did. He did not want to become somebody else, even though that person was a success in his own line of business. (The film has him envying the "other man," with his original body, wife, and job.)

It may be said that the want-satisfaction thesis is too weak to be worth maintaining in the form in which I have stated it. But I believe that it is quite possible to reject it and that this makes it important to defend it. What makes the thesis important is its connection with liberalism. Let me make it clear that it is not in my view essential to the defense of liberal institutions. For I wish to maintain that arguments from the fair treatment of those with conflicting conceptions of the good can underwrite liberal conclusions.[5] Those who reject the want-satisfaction thesis can still therefore be liberals, provided they accept a certain idea of what fairness demands. But what about those who think a society can legitimately pursue a certain conception of the good? As long as they accept the want-satisfaction thesis, they will still finish up with liberal policy prescriptions. But if they reject the want-satisfaction thesis, the way is open to characteristically illiberal "communitarian" conclusions. For what is an appeal to "community standards" but an excuse for preventing people from doing things they want to do and that do not harm others? (If they did harm others, that would itself be a reason for preventing people from doing them, and the invocation of community standards would be unnecessary.)

[4] The film is called *Oh God, You Devil* and should on all accounts be missed.

[5] I have put forward an argument along these lines in *Justice as Impartiality* (Oxford: Clarendon Press, 1995).

Communitarians like to take restrictions on pornography as their example, and this serves their purpose well since they can darken counsel by talking about bans on public displays of pornography (e.g. in newsagents), which can be supported by liberals, and helping themselves to the feminist criticism of pornography, which would – if it could be established – show some pornography to be harmful and thus make its prohibition acceptable on liberal grounds. (The liberal objection to this argument is simply that the connection claimed between pornography and harmful acts is too tenuous to warrant the infringement of liberty proposed.) For these reasons and others – for example, the almost inevitable concomitant of the exploitation of minors – it is possible to soft-pedal the "community standards" arguments here. But regardless of the private opinions of communitarian philosophers, they must, if they are honest, accept that the case in which the appeal to community standards has the greatest potential for curtailing liberty is that of consensual homosexual sex.

In the case of *Bowers v. Hardwick*, a majority of the U.S. Supreme Court held that, if there is a strong sentiment in a state to the effect that homosexual sex is depraved, the legislature is entitled to turn that sentiment into law by making it a criminal offense.[6] The appeal to the upholding of community standards could not be clearer. It is not to my purpose to argue that the result in this case was constitutionally erroneous. For, *pace* Ronald Dworkin, I see no reason for assuming that the U.S. constitution is a liberal document in the contemporary sense. The point I want to make is simply that the want-satisfaction thesis would rule out the idea that homosexual relations between consulting adults could be intrinsically bad as long as they satisfied the wants of the parties involved. (Provided these relations are consensual there is a strong presumption that this will be so.) And it appears to me that all appeals to community standards turn on exactly the idea that some want-satisfaction is intrinsically bad – indeed, so bad that it is legitimate to employ the penalties of the law against it. If this is so, communitarianism must, as I have suggested, depend on the rejection of the want-satisfaction thesis.

Let me repeat the thesis that I wish to defend. This is as follows: it is better for the wants that people actually have to be satisfied than for them not to be satisfied. That is to say, if we take the wants that people actually have as given, it is a better state of affairs if one person's wants are more fully

[6] *Bowers v. Hardwick*, 106 S. Ct. 2841 (1986). "An adequate basis for the Georgia ban on sodomy is found, White said [speaking for the Court] in 'the presumed belief of a majority of the electorate in Georgia that homosexual sodomy is immoral and unacceptable' in presumed 'majority sentiments about the morality of homosexuality.'" Stephen Macedo, *Liberal Virtues* (Oxford: Clarendon Press, 1990), pp. 193–194, citing *Bowers*, p. 2846.

satisfied, as long as nobody else's are satisfied less. (This is in effect the Pareto principle for want-satisfaction.) The fundamental assumption is that want-satisfaction is good for the person who experiences it, and that (holding the want-satisfaction of others fixed) it is a better state of affairs if one person is better off in this sense than if he is worse off. Some philosophers wish to resist the second move here, claiming that it makes no sense to speak of better and worse states of affairs. They urge us to stick to talking about what is better or worse for particular people. However, I find it hard to believe that even those who propose this can really live by it. We try to bring about good results in our own lives, and we support public policies that we believe will have good outcomes. We could not do these things, nor could we argue with one another about them, if we were reduced to enumerating the effects of an action or public policy on a set of individuals, taken separately. To get beyond that, we must make use of the idea that there are better and worse states of affairs.

Obviously, there are immense limitations to a criterion for a better state of affairs that cannot deal with any change that makes at least one person better off (in want-satisfaction terms) and at least one worse off. We can extend the applicability of the criterion if we add that amounts of want-satisfaction can be quantified and that the amounts of want-satisfaction enjoyed by different people can be compared. For that enables us to propose that a better state of affairs is one in which the amount of want-satisfaction is greater. Two points should, however, be noted. The first is that it is perfectly possible to hold the want-satisfaction thesis, as I originally stated it, without accepting that aggregate want-satisfaction is an appropriate criterion for evaluating states of affairs. The second point is that the extended criterion will still be capable of yielding only a partial ordering over states of affairs. For it will give us an answer only for cases in which the preferences are the same. That is to say, given a certain set of preferences, we can now compare alternative states of affairs according to the degree to which wants are satisfied in aggregate. But we cannot say that one state of affairs is better or worse than another if preferences are different between them because my formulation of the want-satisfaction thesis does not provide us with the resources to enable us to compare situations in which preferences are different.

Of course, once we grant that degrees of want-satisfaction can be quantified and aggregated, it must be possible to make comparisons between situations in which preferences are different. We can, in principle, say that the total amount of want-satisfaction is greater in situation *y*, given the preferences people have in it, than in situation *x*, given the different preferences people have there. The question is what evaluative significance this comparison has. If we say that it shows *y* to be better than *x*, we must be assuming precisely the idea criticized by Rawls: want-satisfaction is to be maximized even if this simply means bringing preferences into line with what is available. For it is perfectly possible that the only difference between *x* and *y* is that the preferences have been engineered in the second case to bring about greater satisfaction with what is available, while what is available is exactly the same in both situations.

The want-satisfaction thesis in its acceptable form offers only incomplete rankings of states of affairs, while extending it so that it can rank all states of affairs makes it unacceptable. But it does not have to stand on its own and will typically play a subsidiary role within some richer ethical theory. Thus, the want-satisfaction thesis could very plausibly figure as a theorem in hedonistic or eudaimonistic utilitarianism, or a more pluralistic but still broadly utilitarian conception of the good such as that embodied in James Griffin's idea of "well-being."[7] For we might well think that, given that somebody does have a definite want for something, his life will in general go better – he has more pleasure, happiness, or well-being – if he gets it. This does not entail that aggregate want-satisfaction is to be maximized unless it is postulated that aggregate pleasure, happiness, or well-being are maximized when aggregate want-satisfaction is maximized. But this is not at all plausible for happiness and well-being, while equating want-satisfaction with pleasure (which has been done) simply drains the concept of pleasure of any independent connotations. It is certainly not what we would ordinarily mean by the term. Nor is it what the classical utilitarians meant by it.

Thus, although what Rawls is attacking is a doctrine that really is held by some economists and philosophers, he is quite wrong in ascribing it to the leading members of the utilitarian school. The equation of the utilitarian tradition with one that makes want-satisfaction the criterion goes all the way back to *A Theory of Justice*, where Rawls says that the kind of utilitarianism he will discuss is "the strict classical doctrine which receives perhaps its clearest and most accessible formulation in Sidgwick." He describes it as holding that "the principle for society is to advance as far as possible the welfare of the group, to realize to the greatest extent the comprehensive system of desire arrived at from the desires of its members."[8] Although Rawls gives a general citation to *The Methods of Ethics*, he does not refer to any

[7] James Griffin, *Well-Being: Its Meaning, Measurement and Moral Importance* (Oxford: Clarendon Press, 1986).

[8] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), pp. 23–24.

particular passage in support, and I am confident that he would be unable to do so.

For Sidgwick, the utilitarian claim is that "General Good is general happiness," and happiness is to be construed for this purpose as "desirable feelings."[9] And his extensive analysis of hedonism as a criterion for an individual's good makes no sense unless it is assumed that desires can be evaluated according to their contribution to happiness or pleasure (terms that Sidgwick treats as interchangeable). Thus, he says, in pursuing pleasure, we must allow for "the fact that we can change ourselves. For it may be that our past experience has been greatly affected by our being not properly attuned to certain pleasures, as (e.g.) those of art, or study, or muscular exercise, or society, or beneficent action; or not duly hardened against certain sources of pain, such as toil, or anxiety, or abstinence from luxuries: and there may be within our power some process of training or hardening ourselves which may profoundly modify our susceptibilities."[10]

Bentham, too, took the utilitarian criterion to be defined in terms of pleasures and pains, while John Stuart Mill quite explicitly denounced want-satisfaction as a description of the utilitarian end. Thus, in chapter 2 of *Utilitarianism*, he says that "the being whose capacities of enjoyment are low, has the greatest chance of having them fully satisfied." But only somebody who "confounds the two very different ideas, of happiness, and content" will suppose that this being has more happiness than one with more complex (and hence less easily satisfied) desires.[11]

No utilitarian who took pleasure, happiness, or well-being as the criterion of goodness would maintain that it is a good thing to raise children so that they want only things that are extremely easy to achieve. We may well believe that it is a good thing in bringing up children to imbue them with high standards and encourage them to pursue goals that are difficult (though not impossible) to achieve. We should thus be trying to instill preferences that will make it harder for them to reach a high degree of want-satisfaction than if we had encouraged them to be content with achievements easily within their capabilities. Yet, there is surely nothing in the least inconsistent in our hoping that they will be successful in achieving what we have encouraged them to want. Indeed, it is hard to see how we could attach any importance to their having wants of certain kinds if we did not believe that it mattered for them to be satisfied.

[9] Henry Sidgwick, *The Methods of Ethics,* 7th ed. (London: Macmillan, 1907), p. 398.

[10] Ibid., p. 149.

[11] John Stuart Mill, *Utilitarianism,* in *Utilitarianism, Liberty and Considerations on Representative Government,* ed. H. B. Acton (London: Dent [Everyman Library], 1972), p. 9.

Thus, a utilitarian of the kind I am now considering can quite well hold that some people's wants are so impoverished that even satisfying them fully will give them relatively little pleasure, happiness, or well-being. (Let us just call all of these "good," leaving it to be understood that the term is to be understood in the current context to include only these three notions or others akin to them.) The point is that, even granting that satisfying their wants cannot do them as much good as would satisfying wants they do not in fact have (if they had those wants), we can still say that it is for their good that the wants they have should be satisfied.

Thus, consider the example of the "tamed housewife" who has become such a cliché in the literature. Suppose that, after many years in which her horizons have been set by domestic chores, a woman now derives satisfaction only from keeping the house spotlessly clean. It is entirely reasonable for a utilitarian (or anybody else) to say that it is a pity that her wants are so restricted and that she would be better off with more extensive desires. But it is quite possible that (whatever might have been the case at some point in the past) chivvying her to pursue new goals now would merely make her miserable. Whether this is so, it is in any case surely right to say that her good will be advanced, given the wants she actually has at present, if she has the opportunity to satisfy them by keeping the house spotlessly clean.

It is not perhaps surprising that the want-satisfaction thesis is compatible with utilitarianism. But what is more interesting is that it is compatible with the rejection of utilitarianism. There are, to simplify matters as far as possible, two distinctive features of utilitarianism. One is a conception of the good of the kind just discussed; the other is the idea that states of affairs are to be evaluated by aggregating the amount of goodness that is enjoyed by different people and comparing the totals. I myself reject both, as (I am reasonably confident) do most people. But I do not regard that as in the least incompatible with the want-satisfaction thesis.

Let us first take up nonutilitarian conceptions of the good. The point that I wish to make here is that everything that has been said about the relation of utilitarian conceptions of the good to want-satisfaction can be extended to nonutilitarian conceptions. Thus, I have argued that a utilitarian can hold simultaneously that it would be better if people had different wants from those that they have and that it is good for the wants they actually have to be satisfied. Manifestly, a nonutilitarian conception of the good will also give us a basis on which to say that it is less good for people to have the wants they have than it would be for them to have some alternative set. Yet it will still be possible for us to believe that, given the actual wants people have, it is better for those wants to be satisfied than for them not to be satisfied. Thus you might posit some spiritual good as having the highest value, and thus hold that the most good comes about from the satisfaction of the desires of those whose goal is to achieve this spiritual end. But it would be possible to believe at the same time that satisfying the wants of those with other goals still has some value, though less value (perhaps a great deal less) than satisfying the wants of those with the most exalted aims.

Suppose someone shares with utilitarianism the premise that total good should be maximized but has a nonutilitarian conception of the good. There is no reason for expecting that, even if this person adheres to the want-satisfaction thesis, the implication drawn will be that aggregate want-satisfaction should be maximized. But here again there is no conflict between this position and that of someone with a utilitarian conception of the good. A utilitarian can, I have argued, say that wants have changed in such a way as to increase the amount of want-satisfaction but decrease the amount of good. A nonutilitarian can even more obviously say the same thing. The only difference between the two is that the conception of the good in the second case will not belong to the utilitarian family of pleasure, happiness, and well-being. Thus, the utilitarian and the nonutilitarian can agree both in rejecting the proposition that aggregate want-satisfaction should be maximized and in affirming the want-satisfaction thesis in the form in which I have stated it.

The second point is most easily illustrated by considering a theory of justice such as the one that I myself hold.[12] According to this, what justice requires is that rights, opportunities, and resources should be fairly distributed. For this purpose, the amount of good (on whatever conception of the good that one holds) that results from the distribution is not morally relevant. If more aggregate good would be created by shifting resources away from a fair distribution that does not provide a valid reason for doing so. Since the want-satisfaction thesis in its basic form is Paretian rather than aggregative, it is compatible with any distribution of resources. All it says is that, given the distribution of the resources and the actual preferences, it is an inferior state of affairs if there is less want-satisfaction than there might be.

12.3 Elster's Objection

In the rest of this contribution to the symposium, I shall show how the ideas put forward so far bear on the essay by Jon Elster on "Sour Grapes." Here, then, is Elster's statement of the problem that he sees for the want-satisfaction thesis.

[12] This thumbnail sketch is greatly expanded in *Justice as Impartiality.*

> My goal in this chapter will ultimately be to throw light on a problem arising in the foundations of utilitarian theory. It is this: why should individual want satisfaction be the criterion of justice and social choice when individual wants themselves may be shaped by a process that preempts the choice? And in particular, why should the choice between feasible options only take account of individual preferences if people tend to adjust their aspirations to their possibilities? For the utilitarian, there would be no welfare loss if the fox were excluded from the consumption of the grapes, since he thought them sour anyway. But of course the cause of his holding them be sour was his conviction that he would be excluded from consuming them, and then it is difficult to justify the allocation by invoking his preferences.[13]

In fact, the fable of the fox and the grapes does not suit Elster's purposes at all well. The fox, it may be recalled, tried to reach a bunch of grapes that was high up by jumping for them. Having failed to get them, he turned away, and as he did so said "I see now that they were quite sour." The natural reading of this is surely that he thought only that that particular bunch of grapes, which had proved inaccessible, was sour. If a little further along the path he came across another bunch of grapes that was close to the ground, there is nothing to suggest that he would be anything except delighted to eat them. Thus, if (as Elster seems to be imagining) we had some grapes and wanted to distribute them in an equitable way, the fox of the fable would be a perfectly good candidate for the receipt of a fair share. For he wants grapes in general; the only ones he does not want are the ones he is unable to get.

What Elster is gesturing at here, but fails to capture with the fable of the fox and the grapes, is the worry (natural enough in anyone with egalitarian inclinations) that the general "sour grapes" phenomenon will have perverse distributive implications. The underlying process at work in the fable is one of cognitive dissonance reduction. According to cognitive dissonance theory, people tend to adjust their beliefs so as to make themselves feel good.[14] Thus, if someone has just bought a new car, the theory suggests, he will tend to look around for evidence confirming the wisdom of his choice and steer clear of evidence to the contrary. So somebody who does not expect to get something may downgrade its value so as to reduce his disappointment at not getting it: This is the general form of the "sour grapes" phenomenon. The distribution problem is then that, if someone's low expectations have led him not to want much, any theory of just distribution that is sensitive to wants will have the implication that people with limited wants should get less than those with more extensive wants, even if there is no other

[13] Elster, "Sour Grapes," p. 109.

[14] See Leon Festinger, *A Theory of Cognitive Dissonance* (Evanston, IL: Row, Peterson, 1957).

difference between them. Thus, the poor and the oppressed are liable to be discriminated against by any such conception of just distribution.

The idea that the poor and oppressed have limited aspirations is scarcely new, and the "tamed housewife" is simply the currently fashionable exemplar. In fact, I am inclined to doubt that, at any rate in regard to material goods, cognitive dissonance reduction plays a very important role in contemporary societies. I do not mean here only western societies. Indeed, one of the earliest theories of modernization took as the beginnings of modernity the acquisition of a notion (even if somewhat distorted) of a middle-class American lifestyle from watching American films.[15] If the poor in more affluent countries such as Britain could not imagine how life might be better with more money, we would not be able to account for their spending a higher proportion of their incomes on the national lottery than those with more money. Presumably, the explanation is precisely that they have a *more* acute perception of the difference it would make to their lives to win a big prize.

Ambitions are a more complicated matter. If it requires an investment of time, effort, and money to realize an ambition, it is entirely rational to estimate the chances of success before pursuing it. "Don't put your daughter on the stage, Mrs Worthington," as Noel Coward memorably advised the mother of a young woman of limited dramatic talent. Elster, it is worth recalling, stated his alleged problem as arising "in particular" from "people tend[ing] to adjust their aspirations to their possibilities." But it must be emphasized that this is a normal and perfectly reasonable occurrence. There is no plausible conception of justice on which it sheds the least doubt. It should also be emphasized that the phenomenon has nothing to do with the one that is Elster's official concern, that of "sour grapes." Somebody who is tone-deaf will quite rationally decide that he had better not invest in preparing for a musical career. But this does not mean that he must denigrate the value of such a career to those who are suited to it. He may do so, of course, but even if he does not (even if he profoundly wishes he were able to be a musician), he still has good reason for looking elsewhere for a line of work.

If there is discrimination against women in certain occupations, any remotely acceptable conception of justice will enable us to condemn that. Assuming the discrimination is real, women may well be discouraged from training for jobs they are unlikely to get, and that is in itself no more objectionable than the tendency of the tone-deaf to eschew careers as musicians. Again, since women will not then be qualified for certain jobs it will not be unjust if they do not get them, on the assumption that jobs should go

[15] See Daniel Lerner, *The Passing of Traditional Society* (New York: Free Press, 1958).

to those qualified to do them. But we can still, of course, say that the lack of women applicants with the appropriate qualifications is itself the product of discrimination and thus arises from an injustice.

Suppose, alternatively, that the absence of an ambition to pursue certain occupations comes about as a consequence of a widespread social norm to the effect that some jobs are "unsuitable" for women, but that there is no discrimination against those who want them and have competitive qualifications. There are many conceptions of the good (including but by no means confined to the utilitarian ones previously discussed) on the strength of which such a state of affairs can be deplored. But, as with the case of discrimination, there is no injustice per se in women not holding these jobs if the result of the social norm is that qualified women candidates do not present themselves. Moreover, to the extent that the social norm is actually effective in causing women not to want jobs of certain kinds, the want-satisfaction principle says that it would be a bad thing for them to have them, since it is bad for people to be made to do things they do not want to do. But we can still, of course, say that it would be better if some women did want these jobs.

As this example shows, the relation between wants and outcomes is complex, and whenever we are dealing with a social institution it is mediated by some system that allocates rights, opportunities, and resources. But however complicated we make things, the validity of the want-satisfaction principle remains unchallengeable. For the present purpose, what are especially relevant are two corollaries of the basic principle. These both relate to cases in which people are given things they do not want. Consider first a case in which people are given things they do not want but have the right not to make use of them. In such a case, they are no worse off for getting them, but equally they are no better off. (I stipulate that they cannot exchange what they are given for things they do want.) If the same things could have gone to others who would have wanted them, the result would have been that additional wants would have been satisfied with no loss to anybody. The want-satisfaction principle therefore generates the conclusion that it is a less good state of affairs than is attainable to allocate nonexchangeable goods to those who do not want them if they could instead have been given to those who do.

The other case is one in which there is no right of refusal: people simply get something whether they wish to have it or not. Here, the situation is bad in two ways. As before, there is the indirect violation of the want-satisfaction principle in that the same goods could have gone instead to those who wanted them. But there is now a direct violation as well in that the actual recipient gets something positively unwanted. Thus, even if there were no alternative use for the good, it would still be better not to force it on somebody who does not want it. For whatever reason, somebody does not want caviar; if he really does not, then it is wasteful to present him with a pound of beluga and thoroughly obnoxious to force it down his throat. We can say as much as we like about the potential source of delight he is missing out on. That does not alter the implications of the want-satisfaction principle, which are surely completely sound in such a case.

This brings me back to the quotation from Elster. My suggestion is that any advantage Elster gains for his cause is derived entirely from the imprecision of his presentation. In talking about "excluding" the fox from the consumption of the grapes, he seems to imply that somebody has some grapes to allocate (never mind how or why) and is asking whether some should be given to the fox. But the fox of the fable, I have argued, thought only that unattainable grapes were sour. If he knew that these were available, he would want them, and leaving him out would deprive him of a source of want-satisfaction. Of course, if the fox were told that he could not have any grapes, he would not then (we are to suppose) want them. But if we add that twist to the case, what is wrong is giving the fox the false information that there is no possibility of his getting any of the grapes, which (in conjunction with his peculiar way of forming preferences) results in his not wanting them.

The want-satisfaction thesis in its valid form does not, let us recall, allow comparisons between situations in which wants are different. It cannot therefore pronounce on the choice between the situation in which some good is known to be available (and hence wanted) and the one in which it is known not to be available (and hence not wanted). It is perhaps worth observing, however, that the invalid version – which seems to be the "preference utilitarianism" that is the target of Elster's critique – is quite unambiguous. According to it, there can be no doubt that the person concerned has more want-satisfaction in the first case than in the second because he has an additional want and it is satisfied. Moreover, the richer and more plausible versions of utilitarianism that I discussed before will all generate the conclusion that (taking the agent in isolation, as Elster does) there is more good in the first case, in which grapes are wanted and supplied, than in the second case, in which they are not wanted and not supplied.

It is, of course, a familiar objection to all good-maximization theories that they will tend to allocate more resources to those who are more efficient at turning resources into good (however good is defined) than to those who are less efficient. That is indeed a strong objection, and is one reason (though not by any means the only one) for proposing that justice should be defined over allocations of resources and should not take account of the use different people make of them. But it should be observed that the objection is quite independent of the "sour grapes" phenomenon in the form currently under review. A theory of the good call tell us that it would be a better state of affairs if people had different wants from those they actually have. It can therefore certainly tell us that it would be better for somebody to have more wants than he does, especially if they can be satisfied.

I have assumed that the grapes that are (for mysterious reasons) to be distributed are generic grapes. There is, I suppose, an alternative reading of the case, according to which "the grapes" in question are not just any old grapes but are *the very same grapes* as the fox had previously pronounced sour in his frustration at not being able to reach them, and are somehow identifiable as such. Elster may then be reasoning that, having once rejected the grapes, the fox will still reject them even if they are now somehow available to be allocated among a set of potential beneficiaries that includes him. But the theory of cognitive dissonance says nothing about what will happen in such a case. Suppose, to continue the narrative of the fable, just after the fox had turned away from the grapes saying that he saw they were unripe, the bunch he had been jumping for providentially fell off the branch. He might turn back and say, "No, it was a mere trick of the light. Now I see them from close up I realise that they are ripe after all." Only a foolish consistency, that "hobgoblin of little minds," would prevent him from changing his mind in this way.

Let us for the sake of argument imagine that the case is as Elster requires for his purposes, so that the fox's decision against these particular grapes is irrevocable. We then have a (very awkward) representation of the kind of case that is widely thought to spell trouble for the want-satisfaction thesis. This is the case in which a belief that some entire range of goods is inaccessible creates a permanent and irrevocable belief that they are not worth having. I have cast doubt on the prevalence of any such phenomenon, but let us assume for the sake of argument that it does exist.

A better way of adapting the fable to such a case is to suppose the fox, having repeatedly failed to acquire grapes, forms the general belief that all grapes are sour: Even when presented with ripe grapes on a plate he will not now touch them. I still maintain that the want-satisfaction thesis gives the right answer in such a case. If we have some grapes to allocate, we would be wasting them on the fox: the "welfare loss" in this instance is quite clearly created by giving him grapes he does not want. What about distributive justice? The most important thing to he said about distributive justice here is that a theory of distributive justice is a theory about institutions that have distributive implications; it is not a theory about the distribution of particular goods such as grapes. This is above all the lesson taught us by Rawls in *A Theory of Justice*, and if he had done nothing else, he would deserve his fame for having worked through the implication of that idea. Aristotle's discussion of the possible criteria for allocating a flute began a bad tradition that continues to this day and is well represented by Elster's grapes. Like Aristotle's flute, they came from nowhere, and the authority of the person who is to do the allocating to undertake this role is nowhere explained. With this hopeless starting point, it is scarcely surprising if the problem can easily be made to seem insoluble. Aristotle knew no better: He had no conception of distributive justice in the contemporary sense. But there can be no excuse for continuing to follow in his footsteps.

If a theory of justice is (as I believe) one that defines justice in terms of the distribution of rights, opportunities, and resources, the sour grapes phenomenon, even in the sense now being considered, will not have any effect because the just distribution can be defined without alluding to wants. Recall, for example, that if a just distribution of job opportunities is a nondiscriminatory one that remains true even if in fact the expectation of discrimination chokes off qualified applications from members of the group suffering from discrimination.

It is true that the want-satisfaction principle implies that people whose wants have been limited (via the "sour grapes" mechanism) by expectations based on injustice may be unable to take up opportunities that later open up as a result of a shift to a more just society. But to the extent that that is so, all we can say is that is an inertial effect: Changing toward a more just society may not be able to put right all the effects of previous injustice. (To give an obvious example: When the National Health Service was introduced in Britain, it could do only a certain amount for those whose health had already been damaged by previous neglect.) It does not seem to me to be a defect of a theory of justice that it does not promise miracles. It is simply that it may take generations for more just institutions to work their way out fully. All the more reason for having them sooner rather than later!

Good-maximizing theories, I have already pointed out, have the disadvantage of giving most to those who are most efficient at turning resources into good. This may, in conjunction with an irreversible "sour grapes" phenomenon, result in those with limited wants getting less because their marginal production of good starts falling off at a fairly low level of resources. But we can still say what we said for the reversible "sour grapes" case: It would be better if they had more wants and hence were better producers of good. The only difference is one of time scale. In the case of irreversible "sour grapes," it is too late to do anything for those who are already damaged, but that makes it all the more important to change conditions for the next generation. We may still object to the idea that the present generation should get less because previous deprivation has made them inefficient producers of good. But that, to repeat, is a direct consequence of good-maximizing. There is nothing special about "sour grapes": it would be just the same however it came about that somebody was a poor converter of resources into good.

So much for the "sour grapes" phenomenon. I have given it as long as possible a run for its money, and my conclusion is that it does not have any deleterious implications for a sensible version of utilitarianism (either the valid form of want-based utilitarianism or those with some more substantive conception of the good). To the extent that there are problems with utilitarianism, they are problems it has anyhow. At the most, we might say that the last point brings home in a particular way a drawback inherent in all good-maximizing theories.

I want to conclude this discussion of Elster, however, by insisting that "sour grapes" is only one aspect of the general phenomenon of "adaptive preferences," and a quite peculiar and uncharacteristic one. The point about "sour grapes" is that it is irrational in the sense that it is (to say the least) a highly unreliable method of belief formation. If beliefs based on "sour grapes" are false, and if they prevent people from wanting things they would enjoy in the absence of that false belief, it is quite straightforward that people would be better off without them. The same may be said of any other false belief that prevents people from wanting things they would otherwise enjoy. If you believe (falsely, let us suppose) that all grapes have been sprayed with harmful chemicals, that will put you off them just as effectively as believing they are all sour. We can again say that you would be better off without this belief.

But then why stop at beliefs? Phobias prevent people from doing things that they would otherwise want to do, such as leaving the house or flying on an airplane. People would be better off without them too. The point is that there may be all kinds of pathologies about what people want. And no doubt it would be worth a good deal of effort in some of these cases to change these wants. But the fact remains that, given their wants, they are better off getting what they want and not getting what they do not want; and it remains true that giving them what they do not want is a waste of resources. So there is no point in giving a ticket to the opera to somebody who suffers from agoraphobia (even if they love opera), and it would be even less of a kindness to avoid wasting the ticket by forcing them to go.

Of all these pathologies, only "sour grapes" could be described (and then rather misleadingly) as falling under the description of "adaptive preference formation." Elster, however, uses "sour grapes" as if it could be treated as a paradigm of all adaptation of preferences to circumstances. But clearly it is not. I have already pointed out that in general adjusting our aspirations to our possibilities – taken by Elster as somehow problematic – is highly rational and nothing to do with sour grapes. Most adaptive preference formation, however, is neither rational nor irrational. According to Elster, preference change that takes place by "a purely causal process of adaptation, . . . 'behind the back' of the person concerned" is irrational, and only preference change that takes place as the result of a process of "character planning" is rational.[16] This is, I submit, absurd. Wants are not immaculately conceived: All of them have causal antecedents. Character planning has to arise from a second-order want to have different first-order wants and that itself must come from somewhere. (Adding further levels will not help.) Whether we do well or ill by character planning depends on the nature of our aim. But it can equally well be said that whether a want that arose from adaptation is good depends on its content and not its ancestry.

The general point is that there is absolutely nothing wrong with adaptive preferences as such. People in Mexico grow up liking enchiladas and refried beans; people in India grow up with a taste for curry and rice; people in England grow up to want fish fingers or hamburgers, and so on. We tend to like what we get. The human race would not have been very successful if it did not have that kind of adaptability. Contrary to what Elster suggests, then, there is nothing irrational about preference formation that takes place "behind people's backs" – nor is planned character change necessarily rational. You might, for example, deliberately develop a taste for margarine in the belief that it is better for you than butter when in fact the hydrogenization process makes it quite harmful – or so I have read. (I give this example in tribute to our conference hosts in Normandy.) This is, of course, once again an example of a want (in this case deliberately cultivated) based on a false belief.

12.4 Conclusion

Let me draw these remarks to a conclusion. I have not tried to show that the satisfaction of wants is valuable. I have no idea how one would set about trying to show it, except by pointing to the obvious fact that each of us thinks it is valuable for himself, and we all make efforts to satisfy our wants. What I have tried to do is disprove two claims that there is some fundamental problem about want-satisfaction. Both of the arguments I have looked at

[16] Elster, "Sour Grapes," p. 117.

involve preference shaping. Elster claims that only deliberate preference change is rational and that adaptive preference formation is irrational. I have suggested that this is simply not true and that plausible cases of what might loosely be called irrational preference formation invariably involve some other factor - often some defect in the relevant beliefs. There is nothing wrong with preferences that came about as a result of the standard psychological processes of reinforcement and extinction. This is in fact how almost all the preferences we actually have came about, so if Elster were correct the world would be awash in irrationality. Fortunately, he is wrong.

For Rawls, utilitarians have to engage in character planning to be rational – so far he is in agreement with Elster. But Rawls claims that this character planning has to take one particular form. Utilitarians must if they are to be rational change their preferences so as to have wants that can easily be satisfied. Rawls suggests that this is a perverse way of behaving. So if it is what utilitarianism demands as the price of rationality, that amounts to a disproof of utilitarianism. I have denied that utilitarianism has the implications that Rawls claims for it. To say that it is a good thing for the wants we have to be satisfied does not entail that to be rational we must have easily satisfied wants. Indeed, I have argued that the former proposition can be held by nonutilitarians as well as utilitarians and has much to he said for it in its own right.[17]

[17] I am grateful for comments on the original version of the paper to participants in the Caen symposium and also for comments on a revised version to John Broome and Jonathan Riley.

PART FOUR

SHARING THE GAINS FROM SOCIAL COOPERATION

13

Naturalizing Harsanyi and Rawls

Ken Binmore

> He who would understand *baboon* would do more towards metaphysics than John Locke.
>
> Charles Darwin

13.1 Introduction

This chapter reviews a number of ideas drawn in two books: *Game Theory and the Social Contract* and *Natural Justice* (Binmore, 1994, 1998, 2005). The first volume of *Game Theory and the Social Contract* was published in 1994 with the subtitle *Playing Fair*. The second volume was published in 1998 with the subtitle *Just Playing*. *Natural Justice* is a popular acccount of the same material. The current chapter focuses on three issues.

Harsanyi scholarship. Some of the criticisms leveled at Harsanyi's (1977) utilitarian theory are answered by offering an account of his ideas to which the criticisms do not apply. In doing so, this chapter distinguishes two separate lines of thought to which Harsanyi has contributed: a teleological and a nonteleological approach.

Comparison of Harsanyi and Rawls. Harsanyi's nonteleological approach to utilitarianism is compared with Rawls's (1971) defense of egalitarianism. It is argued that Rawls would also have been led to utilitarianism if he had not adopted the iconoclastic expedient of abandoning Bayesian decision theory in favor of the maximin criterion. But the Rawlsian conclusion survives if one abandons instead the assumption that citizens are committed to the

I am grateful to the Economic and Social Research Council and to the Leverhulme Foundation for funding this research through the Centre for Economic Learning and Social Evolution at University College London.

hypothetical deal reached in the original position. The distinction between Rawls and Harsanyi then becomes a question of whether an external enforcement agency is available to police the rules that citizens make for themselves under ideally fair circumstances.

Naturalistic foundations. The Kantian or metaphysical foundations of Harsanyi and Rawls are replaced by a Humean or naturalistic perspective. From this perspective, the device of the original position is seen as a stylized version of a fairness norm that evolved along with the human race. The empathetic preferences that are necessary as inputs when the device is employed are then envisaged as being shaped by the forces of social evolution. These forces will tend to equip everyone with the same empathetic preferences, which then provide a standard for making interpersonal comparisons of utility. Such an approach makes it possible to attack the problem of how to weigh one person's utility against another's. Unfortunately, space does not allow more than a brief outline of the approach pursued in my books, notably *Just Playing* (1998).

Debates between egalitarians and utilitarians typically reduce to slanging matches in which each side pokes fun at the other by inventing bizarre moral problems that they assert will be solved by the other in a counterintuitive manner. I find such debates uninstructive because each side is so sure that the other is wrong that neither feels any need to propose models that allow the methodological apparatus of the other to be brought to bear on the problems in an honest fashion. Nor do I see the relevance of what people claim to be moral intuitions about situations that have occurred too rarely in the past to have had any significant impact on the evolution of our culture. However, if taken seriously, my approach to the issues may dampen the enthusiasm for conducting the controversy at this level because I shall be arguing that the egalitarian and the utilitarian solutions to a moral problem can be *precisely the same* within my framework if both use the standard for making interpersonal comparisons that will evolve in the medium run. Under such circumstances, neither side can ridicule the other without ridiculing itself. Perhaps attention can then be turned to the question on which I think the differences between egalitarians and utilitarians really turn; namely, when can we trust an external enforcement agency to execute its mandate without becoming corrupt?

It will be evident that many loose ends will be left when addressing such large questions in such a small compass. Dissatisfied readers may find some of their concerns addressed in my books, but I am conscious of many defects in my analysis, of which I shall mention only two. First, problems of coalition formation are evaded by confining attention to societies with only two citizens, Adam and Eve. Second, informational problems are neglected altogether by the expedient of assuming that everything Adam and Eve are not specifically required to forget on passing behind the veil of ignorance is common knowledge.

13.2 Teleological Utilitarianism

In introducing Harsanyi's (1977) teleological approach to utilitarianism, three questions that all utilitarians need to answer will be considered:

- What constitutes utility?
- Why should individual utilities be added?
- Why should I maximize the sum of utilities rather than my own?

What constitutes utility? In answering the first question, Bentham and Mill felt comfortable in talking about happiness, but modern authors have learned to be more circumspect. One school of thought led by Scanlon (1975), Sen (1980, 1988), Dworkin (1981), Cohen (1993), and others, argue that welfarism is an inadequate foundation on which to build a moral theory. Since Sen (1988) defines welfarism to be the approach to social choice that pays no attention to anything but the utilities that citizens assign to the available social alternatives, this postwelfarist movement challenges the claim of neoclassical economics that a person's well-being can be adequately assessed in terms of the extent to which his preferences are satisfied. Some authors even deny that subjective criteria like individual preferences have any place at all in a moral theory.

As an alternative to utility as normally understood in economics, postwelfarists give lists of supposedly objective criteria that need somehow to be weighed against each other in determining how well off someone is. The most famous example is Rawls's (1971) index of primary goods. Attempts are then made to deny legitimacy to those who stick by orthodox utility theory. For example, Sen (1976) and Roemer (1996) deny that Harsanyi (1977) can properly be counted as a utilitarian because he interprets utility like Von Neumann and Morgenstern.

My own view is that the movement away from welfarism is retrograde. Apart from their incipient ipsedixism,[1] postwelfarists ignore the fact that

[1] Bentham (1987) defines an ipsedixist to be someone who proposes his own prejudices as moral imperatives for others to follow. Such prejudices are all too apparent in the lists that postwelfarists compile when seeking to characterize the good life. But, as Bentham (1987) observes, who are we to urge our preference for poetry on those who prefer pushpin?

modern utility theory was invented *because* attempts to encompass human aspirations in terms of a priori objective criteria have been universally unsuccessful.

I agree with postwelfarists that the Victorian ideas on utility held by Bentham and Mill are an inadequate basis on which to make moral judgments. If hedonistic dials could somehow be wired into Adam and Eve's pleasure centers, justice would not be achieved merely by ensuring that both dials give the same reading. Among many other objections to such a naive approach to justice, there is the point made by Scanlon (1975) that some people have "champagne tastes," which are costly to satisfy, and that such costs need to be taken into account along with the benefits that people enjoy when considering how social decisions are made.

However, such criticisms miss the target when directed at versions of welfarism based on modern theories of utility. In particular, it is incoherent to attack modern welfarism on the grounds that the utilities attributed to the citizens in a society do not adequately reflect their motivations. According to the modern view, if their utilities did not adequately reflect their motivations, they would be the wrong utilities. In particular, the charge that utility theory necessarily pays attention only to the benefits that citizens receive while neglecting the costs that gratifying these benefits may impose on other citizens seems very strange to a neoclassical economist. Modern utility theory was invented as a theoretical tool precisely to assure that nothing that matters when decisions are taken is left out of account.

One would have thought that a more appropriate criticism from an objectivist would be to complain that utility theory takes *too much* into account for it to be useful as a guide to practical decision making. But the appropriate response to the latter very reasonable criticism is not to invent a bad theory which pretends that we know how to characterize all the fears and aspirations that people nurse in their bosoms in terms of a simple set of universally valid, objective criteria. The appropriate response is to continue to work with a good theory, but to recognize that each application requires a new search for objective criteria that adequately capture the subjective components of the theory. Such objective criteria will usually depend very strongly on the *context* in which the practical problem arises.

Having argued that Harsanyi (1977) should not be denied the utilitarian label for interpreting utility in the sense of Von Neumann and Morgenstern, I now need to defend him for sticking by their definition of a utility function. To discuss the criticism that Weymark (1991) and others direct at Harsanyi's approach to interpersonal comparison, it is necessary to begin by identifying a potentially confusing ambiguity in terminology. Recall that

Von Neumann and Morgenstern's consistency postulates imply that a rational player makes risky decisions as though seeking to maximize the expected value of a utility function. More precisely, a function $\phi\colon\ \Omega \to \mathbb{R}$ exists such that his preference relation $\preceq$ over the set lott(Ω) of lotteries with prizes in Ω is described by $\mathcal{E}\phi\colon\ \text{lott}(\Omega) \to \mathbb{R}$.[2] The question at issue is whether ϕ or $\mathcal{E}\phi$ is said to be a Von Neumann and Morgenstern utility function. The choice matters because $\phi\colon\ \Omega \to \mathbb{R}$ is a *cardinal* utility function for the restriction of $\preceq$ to Ω, whereas $\mathcal{E}\phi\colon\ \text{lott}(\Omega) \to \mathbb{R}$ is only an *ordinal* utility function for the unrestricted preference relation $\preceq$ on lott(Ω). Because it can only make sense to compare utils on cardinal scales, someone who follows Harsanyi in comparing Von Neumann and Morgenstern utility functions must therefore necessarily have settled the nomenclature question in favor of ϕ.

A second difficulty with Harsanyi's (1977) use of the Von Neumann and Morgenstern theory of utility arises from the fact that it is commonly held that a function ϕ that measures individual welfare should admit the interpretation that $\phi(b) - \phi(a) > \phi(d) - \phi(c)$ if and only if the citizen's preference for b over a is more intense than his preference for d over c. Otherwise a util given to a person when his welfare is high might be worth more or less than when his welfare is low. Those who draw this conclusion merely from the fact that ϕ is a Von Neumann and Morgenstern utility function fall prey to one of the standard fallacies listed in Luce and Raiffa's (1957) textbook. To obtain the conclusion legitimately, one has to make an *assumption* about what intensities of preference are to be taken to mean. Harsanyi's use of Von Neumann and Morgenstern's risk-based definition of intensity provides just such a legitimizing assumption and hence provides one of many possible foundations for the concept of individual welfare.

Suppose that $a \prec b$ and $c \prec d$. Then Harsanyi follows Von Neumann and Morgenstern (1944, p. 18) in arguing that a citizen should be deemed to hold the first preference more intensely than the second if and only if he

[2] To say that a utility function $\Phi\colon\ \text{lott}(\Omega) \to \mathbb{R}$ describes a preference relation $\preceq$ on the set lott(Ω) means that $\mathbf{L} \preceq \mathbf{M}$ if and only if $\Phi(\mathbf{L}) \leq \Phi(\mathbf{M})$. A Von Neumann and Morgenstern utility function $\phi\colon\ \Omega \to \mathbb{R}$ is defined by the requirement that $\mathcal{E}\phi\colon\ \text{lott}(\Omega) \to \mathbb{R}$ is a utility function for $\preceq$ on lott(Ω). Note that the Von Neumann and Morgenstern utility function $\phi\colon\ \Omega \to \mathbb{R}$ does *not* describe $\preceq$ on lott(Ω): it describes the *restriction* of $\preceq$ to Ω. A function ϕ is a Von Neumann and Morgenstern utility function if and only if the same is true of $A\phi + B$, for all constants $A > 0$ and B. As with a temperature scale, one is therefore free to choose the zero and the unit on a Von Neumann and Morgenstern utility scale, but then everything else is determined. After zeros and units have been chosen on Adam and Eve's utility scales, it then makes sense to compare Adam and Eve's utils—just as it makes sense to compare the degrees on a Centigrade thermometer with the degrees on a Fahrenheit thermometer. In brief, it is meaningful to compare *cardinal* utility functions.

would always be willing to swap a lottery **L** in which the prizes a and d each occur with probability 1/2 for a lottery **M** in which the prizes b and c each occur with probability 1/2. To see why Harsanyi proposes this definition, imagine that the citizen is in possession of a lottery ticket **N** that yields the prizes b and d with equal probabilities. Would he now rather exchange b for a in the lottery or d for c? Presumably, he should prefer the latter swap if and only if he thinks that c is a greater improvement on d than b is on a. But, to say that he prefers the first of the two proposed exchanges to the second is to say that the citizen prefers **M** to **L**.

In terms of the citizen's Von Neumann and Morgenstern utility function ϕ, the fact that $\mathbf{L} \prec \mathbf{M}$ reduces to the proposition that

$$\tfrac{1}{2}\phi(a) + \tfrac{1}{2}\phi(d) < \tfrac{1}{2}\phi(b) + \tfrac{1}{2}\phi(c).$$

Thus, the citizen holds the preference $a \prec b$ more intensely than the preference $c \prec d$ if and only if $\phi(b) - \phi(a) > \phi(d) - \phi(c)$.

Of course, other definitions of intensity of preference will lead to other conceptions of individual welfare. One might then be led, as is Roemer in this volume, to replace ϕ as a welfare measure by some strictly increasing transformation of ϕ. But Harsanyi's definition would seem the obvious first avenue of exploration.

Why add utilities? This question includes the problem of comparing utils across individuals. Unless utilities are suitably weighted before being added, you might as well pretend, as Bentham observes, to add twenty apples to twenty pears. Postwelfarists have no answer to the problem of interpersonal comparison. The criteria for human well-being they compile are simply asserted to be universally applicable. Harsanyi has two answers. This section examines his teleological answer.

A teleological moral theory postulates an a priori conception of the "common good". Harsanyi (1977) follows a well-trodden path by writing down a list of axioms that are said to characterize the common good. The nature of the common good is then deduced from these axioms, rather than simply being asserted to be self-evident. The axioms used by utilitarians like Harsanyi take the personal preferences of individual citizens as given and describe how these are to be related to the preferences of society as a whole. Society is therefore treated as a single person written large. Or, to follow Adam Smith (1759/1975), the interests of society are identified with those of an *ideal observer*, whose preferences are some kind of average or aggregate of the preferences of all the individuals in the community.[3] This ideal observer's utility function is then said to be a welfare function for the society as a whole.

The mathematics with which Harsanyi (1955, 1977) defends his teleological approach have been criticized by various authors. However, the trivial mathematics offered here do not seem vulnerable to similar criticism. Nor are the axiomatic approaches of Broome (1991), Hammond (1988, 1992), Maskin (1978), and numerous other authors.

Consider a society with two citizens, Adam and Eve. If Adam and Eve honor the Von Neumann and Morgenstern rationality conditions, their preferences over lotteries can be summarized by Von Neumann and Morgenstern utility functions defined on a finite set S of possible social states. To keep things simple, let us assume that Adam and Eve agree that there is a worst social state $\mathcal{L}$ and a best social state $\mathcal{W}$. We can then select the zeros and the units on Adam and Eve's utility scales so that $u_A(\mathcal{L}) = u_E(\mathcal{L}) = 0$ and $u_A(\mathcal{W}) = u_E(\mathcal{W}) = 1$. The Von Neumann and Morgenstern utility functions u_A and u_E will serve as inputs to our algorithm for determining the preferences of the ideal observer.

The next step is very much less innocent. Not only are Adam and Eve assumed to honor the Von Neumann and Morgenstern rationality conditions, but the same is taken to be true of the ideal observer, who is assumed to be no different in kind from Adam and Eve. In particular, he may be assigned a Von Neumann and Morgenstern utility function u_i that describes his preferences over lotteries in which the prizes are social states. As with Adam and Eve, the zero and the unit on the ideal observer's utility scale will be chosen so that $u_i(\mathcal{L}) = 0$ and $u_i(\mathcal{W}) = 1$.

The third step requires that, in assessing the consequences of choosing an action that leads to one lottery over social states rather than another, the ideal observer is assumed to take into account *only* Adam and Eve's personal preferences over the lotteries. One can translate this requirement into formal terms by requiring that, for each lottery **L**,

$$\mathcal{E}u_i(\mathbf{L}) = v_i(\mathcal{E}u_A(\mathbf{L}), \mathcal{E}u_E(\mathbf{L}))\,, \tag{13.1}$$

where the values $v_i(x_A, x_E)$ of the function v_i depend only the pair (x_A, x_E) of utilities that Adam and Eve receive as a consequence of the ideal observer's decision. In brief, the ideal observer's expected utility for a lottery is assumed to depend only on Adam and Eve's expected utilities for the lottery.

[3] I apologize for introducing a possible source of confusion. Weymark (1991) speaks of an impartial observer when referring to Harsanyi's *nonteleological* defense of utilitarianism.

Equation (13.1) implies that the function v_i is *linear* on its domain of definition D.[4] A standard result from elementary linear algebra then tells us that constants U and V must exist for which

$$u_i(x_A, x_E) = Ux_A + Vx_E.$$

The ideal observer will therefore make his choice from S by maximizing a suitably weighted sum of Adam and Eve's utilities. The common good he personifies is therefore utilitarian.

Why not maximize your own utility? When asked why a citizen should selflessly pursue the common good rather than pursue his own interests, Harsanyi (1977) has nothing to say except that the citizen has a moral obligation to do so. But is this not simply to say that he ought to do it because he ought to do it?

Personally, I think it a major error for utilitarians to fudge the issue of why citizens should pursue the aims of some ideal observer rather than furthering their own individual interests. The question that needs to be decided is whether utilitarianism is a moral system to be employed by *individuals* in regulating their interactions with others, or whether it is a set of tenets to be followed by a *government* that has the power to enforce its decrees. It is understandable that utilitarians are reluctant to argue that their doctrine should be forced down the throats of people who find it hard to swallow. They prefer to imagine a world in which their thoughts are embraced with open arms by all the citizens of a utilitarian state. As with Marxists, there is sometimes talk of the state withering away when the word has finally reached every heart. However, most utilitarians see the practical necessity of compulsion. As Mill (1859/1962) puts it: "For such actions as are prejudicial to the interests of others, the individual is accountable, and may be subjected to social or legal punishment."

My own view is that utilitarians would be wise to settle for the public policy option, which Hardin (1988) refers to as *institutional utilitarianism.* If those in power are inclined to personify the role of the government of

[4] If x is in D, then there exists a lottery $\mathbf{M}$ such that $x_A = u_A(\mathbf{M})$ and $x_E = u_E(\mathbf{M})$. A corresponding lottery $\mathbf{N}$ can be associated with any y in D. Let $\mathbf{L}$ be the compound lottery that yields the prize $\mathbf{M}$ with probability p and $\mathbf{N}$ with probability $1 - p$. Then the left-hand side of (13.1) is $\mathcal{E}u_i(\mathbf{L}) = p\mathcal{E}u_i(\mathbf{M}) + (1 - p)\mathcal{E}u_i(\mathbf{N}) = pv_i(x) + (1 - p)v_i(y)$. The right-hand side of (13.1) is given by $v_i(\mathcal{E}u_A(\mathbf{L}), \mathcal{E}u_E(\mathbf{L})) = v_i(pu_A(\mathbf{M}) + (1 - p)u_A(\mathbf{N}), pu_E(\mathbf{M}) + (1 - p)u_E(\mathbf{N})) = v_i(px + (1 - p)y)$. It follows that $v_i(px + (1 - p)y) = pv_i(x) + (1 - p)v_i(y)$. Mathematicians will recognize this equation as the requirement that v_i be affine. But $u_A(\mathcal{L}) = u_E(\mathcal{L}) = u_i(\mathcal{L}) = 0$. Hence, $v_i(0) = 0$, and so v_i is linear.

which they form a part, then they may be open to the suggestion that its actions should be rational in the same sense that an individual is rational. In a society with liberal traditions, Bentham's (1789/1987) "everyone to count for one, nobody to count for more than one" will also be attractive. If the powerful are persuaded by such propaganda, then Harsanyi's (1977) teleological argument will then require that the government act as though it had the preferences of a utilitarian ideal observer.

A bourgeois liberal like myself will remain unpersuaded, but not because we are repelled by the idea that citizens need to be coerced into compliance. As long as someone is guarding the guardians, we see coercion as a practical necessity in a large modern state. Who would pay their taxes on time and in full unless compelled to do so? Even Hayek (1960, p. 153) creates no difficulties on this score. Provided laws are framed impersonally, he is willing to say, "When we obey the laws . . . we are not subject to another man's will and are therefore free." There is therefore no reason why enforcement should be a painful issue for teleological utilitarians – provided that they are willing to grasp the nettle firmly by facing up to the intrinsic paternalism of their doctrine.

In summary, Harsanyi's (1977) teleological defense of utilitarianism applies best to the problem that welfare economics sets out to solve. In its idealized form, a benign but paternalistic government asks what behavior it should enforce on the citizens subject to its authority. Harsanyi's answer makes sense when the government regards itself as an individual written large who is immune to personal prejudice.

13.3 Nonteleological Moral Theories

Evolutionary biologists shrink with horror from teleological explanations of natural phenomena. The idea that evolution is designed to fulfill some a priori purpose is nothing less than heretical. Authors like myself, who offer naturalistic explanations of moral phenomena in human societies, feel much the same about teleological explanations of social phenomena. However, it is commonplace for critics to fail to understand a nonteleological approach to moral theory at all, since they see no point in discussing a common good function G unless the fact that $G(a) < G(b)$ provides a reason why a society should prefer state b to a. But a nonteleological theory follows Rawls (1971) in taking the *procedure* by means of which a society makes its decisions as fundamental. A common good function that arises in such a theory has no causal role. One simply asserts that to use the procedure under study is to behave *as though* maximizing a particular common good function. Rawls

(1971) refers to the process of constructing a common good function from a just procedural system as deducing the Good from the Right.

In this section, I plan to develop a version of Rawls's (1971) egalitarian theory alongside Harsanyi's (1977) nonteleological utilitarian theory. However, his use of the maximin principle as a criterion for making decisions in risky situations will be applied to utilities instead of an index of primary goods. Since the utility theory of Von Neumann and Morgenstern only makes sense when decision-makers maximize expected utility, the prognosis in the Rawlsian case is not promising. But we shall find that the conclusions to which we are led in this case make more sense after a radical restructuring of his approach.

Political legitimacy. The procedure I plan to study is based on the current consensus on political legitimacy in western democracies. The people use a fair process to deliver a mandate to the government, which then acts to enforce the laws that the people have made for themselves. In seeking to model this consensus, I invent an impartial philosopher-king, who is all-powerful but entirely benign. He has no largesse of his own to distribute, all the productive potential of the state being vested in the hands of his subjects. His role is therefore entirely organizational. First, he receives a mandate from the people to pursue certain ends, and then he insists that each person take whatever action is necessary to achieve these ends. In real life, people are only too ready to vote for an end, but against the means for attaining it. However, in a rational society, people will accept that working together toward an ambitious goal may require some surrender of their personal freedom. Without a philosopher-king to police their efforts at self-discipline, the citizens would have no choice but to rely on their own feeble powers of commitment to prevent any free-riding. The ends that they could jointly achieve would then be severely restricted. But with a philosopher-king to enforce the laws that they make for themselves, the citizens of a society will have a much larger feasible set open for them to exploit.

The fair procedure that the people employ to determine their mandate is the device of the original position. This idea is independently employed by both Harsanyi (1977) and Rawls (1971). The two citizens, Adam and Eve, pretend that their current and future roles in society are concealed behind a veil of ignorance. In this state, Adam becomes player 1 and Eve becomes player 2. Players 1 and 2 then negotiate a deal that I shall call a social contract. This social contract then serves as a system of rules for the philosopher-king to enforce. It is fair to the extent that it has been negotiated without Adam or Eve knowing who would occupy any position of privilege that the social contract may specify.

Neither Harsanyi (1977) nor Rawls (1971) feel the need to postulate a philosopher-king. Both argue that Adam and Eve are somehow committed to the hypothetical deal reached behind the veil of ignorance. However, it seems to me obvious that Hume (1739/1978, p. 156) was correct to argue that "a promise wou'd not be intelligible, before human conventions had established it ... even if it were intelligible, it wou'd not be attended with any obligation."

Empathetic preferences. Behind the veil of ignorance, Adam and Eve's personal preferences are common knowledge, but neither player 1 nor player 2 knows who he is. To evaluate social contracts in the original position, Harsanyi (1977) points out that they therefore need what I call empathetic preferences.

Empathetic preferences are identical to the extended sympathy preferences developed by Suppes (1966), Arrow (1978), Sen (1970), and Harsanyi (1977). Adam is expressing a personal preference when he says that he would rather have a fig leaf to wear than an apple to eat. However, if someone says that he would rather be Adam wearing a fig leaf than Eve eating an apple, then he is expressing an empathetic preference.

Modeling an individual's empathetic preferences using a Von Neumann and Morgenstern utility function v_i is easy, provided one bears in mind that his empathetic utility function is quite distinct from his personal utility function u_i. Let C be the set of possible consequences or prizes. Let $\{A, E\}$ be the set consisting of Adam (A) and Eve (E). A personal utility function u_i assigns a real number $u_i(\mathcal{C})$ to each $\mathcal{C}$ in the set C. By contrast, an empathetic utility function v_i assigns a real number $v_i(\mathcal{C}, j)$ to each pair $(\mathcal{C}, j)$ in the set $C \times \{A, E\}$. The number $u_i(\mathcal{C})$ is the utility the individual will get if $\mathcal{C}$ occurs. The number $v_i(\mathcal{D}, E)$ is the utility he would derive *if he were Eve* and $\mathcal{D}$ occurs. To write $u_i(\mathcal{C}) > u_i(\mathcal{D})$ means that the individual prefers $\mathcal{C}$ to $\mathcal{D}$. To write $v_i(\mathcal{C}, A) > v_i(\mathcal{D}, E)$ means that he would rather be Adam when $\mathcal{C}$ occurs than Eve when $\mathcal{D}$ occurs.

The zero and the unit on Von Neumann and Morgenstern utility scales can be chosen at will, but then our freedom for maneuver is exhausted. As previously, assume that everybody at least agrees that there is a worst outcome $\mathcal{L}$ and a best outcome $\mathcal{W}$ in the set C of feasible social contracts. Perhaps $\mathcal{L}$ is the event that everybody goes to hell and $\mathcal{W}$ is the event that everybody goes to heaven. We may then take $u_A(\mathcal{L}) = u_E(\mathcal{L}) = 0$ and $u_A(\mathcal{W}) = u_E(\mathcal{W}) = 1$.

Harsanyi (1977) argues that if an individual is totally successful in empathizing with Adam, then the preferences he will express when imagining himself in Adam's shoes will be identical to Adam's personal preferences. However, the Von Neumann and Morgenstern theory of expected utility tells us that two Von Neumann and Morgenstern utility scales which represent the same preferences can only differ in the location of their zeros and units. It follows that

$$v_i(\mathcal{C}, A) = \tilde{u}_A(\mathcal{C}) = \alpha u_A(\mathcal{C}) + \gamma\,, \tag{13.2}$$

where $\alpha > 0$ and γ are constant. Similarly, for suitable constants $\beta > 0$ and δ,

$$v_i(\mathcal{C}, E) = \tilde{u}_E(\mathcal{C}) = \beta u_E(\mathcal{C}) + \delta\,. \tag{13.3}$$

Although the zeros and units on Adam and Eve's personal scales have been fixed, the zero and unit on the observer's empathetic utility scale remain undetermined. Somewhat arbitrarily, I fix this scale so that $v_i(\mathcal{L}, A) = 0$ and $v_i(\mathcal{W}, E) = 1$. We are then not free to meddle anymore with the observer's empathetic utility scale. It follows that the two constants $U_i > 0$ and $V_i > 0$ defined by

$$U_i = v_i(\mathcal{W}, A) \text{ and } 1 - V_i = v_i(\mathcal{L}, E) \tag{13.4}$$

tell us something substantive about his empathetic preferences. Indeed, these two parameters characterize the observer's empathetic preferences. To see this, substitute the four values $v_i(\mathcal{L}, A) = 0$, $v_i(\mathcal{W}, E) = 1$, $v_i(\mathcal{W}, A) = U_i$ and $v_i(\mathcal{L}, E) = 1 - V_i$ into Eqs. (13.2) and (13.3). The result will be four equations in the four unknowns α, β, γ, and δ. Solve these equations and substitute the resulting values for α, β, γ, and δ back into Eqs. (13.2) and (13.3). We then find that the observer's empathetic utility function v_i can be expressed entirely in terms of the two parameters U_i and V_i:

$$\begin{aligned} v_i(\mathcal{C}, A) &= U_i u_A(\mathcal{C}), \\ v_i(\mathcal{C}, E) &= 1 - V_i\{1 - u_E(\mathcal{C})\}\,. \end{aligned} \tag{13.5}$$

Intrapersonal comparison of utility. Before interpreting these equations, we need to return to the three questions that open Section 13.2. The answer to the first question is still that utility is to be interpreted in the sense of Von Neumann and Morgenstern. The philosopher-king represents an external enforcement agency that provides an answer to the third question. But it may have passed unnoticed that the answer to the second question given in Section 13.2 will no longer suffice. The ideal observer's preferences incorporated a standard for making interpersonal comparisons of utility. But we no longer haved an ideal observer. The philosopher-king certainly will not suffice for this purpose because his entire mandate derives from the people. This also includes the standard for interpersonal comparison.

A major part of Harsanyi's (1977) achievement was to notice the relevance of empathetic preferences to this question. Equations (13.5) imply that U_i of Eve's utils are equivalent to V_i of Adam's utils. However, this standard for comparing utils is *intra*personal rather than *inter*personal. It is the observer's own idiosyncratic standard and need not be shared by anyone else. But how does it help that we should each have our own private intrapersonal standards for comparing utils if we are unable to agree on a common interpersonal standard to be applied when joint decisons are made? Nor is social choice theory very encouraging about the possibility of aggregating our intrapersonal standards. Hylland (1991) has shown that a version of Arrow's paradox applies, and so the the only aggregation methods that meet the usual criteria of social choice theory are essentially dictatorial.[5] Such a procedure fits comfortably into an ipsedixist philosophy. The ideal observer approach can also be accommodated if the ideal observer is treated as an imaginary citizen. But we are now exploring the notion of a philosopher-king who must derive *all* his standards from the people.

A serious problem therefore awaits attention. But this will be put on hold until a model has been developed that allows the bargaining problem behind the veil of ignorance to be analyzed. Two versions will be examined: one for Harsanyi and one for Rawls.

13.3.1 Harsanyi's Nonteleological Utilitarianism

All that will matter about a social contract $\mathcal{C}$ in the analysis that follows is that it generates a payoff pair $x = (u_A(\mathcal{C}), u_E(\mathcal{C}))$. Figure 13.1(a) shows the feasible set X of all such contracts. The set X is assumed to be closed, bounded above, and comprehensive for the usual reasons. The asymmetries of the set X register the ineradicable inequalities between Adam and Eve for which the device of the original position provides redress.

Behind the veil of ignorance, the players face an uncertain situation. They do not know which of two events, *AE* or *EA*, will be revealed after their negotiations are concluded. Because we chose the notation, we know that they will actually observe event *AE*, in which player 1 is Adam and

[5] Arrow's theorem is irrelevant in the later part of the chapter because the condition of unrestricted domain does not apply.

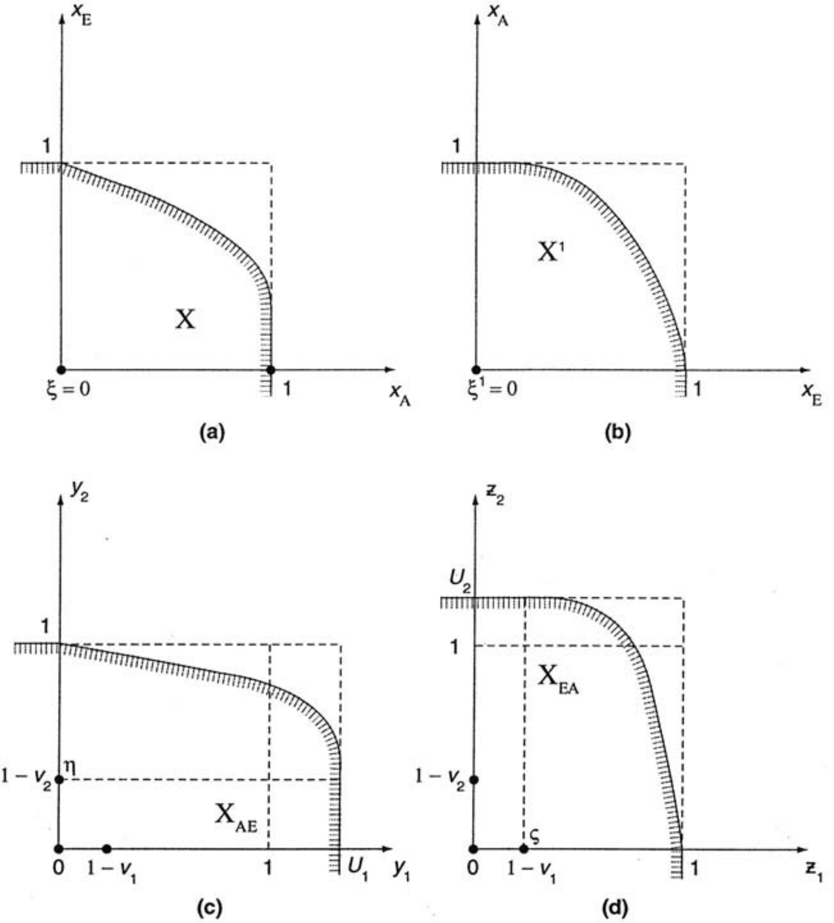


Figure 13.1. Various transformations of the set X.

player 2 is Eve. But the protagonists themselves must also take account of the event *EA* in which player 1 turns out to be Eve and player 2 to be Adam. Unlike Harsanyi (1977) and Rawls (1971), I allow Adam and Eve to come to an arrangement that makes the social contract contingent on who turns out to occupy which role. That is to say, they are assumed free to agree to operate one social contract $\mathcal{C}$ if event *AE* occurs and another social contract $\mathcal{D}$ if event *EA* occurs. This contingent social contract will be denoted by $(\mathcal{C}, \mathcal{D})$.

Suppose that player 1's preferences are given by his empathetic Von Neumann and Morgenstern utility function v_1. His expected utility for the contingent social contract $(\mathcal{C}, \mathcal{D})$ is then

$$w_1(\mathcal{C}, \mathcal{D}) = \tfrac{1}{2}v_1(\mathcal{C}, A) + \tfrac{1}{2}v_1(\mathcal{D}, E)\,. \tag{13.6}$$

In this expression, $v_1(\mathcal{C}, A)$ is the utility player 1 derives if the social contract $\mathcal{C}$ is operated with him in Adam's role. Similarly, $v_1(\mathcal{D}, E)$ is the utility he derives if the social contract $\mathcal{D}$ is operated with him in Eve's role.

We have seen that the empathetic utility function of a person who empathizes fully with both Adam and Eve can be expressed in terms of their personal utility functions u_A and u_E. Writing $w_1(\mathcal{C}, \mathcal{D})$ of (13.6) and the corresponding formula for $w_2(\mathcal{C}, \mathcal{D})$ in terms of Adam and Eve's *personal* preferences, we obtain that

$$w_1(\mathcal{C}, \mathcal{D}) = \tfrac{1}{2}U_1 u_A(\mathcal{C}) \;+\; \tfrac{1}{2}\{1 - V_1(1 - u_E(\mathcal{D}))\}\,, \tag{13.7}$$

$$w_2(\mathcal{C}, \mathcal{D}) = \tfrac{1}{2}\{1 - V_2(1 - u_E(\mathcal{C}))\} \;+\; \tfrac{1}{2}U_2 u_A(\mathcal{D})\,. \tag{13.8}$$

Bargaining in the original position. The simplest type of bargaining problem can be formulated as a pair (T, τ), where T consists of all payoff pairs on which the two players can agree, and τ is the payoff pair that will result if there is a disagreement. Our first problem is therefore to determine the set T of feasible payoff pairs for players 1 and 2 in the original position.

Figure 13.1(a) shows the set X of feasible personal payoffs pairs that Adam and Eve can achieve by coordinating on a suitable social contract. The point ξ corresponds to the state of nature – the payoff pair the players receive if they cannot agree on a social contract. We identify $\mathcal{L}$ with the state of nature to allow the normalization $\xi = 0$. Behind the veil of ignorance, players 1 and 2 attach probability 1/2 to both events *AE* and *EA*. Since they evaluate an uncertain prospect by calculating its expected utility, they regard a contingent social contract that leads to the payoff pair y when *AE* occurs and z when *EA* occurs as equivalent to the pair

$$t = \tfrac{1}{2}y + \tfrac{1}{2}z\,, \tag{13.9}$$

which is simply a compressed version of Eqs. (13.7) and (13.8).

The payoff pair t is shown in Figure 13.2(a). It lies halfway along the line segment joining the payoff pairs y and z. The pair y lies in the set X_{AE} consisting of all payoff pairs that players 1 and 2 regard as attainable should *AE* occur. The set X_{AE} has a similar shape to X, but needs to be rescaled to reflect the relative worth that player 1 places on Adam's utils and player 2

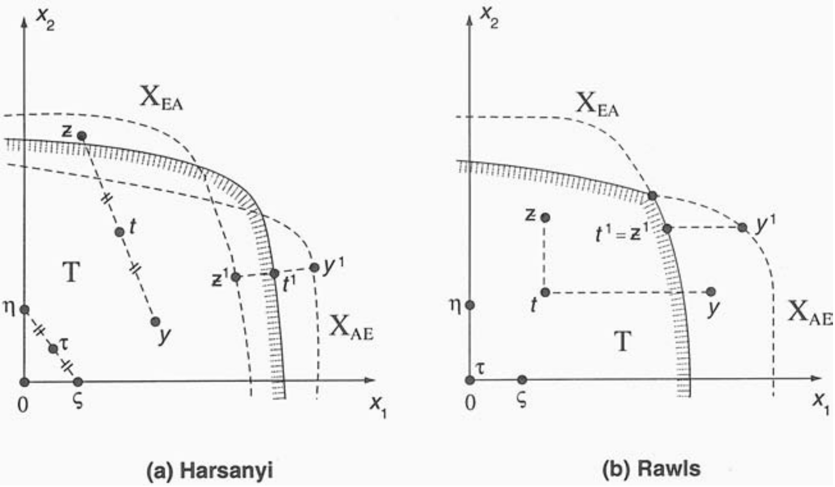


Figure 13.2. Constructing the set T.

places on Eve's utils. From Eq. (13.5), we know that player 1 regards a payoff of x_A to Adam as being worth $y_1 = U_1 x_A$. Similarly, player 2 regards a payoff of x_E to Eve as being worth $y_2 = 1 - V_2(1 - x_E)$. To obtain X_{AE} from the set X, we must therefore replace each payoff pair $x = (x_A, x_E)$ in X by the rescaled pair $y = (y_1, y_2) = (U_1 x_A, 1 - V_2(1 - x_E))$. The resulting set X_{AE} is shown in Figure 13.1(c).

To obtain X_{EA} from the set X is slightly more complicated because player I will become Eve if EA occurs. However, his payoffs are measured on the horizontal axis while hers are measured on the vertical axis. As shown in Figure 13.1(b), it is therefore necessary to begin by swapping over Adam and Eve's axes to obtain the set X' (which is simply the reflection of X in the line $x_A = x_E$). After this transformation, player 1's payoffs and Eve's payoffs are both measured on the horizontal axis, and so we can proceed as before. The set X_{EA} has a similar shape to X', but it needs to be rescaled in order to reflect the relative worth that player 1 places on Eve's utils and player 2 places on Adam's utils. To obtain X_{EA} from X', replace each payoff pair $x = (x_E, x_A)$ in X' by the rescaled pair $z = (z_1, z_2) = (1 - V_1(1 - x_E), U_2 x_A)$. The set X_{EA} is shown in Figure 13.1(d).

The preceding discussion of how X_{AE} and X_{EA} are constructed from X is a necessary preliminary to drawing the set T of all payoff pairs that are feasible for players 1 and 2 in the original position. As Figure 13.2(a) illustrates, T is the set of all $t = \frac{1}{2}y + \frac{1}{2}z$, with y in X_{AE} and z in X_{EA}. (The Pareto frontier of T can be characterized as the set of all $t = \frac{1}{2}y + \frac{1}{2}z$ with the property

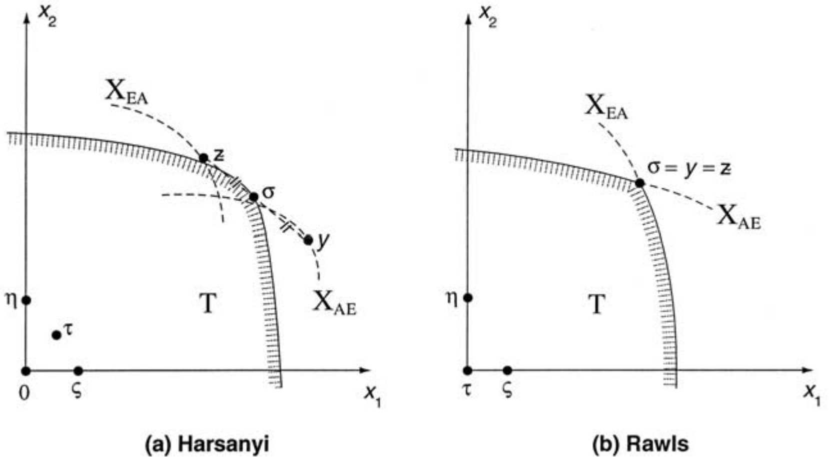


Figure 13.3. Applying the Nash bargaining solution to (T, τ).

that the tangent to the Pareto frontier of X_{AE} at y has the same slope as the tangent to the Pareto frontier of X_{EA} at z.)

Having tied down the set T, the next step in specifying the bargaining problem (T, τ) is to determine the status quo τ. Chapter 2 of Binmore (1998) describes how Nash's variable threats theory can be used in principle to determine τ. But such considerations are short-circuited here by simply assuming that Adam and Eve's alternative to agreeing to a social contract is that they revert to the state of nature represented in Figure 13.1(a) as the payoff pair $\xi = (\xi_A, \xi_E)$. This is mirrored in Figure 13.1(b) by the payoff pair $\xi' = (\xi_E, \xi_A)$. Behind the veil of ignorance, players 1 and 2 therefore evaluate the consequences of a disagreement as being equivalent to the payoff pair $\tau = \frac{1}{2}\eta + \frac{1}{2}\zeta$ illustrated in Figure 13.2(a).

Solving the bargaining problem. The bargaining problem faced by players 1 and 2 in the original position has been formulated as the pair (T, τ) shown in Figure 13.2(a). It is easy to describe the solution to this bargaining problem in geometric terms. As illustrated in Figure 13.3(a), Nash's theory of bargaining with commitment predicts that the bargaining outcome will be the symmetric Nash bargaining solution σ for the bargaining problem (T, τ).

Before discussing what this solution implies for Adam and Eve's personal payoffs, it is as well to emphasize the strong informational assumptions required by the argument leading to the payoff pair σ in Figure 13.3(a). In the original position, players 1 and 2 forget whether they are Adam or Eve.

Since their own empathetic preferences are common knowledge, it follows that they must also forget which empathetic preference derives from Adam and which from Eve. Everything else is assumed to be common knowledge between players 1 and 2. This assumption is essential in the case of *all* the data used to construct the bargaining problem (T, τ). Players 1 and 2 therefore know the rules of the game of life, and hence which potential social contracts are feasible. Each also knows Adam and Eve's personal preferences, and his own empathetic preferences together with those of his bargaining partner.

Returning to the question of how an agreement on σ in the original position translates into personal payoffs to Adam and Eve, it is necessary to recall that for σ to be admissible as a member of the set T, it must be of the form $\sigma = \frac{1}{2}y + \frac{1}{2}z$, where y is in X_{AE} and z is in X_{EA}. One must also remember that the bargaining that supposedly takes place behind the veil of ignorance is only hypothetical. Adam and Eve only pretend to forget their identities when using the device of the original position to compute a fair social contract. In fact, player 1 is actually Adam and player 2 is actually Eve. Of the two events *AE* and *EA*, it is therefore the former that actually obtains.

It follows that the social contract $\mathcal{C}$ that will actually be operated corresponds to the payoff pair $y = (y_1, y_2)$ in X_{AE} illustrated in Figure 13.3(a). In terms of Adam and Eve's original personal utility scales, the social contract $\mathcal{C}$ yields the payoff pair $h = (h_A, h_E)$ defined by $y_1 = U_1 h_A$ and $y_2 = 1 - V_2(1 - h_E)$. As far as I know, there is no neat way to summarize the payoff pair h in terms of the set X and the underlying game of life. However, matters become more promising in the symmetric case illustrated in Figure 13.4(a).

The symmetric case. When $U_1 = U_2 = U$ and $V_1 = V_2 = V$, Figure 13.3(a) translates into the symmetric Figure 13.4(a). In particular, the bargaining problem (T, τ) becomes symmetric, and so the outcome

$$N = \tfrac{1}{2}H + \tfrac{1}{2}K \tag{13.10}$$

obtained by using the symmetric Nash bargaining solution is symmetric as well.

The symmetry ensures that the payoff pair N can be achieved using the same social contract $\mathcal{C}$ whether *AE* or *EA* occurs. However, the event that actually obtains is *AE* and so the personal payoff pair $h = (h_A, h_E)$ that Adam and Eve actually experience when $\mathcal{C}$ is implemented is given by $H_1 = U h_A$ and $H_2 = 1 - V(1 - h_E)$.

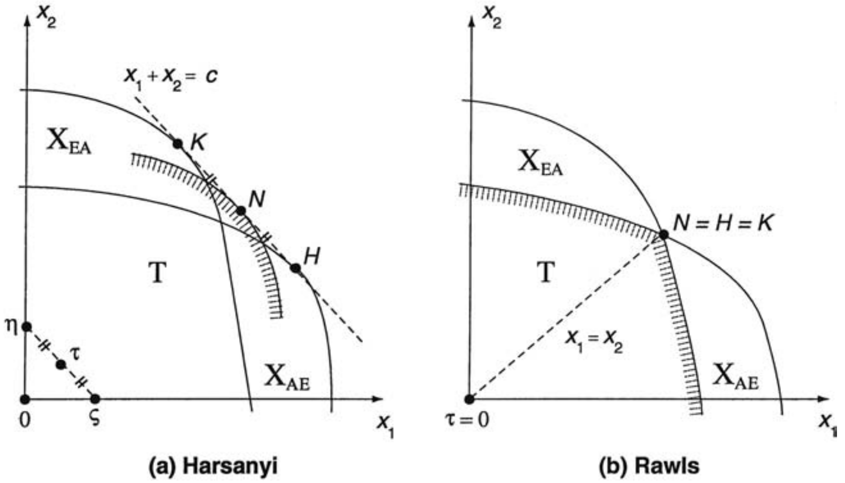


Figure 13.4. The symmetric case: $U_1 = U_2 = U$ and $V_1 = V_2 = V$.

In the asymmetric case, it proved difficult to characterize the payoff pair h as a point of X. But here it is easily identified as the point x in X at which the weighted utilitarian social welfare function

$$W_h(x) = Ux_A + Vx_E$$

is maximized. To see this, observe that H and K in Figure 13.4 lie on a common tangent $x_1 + x_2 = c$ to the Pareto-frontiers of X_{AE} and X_{EA}. It follows that H is the point in X_{AE} at which the social welfare function $W_H(x) = x_1 + x_2$ is maximized. But the function defined by $x_1 = Ux_A$ and $x_2 = 1 - V(1 - x_E)$ that maps X_{AE} to X transforms $x_1 + x_2$ into $Ux_A + 1 - V(1 - x_E)$. The constant $1 - V$ is irrelevant to the maximization, and so to maximize W_H on X_{AE} is the same as maximizing W_h on X. Figure 13.5(a) shows the location of h as the point x in X at which W_h is maximized.

The argument so far generalizes Harsanyi's nonteleological defense of utilitarianism to the case of contingent social contracts, provided that one is willing to swallow the assumption that $U_1 = U_2 = U$ and $V_1 = V_2 = V$. But to make this assumption is to accept that Adam and Eve's intrapersonal standards for making utility comparisons are the same, and hence provide a basis for an interpersonal standard. The question of how Harsanyi (1977) justifies this large assumption is put aside until Section 13.3.4, while the Rawlsian model is advanced to the same stage.

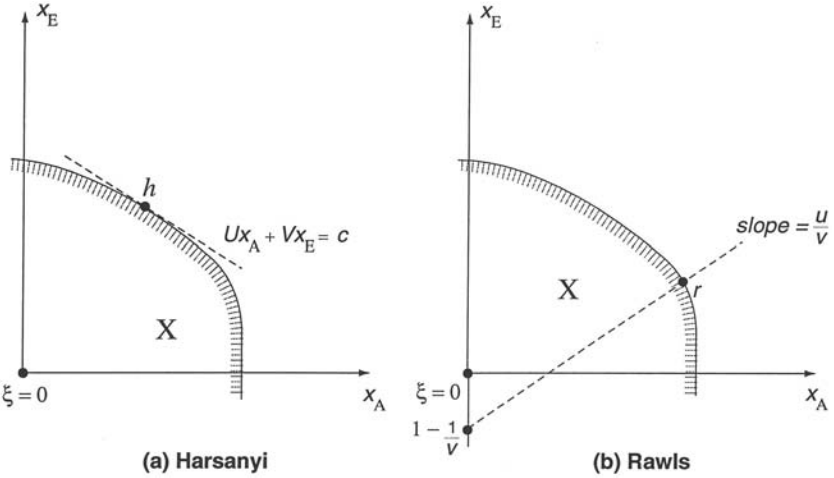


Figure 13.5. Utilitarian and Rawlsian solutions.

13.3.2 Rawls' Model

Recall that Rawls (1971) replaces Harsanyi's (1977) use of Bayesian decision theory by the maximin principle. Harsanyi's model must therefore be modified so that a player in the original position proceeds as though whichever of the two events *AE* and *EA* he dislikes more were certain to occur. This modification has no impact on the analysis until Eq. (13.9) is reached. At this point, it is necessary to diverge from Harsanyi's argument because we are no longer to make the Bayesian assumption that player i seeks to maximize $t_i = \frac{1}{2}y_i + \frac{1}{2}z_i$. Instead, we must take the most pessimistic of all possible views and assume that player i seeks to maximize

$$t_i = \min\{y_i, z_i\}\,. \tag{13.11}$$

The payoff pair $t = (t_1, t_2)$ defined by Eq. (13.11) is shown in Figure 13.2(b).

The set T is easier to describe than in Harsanyi's case because it is simply the set of all payoff pairs that lie in both X_{AE} and X_{EA}. That is to say, $T = X_{AE} \cap X_{EA}$. Figure 13.3(b) shows how to compute the symmetric Nash bargaining solution σ in this new case. When the situation is not too far from being symmetric, σ lies at the point where the Pareto frontiers of X_{AE} and X_{EA} cross. Figure 13.4(b) illustrates the fully symmetric case when $U_1 = U_2 = U$ and $V_1 = V_2 = V$.

The personal payoff pair $r = (r_A, r_E)$ shown in Figure 13.5(b) tells us how Adam and Eve evaluate the social contract $\mathcal{C}$ after emerging from behind the veil of ignorance to discover that *AE* actually obtains. It is determined by the requirement that

$$Ur_A = 1 - V(1 - r_E)\,.$$

One may characterize r as the proportional bargaining solution with weights U and V for the bargaining problem (X, α) in which the status quo α is $(0, 1 - 1/V)$.

I think that the emergence of such a nonsensical status quo simply signals that something is awry with the underlying assumptions. Nor is the problem hard to find. As chapter 4 of Binmore (1994) argues at length, Rawls's (1971) reasons for favoring the maximin principle as a criterion for making decisions under risk lack any conviction. However, if he were to maintain his other assumptions while replacing the maximin principle by Bayesian decision theory, then his argument would reduce to Harsanyi's. Rawls would then find that he was a utilitarian!

Retelling the Rawlsian tale without a philosopher-king. However, I do not think that egalitarianism should therefore be rejected as a viable option. On the contrary, a sound nonteleological defense of egalitarianism can be obtained by taking Rawls's (1971) concerns about the "strains of commitment" to their logical extreme. This means abandoning the fiction of the philosopher-king altogether. After all, when constitutional issues are in question, there is no external enforcement agency to whom one can appeal. But without a philosopher-king, any social contract must be self-policing.

Binmore (1994, 1998) models such self-policing social contracts as equilibria in a repeated game. The device of the original position then serves simply to coordinate behavior on one of these equilibria. But the issue on which I want to concentrate here is much simpler. Without commitment, the players have no reason to honor the fall of the hypothetical coin that determines their role in the social contract. If it falls to their disadvantage, why do they not call for it to be tossed again? This simple question is fatal to the logic of the original position unless the feasible set of contingent social contracts is restricted to cases where players 1 and 2 each expect the *same* payoff whether *AE* or *EA* occurs. Neither can then gain an advantage from tossing the coin again.

The consequences of restricting the feasible set in this way are severe. Equation (13.9) is no longer replaced by Eq. (13.11). We still require that $t = \frac{1}{2}y + \frac{1}{2}z$, but this equation must be supplemented by the constraint $y = z$. It follows that t must belong to both X_{AE} and X_{EA}. The conclusion that $T = X_{AE} \cap X_{EA}$, which Rawls obtained by appealing to the maximin principle, is therefore obtained here by throwing out all commitment assumptions that supposedly constrain the players in the models of Harsanyi and Rawls.

Similar considerations also lead to the conclusion that $\xi = \eta = \zeta$. Thus $V_1 = V_2 = 1$. Recycling the argument of the Rawlsian model, we are therefore led again to the proportional bargaining solution r of Figure 13.5(b) with weights U and V, but now applied to the bargaining problem (X, ξ). If X is strictly comprehensive, r is the point at which the Rawlsian social welfare function

$$W_r(x) = \min\{U(x_A - \xi_A),\ V(x_E - \xi_E)\}$$

is maximized subject to the requirement that x lies in X.

In brief, Rawls's moral intuition is vindicated, but at a very heavy cost for a theory supposedly based on deontological principles. My justification denies, not only that we have a natural duty to honor hypothetical deals reached in the original position, but that we have any natural duties at all.

13.3.3 Kantian Foundations for Interpersonal Comparison

The nonteleological defenses of Harsanyi and Rawls I have offered rely on the players having the same intrapersonal standards for comparing utilities. Their common attitude than defines an *inter*personal standard for this purpose. But why should we assume that $U_1 = U_2 = U$ and $V_1 = V_2 = V$? And what determines the values of U and V?

Rawls (1971) deploys his index of primary goods as though this problem did not exist, but Harsanyi (1977) is more true to the Kantian ideas that both take as their foundation stone. He postulates a very thick veil of ignorance behind which people forget the empathetic preferences they have in real life and so find it necessary to adopt new empathetic preferences. He then appeals to a principle that has become known as the *Harsanyi doctrine.* In its extended form, the doctrine asserts that rational individuals placed in identical situations will necessarily respond identically. The empathetic preferences that Adam and Eve adopt in the original position will therefore be the same, and the problem of interpersonal comparison is solved. Since both players will be led to the same standard for comparing utils, no difficulties can arise in taking this as their common standard.

Binmore (1994, p. 210) criticizes the use of the Harsanyi doctrine in such a context at length. However, even if its use could be adequately defended, how would poor mortals like ourselves be able to predict the empathetic preferences that the super-rational players of game theory would adopt behind the thick veil of ignorance envisaged by Harsanyi? Even if we could, why should we substitute these empathetic preferences for those we already have? I think that the fact that Kantians have no answers to these questions is fatal for their endeavor.

My own view is that the reason that the device of the original position is so attractive as a fairness criterion is not because the Kantian arguments offered in its favor are overwhelmingly convincing. In fact, they seem to me so tenuous as to be almost invisible. I think we are attracted to the original position because it captures in a stylized form the essence of do-as-you-would-be-done-by principles that are *already* firmly entenched as joint decision-making criteria within the system of commonly understood conventions that bind society together. It is distinguished from other versions by its power to answer objections like: don't do unto others as you would have them do unto you – they may have different tastes from yours. I think that we have empathetic preferences at all only because we need them as inputs when using rough-and-ready versions of the device of the original position to make fairness judgments in real life. Insofar as people from similar cultural backgounds have similar empathetic preferences, it is because the use of the original position in this way creates evolutionary pressures that favor some empathetic preferences at the expense of others. In the medium run, the result is that everybody tends to have the same set of empathetic preferences.

Our guide to analyzing the device of the original position should therefore be how it is actually used in real life. If so, then we must not follow Harsanyi and Rawls in postulating a thick veil of ignorance. On the contrary, the veil of ignorance needs to be taken to be as thin as possible. Adam and Eve must certainly still forget their personal preferences along with their identities, but it is essential that they do not forget the *empathetic* preferences with which their culture has equipped them. To isolate Adam and Eve in a Kantian void from the cultural data summarized in their empathetic preferences, and then to ask them to make interpersonal comparisons seems to me like inviting someone to participate in a pole-vault competition without a pole.

13.4 The Original Position as a Natural Norm

Chapter 2 of Binmore (1998) sketches a possible evolutionary history for the device of the original position as a fairness norm. Very briefly, it is argued that the origins of the device are to be found in primitive food-sharing agreements. If player 1 is lucky enough to have an excess of food this week, then it makes sense for him to share with player 2 in the expectation that she will be similarly generous when she is lucky in the future.

To see the similarity between bargaining over mutual insurance pacts and bargaining behind the veil of ignorance, think of players 1 and 2 as not knowing whether tomorrow will find them occupying the role of Mr. Lucky or Ms. Unlucky. It then becomes clear that a move to the device of the original position requires only that the players put themselves in the shoes of somebody else – either Adam or Eve – rather than in the shoes of one of their own possible future selves. The same distinction separates Buchanan and Tullock's (1962) "veil of uncertainty" from Rawls' (1971) veil of ignorance. Dworkin (1981b) similarly distinguishes between "brute luck" and "opportunity luck".

But what of the origins of the capacity to empathize with a fellow man rather than a possible future self ? On this subject, one has to look no further than the relationships that hold within families. In accordance with Hamilton's (1963, 1964) rule for games played within families, each player's payoff consists of a weighted sum of the fitnesses of himself and his kinfolk, with each weight equal to his degree of relationship to the relative concerned. For example, the weight that Adam will attach to the fitness of a first cousin is 1/8 because they share this fraction of their genes. To express empathetic preferences outside the family, Adam has therefore only to adapt the mechanisms that evolved within the family to a new purpose.

However, a problem still remains. The weights we use when discounting the fitnesses of our partners in a family game are somehow obtained from estimating our degree of relationship to our kinfolk from the general dynamics of the family and our place in it. But where do we get the weights to be used when discounting Adam and Eve's personal utils in constructing an empathetic utility function?

I see the empathetic preferences held by the individuals in a particular society as an artifact of their upbringing. As children mature, they are assimilated to the culture in which they grow up largely as a consequence of their natural disposition to imitate those around them. One of the social phenomena they will observe is the use of the device of the original position in achieving fair compromises. They are, of course, no more likely to recognize the device of the original position for what it is, than we are when we use it in deciding such matters as who should wash how many dishes. Instead, they simply copy the behavior patterns of those they see using the device. An internal algorithm then distills this behavior into a preference-belief model against which they then test alternative patterns of behavior. The preferences in this model will be empathetic preferences – the inputs required when the device of the original position is employed.

I plan to short-circuit the complexities of the actual transmission mechanism by simply thinking of a set of empathetic preferences as being packaged in a social signal or meme – which is Dawkins's (1976) name for the social equivalent of a gene. The imitative process is seen as a means of propogating such memes in much the same way that the common cold virus finds its way from one head to another. As always, I keep things simple by assuming that all games to be played are games of complete information, which means that the rules of the game and the preferences of the players are taken to be common knowledge. In particular, it is assumed that the hypothetical bargaining game played behind the veil of ignorance has complete information. The empathetic preferences with which the players evaluate their predicament in the original position are therefore taken to be common knowledge. Along with a set of empathetic preferences, a meme will also carry a recommendation about which bargaining strategy to use in the original position. Only when the stability of the system is threatened by the appearance of a "mutant" meme will they have reason to deviate from this normal bargaining strategy.

To explore the issue of evolutionary stability, imagine that all players are currently controlled by a normal meme *N*. A mutant meme *M* now appears. Will it be able to expand from its initial foothold in the population? If so, then the normal population is not evolutionarily stable.

Only one of the two standard conditions for evolutionary stability is needed here – namely, the condition that (*N*, *N*) should be a *symmetric* Nash equilibrium of the underlying game in which *M* and *N* are strategies (Binmore, 1992, p. 414). In brief, *N* must be a best reply to itself.

To interpret this condition, imagine that player 1 has been infected with the mutant meme *M* while player 2 remains in thrall to the normal meme *N*. Both players will test their recommended bargaining strategy against the empathetic preferences they find themselves holding, and adjust their behavior until they reach a Nash equilibrium of their bargaining game. I shall be assuming that this Nash equilibrium implements a suitable version of the Nash bargaining solution of the game. As a consequence, player 1 and player 2 will each now receive some share of the benefits and burdens in dispute.

The imitation mechanism that determines when it is appropriate to copy the memes we observe others using will take account of who gets what. Onlookers will almost all currently be subject to the normal meme *N* and so will evaluate the shares they see players 1 and 2 receiving in terms of the empathetic preferences embedded in *N*. If player 1's payoff exceeds player 2's, then I assume that onlookers who are vulnerable to infection are more likely to be taken over by the meme *M* controlling player 1 than by the meme *N* controlling player 2. But then *M* will be a better reply to *N* than *N* is to itself.

A necessary condition for the evolutionary stability of a normal population is therefore that the empathetic preferences originally held by players 1 and 2 constitute what I call a symmetric *empathy equilibrium.* To test whether a pair of empathetic preferences constitutes an empathy equilibrium each player should be asked the following question:

> Suppose that you could deceive everybody into believing that your empathetic preferences are whatever you find it expedient to claim them to be. Would such an act of deceit seem worthwhile to you *in the original position* relative to the empathetic preferences that *you actually hold?*

The right answer for an empathy equilibrium is *no.*

Although the relevant mathematics are suppressed in this chapter, I believe that the fact that insight can be obtained into this question using the concept of an empathy equilibrium is one of the major advantages of my approach. For both Harsanyi's model and the reconstructed Rawlsian model without commitment, Binmore (1998) demonstrates that, at a symmetric empathy equilibrium, the personal payoffs Adam and Eve receive as a result of bargaining as though behind the veil of ignorance are precisely the same as they would have gotten if they had solved the bargaining problem (X, ξ) directly using the symmetric Nash bargaining solution. When allowed to operate for long enough, the effect of social evolution is therefore to leach out all moral content from the device of the original position.

13.4.1 Interpersonal Comparison in the Medium Run

It is helpful to adapt the distinction between short-run and long-run time periods used in the economic theory of the firm. The short run corresponds to economic time – the time period in which decisions are made. As is standard in economics, all preferences are assumed to be fixed in the short run. The long run corresponds to biological time, during which I assume that personal preferences are able to evolve. However, this chapter is more interested in the medium run, which corresponds to social time. In the medium run, empathetic preferences are assumed to evolve to an empathy equilibrium while personal preferences remain effectively fixed.

We have just seen that the result of social evolution operating in the medium run is that Adam and Eve will get precisely the same personal payoffs if they play fair by using the device of the original position as they would if they were to bargain face-to-face with no holds barred. So what use is a fairness norm if it serves only to conceal the iron fist in a velvet glove?

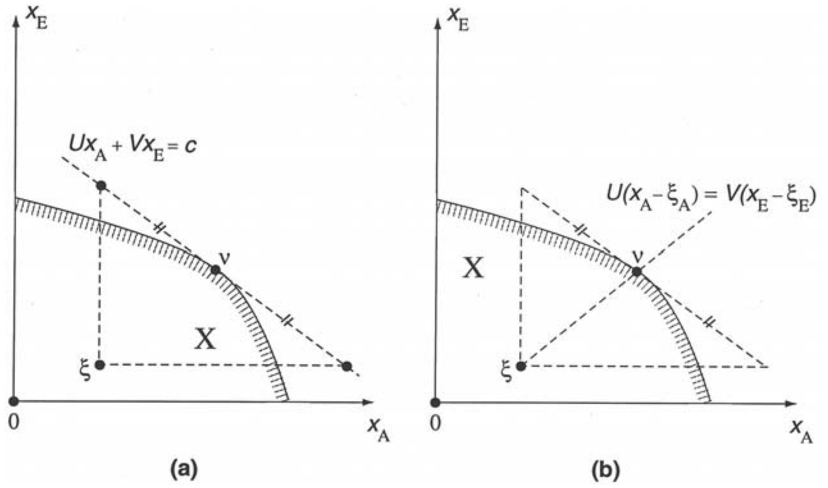


Figure 13.6. Interpersonal comparison in the medium run.

The answer is that the type of morality with which we are concerned has its bite in the *short run.*

To understand how I see fairness norms operating in practice, one must begin by imagining that groups of people assembled in different places at different times for various purposes find themselves continually facing the need to coordinate on Pareto-efficient solutions to new problems. Such minisocieties are simplified in my treatment to pairs of men and women seeking some *modus vivendi* that I refer to as a social contract. In our Garden of Eden fable, Adam is therefore a representative man and Eve is a representative woman.

To keep things simple, it will be assumed that all pairs always face the *same* set X of feasible social contracts, and the same state of nature ξ. Each Adam and Eve choose a social contract using the device of the original position. It will also be assumed that social evolution operates in the medium run to shape the manner in which Adam and Eve make interpersonal comparisons of utility. Eventually, everybody will therefore use the same weights U and V when comparing Adam's utils with Eve's.

In Harsanyi's model, the ratio U/V can then be computed as shown in Figure 13.6(a). One simply selects weights U and V so that the weighted utilitarian solution for the problem (X, ξ) coincides with the symmetric Nash bargaining solution ν. In the reconstructed version of Rawls' model without commitment, the constant V is fixed to be one, but U is computed as shown in Figure 13.6(b). In this case, the weights U and $V = 1$ are selected so

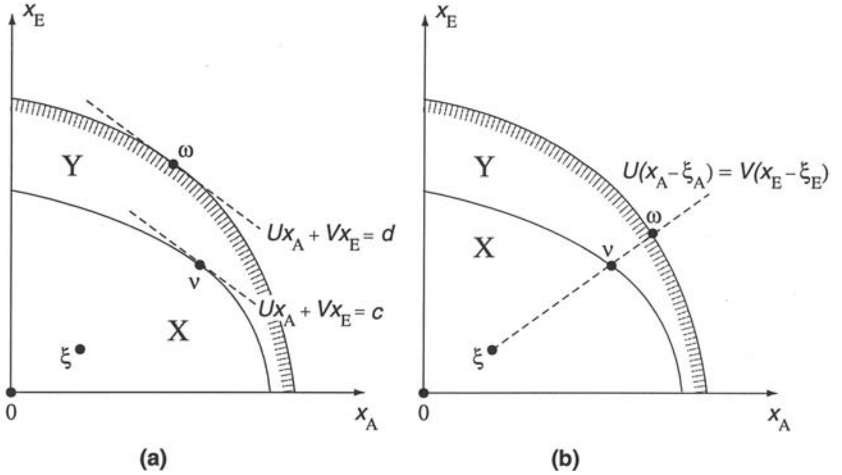


Figure 13.7. Morality in the short run.

that the proportional bargaining solution for the problem (X, ξ) coincides with the symmetric Nash bargaining solution ν.

Yaari (1981) has shown that if any two of the weighted utilitarian solution, the weighted proportional bargaining solution, and the symmetric Nash bargaining solution coincide, then they all coincide. It follows that utilitarians and egalitarians will not only agree on how to make interpersonal comparisons in the medium run, they will recommend the same actions!

13.4.2 Morality as a Short-Run Phenomenon

What does it matter what U and V are, since we can determine the Nash bargaining solution ν of (X, ξ) without their aid? The answer is that the values of U and V are relevant *in the short run* after some change in the underlying environment alters the set of feasible contracts. Perhaps some new innovation results in the set of available social contracts expanding from X to Y, as illustrated in Figure 13.7(a) and 13.7(b). The fairness norm being operated now has a chance to fulfill the function for which it originally evolved – to shift its mini-society to a new Pareto-efficient social contract ω without damaging internal conflict. In the short run, U and V remain fixed, and so the new social contract ω is located as shown in Figure 13.7(a) for the Harsanyi model and in Figure 13.7(b) for the reconstructed Rawls model.

Of course, if the set of feasible social contracts faced by Adam and Eve continues to be Y for long enough, then the standard for making interpersonal comparisons will adjust to the new situation, and so the moral content of the social contract will again be eroded away. But it would be wrong to deduce that morality has only a small and ephemeral role to play in regulating the conduct of our affairs. If matters seem otherwise, it is because we mislead ourselves by thinking of morality as something to be taken out of its glass case only when grand and difficult problems need to be addressed. The real truth is that we use our inbuilt sense of justice all the time in resolving the innumerable short-run coordination problems that continually arise as we try to get along with those around us. Such coordinating problems are usually so mundane and we solve them so effortlessly that we do not even think of them as problems – let alone moral problems. Like Molière's Monsieur Jourdain, who was delighted to discover that he had been speaking prose all his life, we are moral without knowing that we are moral.

Although I find few takers for the claim, I think the observation that morality works so smoothly much of the time that we don't even notice it working is of considerable significance. Just as we only take note of a thumb when it is sore, so moral philosophers tend to notice moral rules only when attempts are made to apply them in situations for which they are ill-adapted. As an analogy, consider Konrad Lorenz's observations of a totally inexperienced baby jackdaw going through all the motions of taking a bath when placed on a marble-topped table. By triggering such instinctive behavior under pathological circumstances, he learned a great deal about what is instinctive and what is not when a bird takes a bath. But this vital information is gained only by avoiding the mistake of supposing that bath-taking behavior confers some evolutionary advantage on birds placed on marble-topped tables. Similarly, one can learn a great deal about the mechanics of moral algorithms by triggering them under pathological circumstances but only if one does not make the mistake of supposing that the moral rules have been designed to cope with pathological problems.

13.5 Reform

The naturalistic account of the original position that I have hastily summarized is descriptive in intent. Evolutionary ethics recognizes no unconditional imperatives. Only metaphysically minded moral philosophers believe that they are able to *justify* their prescriptive recommendations in a manner free from cultural bias. In the suggestion made in this section, I simply observe that a useful social tool has been washed up on the beach by the tide of evolution and invite others with cultural prejudices similar to mine to join in using it to advance our common cause.

We do not need to confine the device of the original position to the small-scale problems for which it evolved. We can deliberately seek to expand its circle of application by trying to apply this familiar social tool on a larger scale. That is to say, we can try to achieve Pareto-improving solutions to large-scale coordinating problems by appealing to the same fairness criteria that we use to solve small-scale coordinating problems. But such an enterprise will not work unless we put aside the temptation to romanticize our fairness intuitions. In particular, people make interpersonal comparisons of utility as they do – not as we would wish them to.

If my descriptive theory is anywhere near correct, appeals to fairness that ignore the realities of power are doomed, because the underlying balance of power is what ultimately shapes the interpersonal comparisons necessary for fairness judgments to be meaningful. Philosophers with utopian ambitions for the human race tell me that such conclusions are unacceptable. But I think this is just another example of an argument being rejected because it has unwelcome implications. In particular, the fact that fairness norms do not work like utopian thinkers would like them to work should not discourage us from trying to use them in the manner in which they actually do work. Others are free to toy with grandiose plans to convert our planet into a new Jerusalem, but bourgeois liberals like myself are content to aim at finding workable ways of making life just a little bit more bearable for everyone.

References

Arrow, K. J. 1978. Extended sympathy and the problem of social choice. *Philosophia* 7, 233–237.

Bentham, J. 1987. An introduction to the principles of morals and legislation. In *Utilitarianism and Other Essays.* Penguin, Harmondsworth, UK. (Introduction by A. Ryan. Essay first published 1789.)

Binmore, K. 1992. *Fun and Games.* D. C. Heath, Lexington, MA.

Binmore, K. 1994. *Game Theory and the Social Contract*, Vol. 1: *Playing Fair.* MIT Press, Cambridge, MA.

Binmore, K. 1998. *Game Theory and the Social Contract*, Vol. 2: *Just Playing.* MIT Press, Cambridge, MA.

Binmore, K. 2005. *Natural Justice.* Oxford University Press, New York.

Broome, J. 1991. *Weighing Goods.* Blackwell, Oxford.

Buchanan, J. M., and Tullock, G. 1962. *The Calculus of Consent.* University of Michigan Press, Ann Arbor.

Cohen, G. A. 1993. Equality of what? On welfare, goods, and capabilities. In *The Quality of Life*, ed. M. Nussbaum and A. Sen. Clarendon Press, Oxford, pp. 9–29.

Dawkins, R. 1976. *The Selfish Gene.* Oxford University Press, Oxford.

Dworkin, R. 1981a. What is equality? Part 1: Equality of welfare. *Philosophy & Public Affairs* 10, 185–246.

Dworkin, R. 1981. What is equality? Part 2: Equality of resources. *Philosophy & Public Affairs* 10, 283–345.

Hamilton, W. D. 1963. The evolution of altruistic behavior. *American Naturalist* 97, 354–356.

Hamilton, W. D. 1964. The genetic evolution of social behaviour. Parts 1 and 2. *Journal of Theoretical Biology* 7, 1–52.

Hammond, P. J. 1988. Consequentionalist foundations for expected utility. *Theory and Decision* 25, 25–78.

Hammond, P. J. 1992. Harsanyi's utilitarian theorem: A simpler proof and some ethical connotations. In *Rational Interaction,* ed. R. Selten. Springer-Verlag, Berlin, pp. 305–319.

Hardin, R. 1988. *Morality within the Limits of Reason.* University of Chicago Press, Chicago.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, Cambridge.

Hayek, F. A. 1960. *The Constitution of Liberty.* University of Chicago Press, Chicago.

Hume, D. 1978. *A Treatise of Human Nature,* 2nd ed. Clarendon Press, Oxford. (Edited by L. A. Selby-Bigge. Revised by P. H. Nidditch. First published 1739.)

Hylland, A. 1991. Subjective interpersonal comparisons. In *Interpersonal Comparisons of Well-Being,* ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge. pp. 337–370.

Luce, R. D., and Raiffa, H. 1957. *Games and Decisions.* Wiley, New York.

Maskin, E. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45, 33–96.

Mill, J. S. 1962. On liberty. In *Utilitarianism.* Collins, London. (Edited by M. Warnock. Essay first published 1859.)

Rawls, J. 1971. *A Theory of Justice.* Harvard University Press, Cambridge, MA.

Roemer, J. E. 1996. *Theories of Distributive Justice.* Harvard University Press, Cambridge, MA.

Scanlon, T. M. 1975. Preferences and urgency. *Journal of Philosophy* 73, 655–669.

Sen, A. 1970. *Collective Choice and Social Welfare.* Holden-Day, San Francisco.

Sen, A. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7, 243–262.

Sen, A. 1980. Equality of what? In *Tanner Lectures on Human Values I,* ed. S. M. McMurrin. University of Utah Press, Salt Lake City, pp. 195–220.

Sen, A. 1988. Utilitarianism and welfarism. *Journal of Philosophy* 76, 463–489.

Smith, A. 1975. *The Theory of Moral Sentiments.* Clarendon Press, Oxford. (Edited by D. D. Raphael and A. L. Macfie. First published 1759.)

Suppes, P. 1966. Some formal models of grading principles. *Synthèse* 6, 284–306.

Von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ.

Weymark, J. A. 1991. A reconsideration of the Harsanyi-Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being,* ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, pp. 255–320.

Yaari, M. E. 1981. Rawls, Edgeworth, Shapley, Nash: Theories of distributive justice re-examined. *Journal of Economic Theory* 24, 1–39.

14

The Social Contract Naturalized

Brian Skyrms

14.1 Introduction

For John Harsanyi and John Rawls – as well as for Thomas Hobbes before them – the theory of the social contract is an application of the theory of rational decision. For Harsanyi and Rawls, it does not matter whether people are rational or whether there ever was or could have been a state of nature of the kind considered. The importance of the concepts *rationality* and the *state of nature* is not descriptive, but rather lies in the role that they play in a counterfactual definition of justice. A just arrangement is one to which rational decision makers would agree in the state of nature.

Justice is customarily depicted with a blindfold, a scale, and a sword. Both Harsanyi and Rawls structure the state of nature as the blindfold. Justice is rational decision behind the "veil of ignorance." This is a departure from Hobbes. Rawls sees it as a Kantian approach, but justice had her blindfold long before Immanuel Kant formulated his categorical imperative.

There is a different tradition exemplified by David Hume. For Hume, the social contract is a tissue of conventions that have grown up over time. I cannot resist reproducing in full this marvelously insightful passage from his *Treatise:*

> Two men who pull on the oars of a boat do it by an agreement or convention, tho' they have never given promises to each other. Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. On the contrary, this experience assures us still more, that the sense of interest has become common to all our fellows, and gives

This chapter treats issues previously discussed in Skyrms (1994, 1996). I would like to thank the Office of Academic Computing at the University of California, Irvine, for a grant of computer time which made the simulations reported here possible.

> us a confidence of the future regularity of their conduct: And 'tis only on the expectation of this, that our moderation and abstinence are founded. In like manner are languages gradually establish'd by human conventions without any promise. In like manner do gold and silver become the common measures of exchange, and are esteem'd sufficient payment for what is of a hundred times their value. (Hume, 1739, p. 490)

Hume is interested in how we actually got the contract we now have. Modern Humeans at this conference include Robert Sugden and Ken Binmore. Modern Humeans take inspiration as well from Darwinian dynamics. The social contract has evolved and will continue to evolve. Different cultures, with their alternative social conventions, may be instances of different equilibria, each with its own basin of attraction. The proper way to pursue modern Humean social philosophy is via dynamic modeling of cultural evolution. Here, I discuss the light that this approach may throw on questions of distributive justice.

14.2 Distributive Justice, Symmetry

Here we start with a very simple problem; we are to divide a chocolate cake between us. Neither of us has any special claim as against the other. Our positions are entirely symmetric. The cake is a windfall for us, and it is up to us to divide it. But if we cannot agree how to divide it, the cake will spoil and we will get nothing. What we ought to do seems obvious. We should share alike.

One might imagine some preliminary haggling: "How about 2/3 for me, 1/3 for you? No, I'll take 60 percent and you get 40 percent," but in the end, each of us has a bottom line. We focus on the bottom line and simplify even more by considering the bargaining game of John Nash (1950). Each of us writes a final claim to a percentage of the cake on a piece of paper, folds it, and hands it to a referee. If the claims total more than 100 percent, the referee eats the cake. Otherwise, we get what we claim.

What will people do when given this problem? I expect that we would all give the same answer – almost everyone will claim half the cake. In fact, the experiment has been done. Nydegger and Owen (1974) asked subjects to divide a dollar among themselves. There were no surprises. All agreed to a fifty-fifty split. The significance of these results might be questioned because Nydegger and Owen had subjects bargain face to face. However Van Huyk et al. (1995) conducted the experiment, substantially as I have described it, with substantially the same results. Subjects submitted a claim of either 40 percent, 50 percent, or 60 percent of a dollar and these and were randomly paired by the experimenter to determine the outcome. Interaction with anonymous opponents still almost always produced equal splits. The results are just what everyone would have expected. Here it is uncontroversial which rule of division is the norm, and this norm of distributive justice is widely respected in practice. This is the first case to examine.

Suppose that, with Harsanyi and Rawls, we take Justice to be rational choice behind a "veil of ignorance."

> Somehow we must nullify the effects of specific contingencies which put men at odds and tempt them to exploit social and natural circumstances to their own advantage. In order to do this I assume that parties are situated behind a veil of ignorance. (Rawls, 1971, p. 36)

Exactly what the veil is supposed to hide is a surprisingly delicate question, which I will not pursue here. Abstracting from these complexities, imagine you and I are supposed to decide how to divide the cake between individuals A and B, under the condition that a referee will later decide whether you are A and I am B or conversely. We are supposed to make a rational choice under this veil of ignorance.

Who is the referee and how will she choose? I would like to know, in order to make my rational choice. In fact, I don't know how to make a rational choice unless I have some knowledge, or some beliefs, or some degrees of belief about this question. If the referee likes me, I might favor 99 percent for A, 1 percent for B or 99 percent for B, 1 percent for A (I don't care which), on the theory that fate will smile upon me. If the referee hates me, I shall favor equal shares.

It might be natural to say, "Don't worry about such things. They have nothing to do with justice. The referee will flip a fair coin." This is essentially Harsanyi's position. Rawls objects that we do not have access to the objective probabilities behind the veil of ignorance, but this seems to be a confusion. The equal probabilities are part of the veil of ignorance – part of the idealized situation used to define justice.

If all I care about is expected amount of cake – if I am neither risk averse nor a risk seeker – I will judge every combination of portions of cake between A and B that uses up all the cake to be optimal: 99 percent for A and 1 percent for B is just as good as fifty-fifty, as far as I am concerned. The situation is the same for you. The veil of ignorance has not helped with this problem (though it would with others).

Rawls doesn't have the referee flip the coin. Rather, he applies the maximin decision rule: maximize your minimum gain. If we both apply this rule, we will agree on the fifty-fifty split. This gets us the desired conclusion, but on what basis? In the early days of game theory, when the paradigm was a zero-sum game, maximin may have seemed to be a rule of rational decision. If both players play maximin in such games, they are each maximizing expected payoff given the other's play. But the situation here is not zero-sum, and maximin is not the hallmark of rational decision. Harsanyi (1975) criticizes Rawls on this point, and Rawls's (1974) reasons for using maximin behind the veil of ignorance must strike decision theorists as inadequate.

Binmore (1993, 1998, 2008) offers a naturalistic account of moral behavior, according to which a systematic use of the veil of ignorance has evolved as a coordination device. According to this account, players maximize expected utility behind the veil as in Harsanyi. However, they may also at any time choose to renegotiate behind the veil. Such an institutionalization of the use of the veil of ignorance allows Binmore to give a new rationale for maximin as a procedure for producing equilibria which do not carry an incentive for renegotiation.

Here, however, I do not want to presuppose the existence of such an institution. Let us ask what can happen when we start before the existence of such moral institutions. Throw away the veil. Throw away the blindfold. People just bargain, over and over, in divide-the-cake, and strategies evolve over time. We suppose that successful strategies are imitated more often than unsuccessful ones, so that this process of cultural evolution follows a dynamics qualitatively similar to the replicator dynamics of evolutionary game theory.

Let $U(S)$ be the average payoff to individuals using strategy S and UBAR be the average payoff to the population. Then according to the (discrete) replicator dynamics, the proportion of the population using S in the next generation is gotten by multiplying the proportion of the population using S in the current generation by the current value of the ratio U/UBAR. This is the dynamics to be used here. More information on replicator dynamics can be found in Hofbauer and Sigmund (1988).

14.3 Evolution of Justice I

What can we say about the evolution of strategies in the simplest bargaining game? The first question to ask is what strategies are *evolutionarily stable* in the sense of Maynard Smith and Price (1973). The intuitive idea of an evolutionary stable strategy is that of a strategy such that a population of individuals playing that strategy cannot be invaded by a small number of mutants playing an alternative strategy. In a context of two-person games with random encounters, Maynard Smith and Price define an evolutionary stable strategy, S, as one such that for any potential mutant strategy, S': *either the native S played against itself gets a better payoff than the mutant, S', played against it, or both S and S' do equally well against S but when played against the mutant S', the native S gets a greater payoff than the mutant.* A state where all the population shares an evolutionary stable strategy is attracting, or strongly stable, in the replicator dynamics.

Sugden (1986) asks this question and notes that the unique evolutionarily stable strategy in the game is equal division. First, we note that a population composed of individuals who demand 50 percent cannot be invaded. A mutant who demanded more would get nothing. A mutant who demanded less would get less. Then we note that any other strategy can be invaded – and indeed can be invaded by this strategy. In a population that demands less, demand 50 percent does better against the natives than they do against themselves. In a population that demands more, neither gets anything playing against the natives, but demand 50 percent does better than the natives when playing against itself. Evolution appears to give equal division that unique status that rational choice behind the veil of ignorance promises but fails to deliver.

The matter, however, is not so simple. Evolution might not lead to a state where every member of the population uses the same strategy, but perhaps to a polymorphic state of the population, a limit cycle, a strange attractor, or something else. The next question is whether there are evolutionarily stable polymorphic states of the population in the divide the cake game. As soon as one asks the question, it is evident that there are an infinite number of such states.

For example, suppose that half the population claims 2/3 of the cake and half the population claims 1/3. Let us call the first strategy *Greedy* and the second *Modest.* A greedy individual stands an equal chance of meeting another greedy individual or a modest individual. If she meets another greedy individual she gets nothing because their claims exceed the whole cake, but if she meets a modest individual, she gets 2/3. Her average payoff is 1/3. A modest individual, however, gets a payoff of 1/3 no matter who she meets.

This polymorphism is a stable equilibrium. If the proportion of Greedys should rise, then Greedys would meet each other more often and the average payoff to Greedy would fall below the 1/3 guaranteed to Modest. And if the proportion of Greedys should fall, the Greedys would meet Modests more often, and the average payoff to Greedy would rise above 1/3. Negative feedback will keep the population proportions of Greedy and Modest at equality. But what about the invasion of other mutant strategies? Suppose that a mutant who demands more than 2/3 arises in this population. This mutant gets payoff of zero and goes extinct. Suppose that a mutant who demands less than 1/3 arises in the population. This mutant will get what she asks for, which is less than Greedy and Modest get, so she will also go extinct, although more slowly than the former one will. The remaining possibility is that a middle-of-the-road mutant arises who asks for more than Modest but less than Greedy. A case of special interest is that of the *fair-minded* mutant who asks for exactly 1/2. All of these mutants would get nothing when they meet Greedy and get less than Greedy does when they meet Modest. Thus they will all have an average payoff less than 1/3 and all, including our fair-minded mutant, will be driven to extinction. The polymorphism has strong stability properties.

This is unhappy news for the population as well as for the evolution of justice because our polymorphism is inefficient. Here everyone gets, on average, 1/3 of the cake, while 1/3 of the cake is squandered in greedy encounters. Compare this equilibrium with the pure equilibrium where everyone demands and gets 1/2 of the cake. In view of both the inefficiency and the strong stability properties of the 1/3–2/3 polymorphism, it appears to be a kind of trap that the population could fall into, and from which it could be difficult to escape. There are lots of such polymorphic traps. For any number, x, between 0 and 1, there is a polymorphism of the two strategies Demand x, Demand $1 - x$, which is a stable equilibrium in the same sense and by essentially the same reasoning as in our example.

Given the infinite number of evolutionarily stable polymorphic states of the population, to go further we need some estimate of the relative size of their basins of attraction under an appropriate dynamics. I pursued some estimates by means of computer simulations. The program picks a vector of population proportions at random from the simplex of population proportions (according to the uniform distribution), lets the system evolve for 1,000 generations, sees if any strategy has taken over 99 percent of the population, tallies the result, and repeats this process many times.

Supposing that the game is to divide 10 dollars, and that permissible claims are whole dollar amounts from $1 through $9, I got the following results:

Total trials:	10,000
Fair division	6,198
4, 6 polymorphism	2,710
3, 7 polymorphism	919
2, 8 polymorphism	163
1, 9 polymorphism	10

Fair division has the largest basin of attraction, with about 62 percent of the initial points evolving to fair division, with the basins of attraction of polymorphic traps being larger, the closer they are to fair division.

Because the continuum of possible choices in the ideal bargaining games must be reduced to a finite number for computation to be feasible, the question arises as to whether the granularity of the discretization makes a difference. It does. Here are the result for simulations involving dividing $20 and dividing $200:

The results for dividing $20 were

Trials	10,000
Fair division	5,720
9, 11 polymorphism	2,496
8, 12 polymorphism	1,081
7, 13 polymorphism	477
6, 14 polymorphism	179
5, 15 polymorphism	38
4, 16 polymorphism	8
3, 17 polymorphism	1

The results for dividing $200 were

Trials	1,000
Fair division	622
99, 101 polymorphism	197
98, 102 polymorphism	88
97, 103 polymorphism	34
96, 104 polymorphism	19
95, 105 polymorphism	14
94, 106 polymorphism	9
93, 107 polymorphism	7
92, 108 polymorphism	5
91, 109 polymorphism	1
90, 110 polymorphism	2
89, 111 polymorphism	2

Only 1,000 trials were done for dividing $200 because of the amount of computing resources needed for this number of strategies. The results for $20 and $200 are compared in Figure 14.1, normalizing trials to 1,000 and size of cake to $20. It is evident that as the granularity of the problem becomes finer, more and more of the initial points go near to an even split.

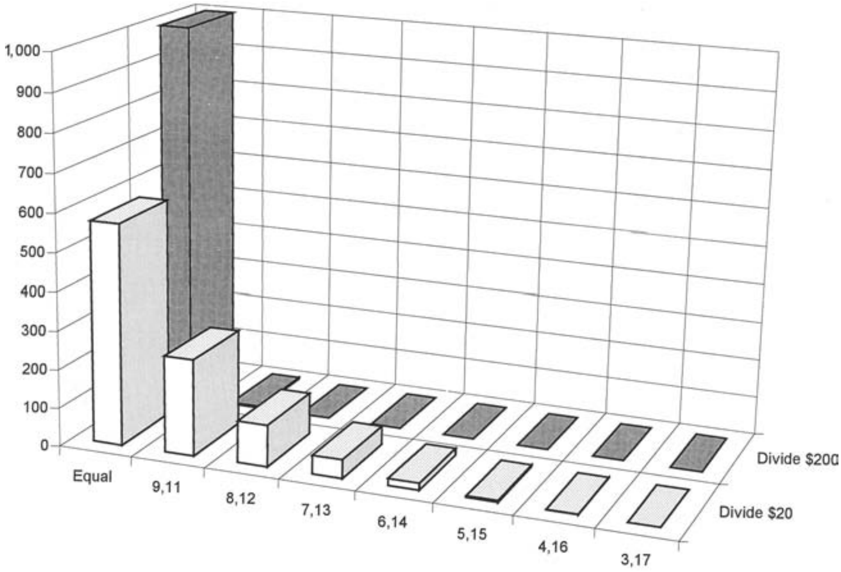


Figure 14.1. Effects of granularity: Symmetric case.

The assumption of a continuously divisible good in problems of distributive justice is an idealization. Social norms will evolve in a setting involving a mix of problems of different granularities. Here the simple model of differential reproduction supplied by the replicator dynamics provides a beginning of an explanation for the evolution of the rule of share and share alike.

We could pursue the matter by adding some detail to our model of the evolutionary process. Pairing between members of the population might not be random. Instead, there might be some population viscosity as in Hamilton (1964). One could make the dynamics explicitly probabilistic to reflect random variation and finite population size, as in Foster and Young (1990) or Kandori, Mailath, and Rob (1993). One could introduce some variability in, and uncertainty about, the size of the cake to be divided, as suggested already by Nash (1953). These refinements of the model in the direction of greater realism only make the case more compelling for the evolution of the norm of equal division in symmetric bargaining situations. See Skyrms (1996) with regard to correlation, Young (1993) with regard to probabilistic evolution, and Binmore (1987) and Van Damme (1987, sec. 7.6) with regard to the variable cake. The last three references give proofs of limiting results. Rather than entering into these matters here, however, I would like to move on to the case of asymmetric bargaining.

14.4 Distributive Justice, Asymmetry

Suppose that two players are to divide a cake, as before, but that now they may derive different benefits from the same amount of cake. Harsanyi and Rawls hold that the way in which benefits, or utility, depend on amount of cake is determined objectively by the kind of person involved or the situation he or she is in. Interpersonal comparisons of benefit are objective. Thus, for Harsanyi and Rawls the just distribution is the one determined by rational choice of a division between A, who has one utility function and B, who has another, where the choice is made with the knowledge of the utility functions of A and B, but in ignorance of who will be A and who will be B. For Rawls, the maximin rule of rational choice yields the *difference principle*. The division should maximize the utility of the worst off. For Harsanyi, equiprobability of being A or B is part of the veil of ignorance, and the rule of maximizing expected utility yields the *utilitarian solution*, which maximizes the sum of the utilities of A and B.

In addition, we might consider two solutions to the bargaining problem which have a different origin, the solutions of Nash (1950) and Kalai and Smorodinsky (1975). Each can be shown to be the unique solution satisfying a set of plausible axioms. Both of the axiom sets contain an axiom that runs counter to the interpersonal comparison of utility or benefit assumed by both Harsanyi and Rawls. That is the axiom that the solution must be invariant with respect to affine transformations of the utility functions of each player. The motivation is the fact that the utility function for a player is not operationally determined by the usual representation theorem, except up to such a transformation. Obviously, neither the difference principle nor the utilitarian solution satisfy this axiom.

Assume that we choose utility scales for A and B which set the utility of the worst outcome to zero. Then Nash's solution is the division of the cake that maximizes the product of the utilities of A and B. The Kalai–Smorodinsky solution looks at the ratio of the utility of A in the outcome best for A and the utility for B in the outcome best for B, and picks the Pareto-efficient division which yields that ratio. This solution has been advocated by Gauthier (1986).

In the case in which utility is a linear function of the amount of cake for each player, each of these solutions coincides with share and share alike. Let us consider two examples that bring out the differences.

Example 1: Player A's utility = amount of cake; player B's utility = 10 × amount of cake. Here Nash and Kalai–Smorodinsky fix on equal division. To them, the problem is just like the symmetric case. The utilitarian solution gives all the cake to B. The difference principle gives most of the cake to A.

Example 2: Player A's utility = amount of cake; player B's utility = amount of cake up to half the cake but remains at that value for larger amounts of cake. (We can think of player B becoming satiated at half the cake, while player A does not.) Nash here still gives one half of the cake to each. This solution also respects the difference principle. Kalai–Smorodinsky gives 2/3 of the cake to A and 1/3 to B. Both solutions are utilitarian.

The Nash and Kalai–Smorodinsky bargaining solutions are elegant answers, but what is the question? Raiffa (1953) discussed the matter in terms of fair arbitration schemes. But Nash had a different view. He regarded a bargaining solution as a predicted outcome, corresponding to an equilibrium in a corresponding noncooperative game. We want to look at the matter from still another perspective: How does evolutionary dynamics discriminate between these bargaining solutions?

14.5 Evolution of Justice II

The evolution of norms of distributive justice in asymmetric cases is a complicated process and applying the replicator dynamics to repeated stylized games is only a first step in investigating that evolutionary process. Still, the first step is worth taking. It is not obvious what the results will be. The considerations that Harsanyi and Rawls bring forth do not speak to the evolutionary process. The axiom of scale invariance used in the justification of both Nash and Kalai–Smorodinsky loses its justification when payoffs are in terms of Darwinian fitness rather than von Neumann–Morgenstern utility.

There are two fundamentally different ways in which we could model the evolutionary dynamics of an asymmetric bargaining game. We could use two populations, where one population has the fitness function of A and the other has the fitness function of B. Or we could use one population, where an individual sometimes finds herself in the role of A, with A's fitness function, and sometime finds herself in the role of B, with B's fitness function. I choose the second alternative here, although one can imagine situations in which the first would also be of interest.

Immediately, we again have a multitude of evolutionarily stable pure strategies. Consider any pure strategy which demands x cake in role A and $(1 - x)$ cake in role B, where $0 < x < 1$. In a population composed of individuals using this strategy, any mutant will do worse against the natives than the natives do against themselves. So we need to compare basins of attraction of all these pure evolutionarily stable strategies.

Suppose that the cake is divided into eighteen equal pieces, and demands must be in terms of whole pieces of cake. Again, initial population proportions were picked at random and the system evolved according to the replicator dynamics. I report the proportion of initial conditions that led to fixation of a strategy in each of our two example games of the last section. I write the strategies in the form "<Demand in role A, Demand in role B>."

Example 3:

Harsanyi	<0, 18>	0.0%
	<1, 17>	0.0%
	<2, 16>	0.0%
	<3, 15>	0.0%
	<4, 14>	0.0%
	<5, 13>	0.0%
	<6, 12>	0.9%
	<7, 11>	12.5%
	<8, 10>	32.4%
Nash, K-S	<9, 9>	38.6%
	<10, 8>	14.6%
	<11, 7>	0.9%
	<12, 6>	0.0%
	<13, 5>	0.0%
	<14, 4>	0.0%
	<15, 3>	0.0%
Rawls	<16, 2>	0.0%
	<17, 1>	0.0%
	<18, 0>	0.0%

Example 4: Here I report the result of 10,000 trials:

	<6, 12>	0
	<7, 11>	0
	<8, 10>	1
Nash, Rawls	<9, 9>	6,164
	<10, 8>	3,374
	<11, 7>	316
K-S	<12, 6>	2
	<13, 5>	0
	<14, 4>	0

The remaining 143 did not converge in the allotted time.

In these simulations the Nash bargaining solution is surprisingly robust. In each case, if one starts from a population in which all strategies are equally represented, the system evolves to the Nash solution. Starting from randomly selected population proportions, the Nash solution displays the largest basin of attraction, and strategies close to Nash attract almost all initial points. But in Example 3, where the utilitarian solution disagrees with Nash, the distribution around the Nash solution is skewed in the direction of the utilitarian solution. And in Example 4, where the Kalai–Smorodinsky solution disagrees with the Nash solution, the distribution around Nash is skewed in the direction of Kalai–Smorodinsky.

It is evident that the evolutionary dynamics of these discrete bargaining games is too rich to be captured by any axiomatic bargaining theory. But if you look for one simple theme to take away from the simulations, it must be the extent to which the dynamics respects the Nash solution.

In the case where payoff = cake, we found that granularity of the discrete bargaining problem made a difference. In the one population model of the asymmetric case, computational investigation of fine-grained discrete bargaining problems is very computationally intensive because of the way the two roles affect the size of the strategy space. However, a small simulation may give some indication of the effect of granularity, even if it cannot be viewed as generating reliable statistics. For Example 4, I ran 1,000 trials with thirty-six indivisible pieces of cake and 1,000 trials with seventy-two pieces of cake. I got the following results:

Nash	<18, 18>	471	<36, 36>	253
			<37, 35>	471
	<19, 17>	410	<38, 34>	216
			<39, 33>	23
	<20, 16>	100		
	<21, 15>	3		

The other trials did not converge. The results are graphed in Figure 14.2. In finer-grained problems, the distribution tends to peak more sharply around the Nash solution.

In the symmetric case, we noted briefly that moving to a probabilistic model only made the case for equal division more compelling. A little bit of random variation in the system will eventually knock the system out of a polymorphic trap, with the result that, in the long run, in an arbitrarily fine-grained symmetric bargaining game, the system will spend most of its time arbitrarily close to the egalitarian solution. In the asymmetric case

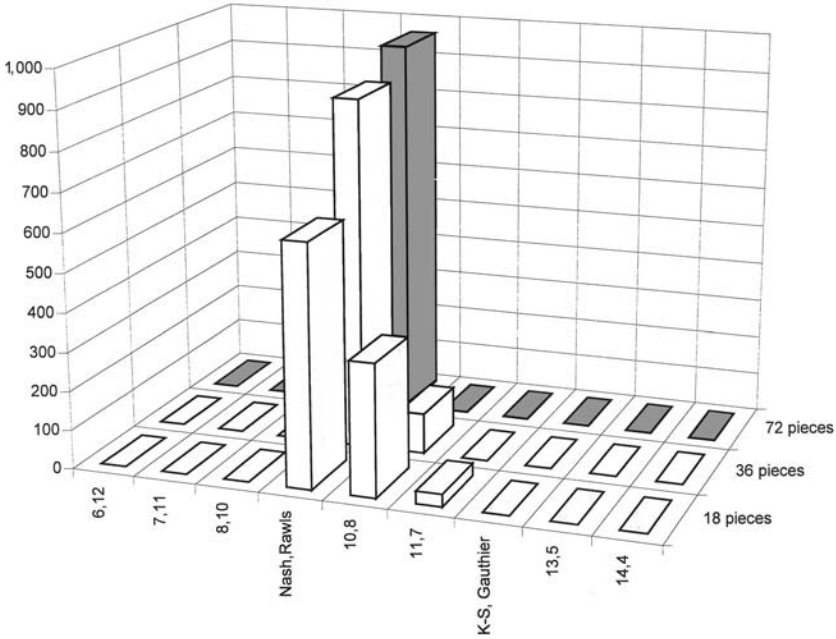


Figure 14.2. Effect of granularity: Asymmetric case.

under consideration here, the Nash bargaining solution enjoys the same distinction. All this and more is established analytically in Young (1993). The simulations reported here serve to show how much robustness those long-run, limiting results have in finite-run discrete cases.

In the symmetric case, we also noted that positive correlation in pairing from the population had the effect of destabilizing polymorphisms and making the evolution of equal division more likely. In the asymmetric case, positive correlation may have a more interesting effect. Consider the extreme case, where each individual interacts with another individual with the same strategy. By hypothesis, the Darwinian fitness of each strategy depends on the sum of the payoffs in the two roles. Accordingly, the utilitarian strategies have the highest fitness. In situations like that of Example 3, where there is a unique utilitarian solution, perfect correlation will cause it to go to fixation. In general, some positive correlation shifts the distribution away from the Nash solution and toward the utilitarian solution in cases where they disagree.

In the symmetric case, we mentioned that a variable cake favors equal division. Nash (1953) suggested how uncertainty about the size of the cake could support the Nash solution in general. Rigorous limiting results supporting the Nash bargaining solution in the asymmetric case as well are given in Binmore (1987) and Van Damme (1987).

14.6 Conclusion

We applied a simple model of differential reproduction, the *replicator dynamics,* to a simple model of the problem of distributive justice, *divide the cake.* In the simplest case, in which payoff in Darwinian fitness equals amount of cake for everyone involved, we found that the rule of equal division was the only evolutionary stable strategy. The importance of this unique status, however, must be qualified in the light of various evolutionarily stable polymorphic states. After considerations of granularity, random variation, and correlation were introduced, we were left with a plausible case for the evolution of something close to the egalitarian norm.

Where the payoff functions are different for players in different roles, the partial results I reported are more complex. In the discrete bargaining game of Example 3, where there is sharp disagreement between the utilitarian solution, the Nash solution and the solution recommended by Rawls's difference principle, the probability distribution of final states (starting from uniform probability on initial states), has its mode at the Nash solution, but is skewed toward the utilitarian solution. In the game corresponding to Example 4, the mode is again at the Nash solution but skewed towards the Kalai–Smorodinsky solution.

Granularity plays a role here too, with the distribution peaking more sharply at the Nash solution in finer-grained bargaining games. Positive correlation in the pairing of individuals from the population to play the game favors the utilitarian solution.

Perhaps it is not going too far to suggest that philosophers – who have devoted considerable effort to discussion and analysis of the difference principle, the utilitarian solution, and even the Kalai–Smorodinsky solution – might turn their attention also to the Nash solution. It is evident, however, that when there are different roles that persist through time as norms evolve, the dynamical behavior is too rich to support unequivocally one bargaining solution, even in the simple models considered here. Many aspects of the dynamics of the evolution of distributive justice remain to be explored.

The social contract is a complex web of conventions and norms. Norms of distributive justice are only one part of the contract. One can also ask about norms of performance of covenants, of property, and of punishment of those who do not obey first-order norms. Evolutionary analysis has already put some of these questions in a new light. For example, Binmore, Gale, and Samuelson (1995) show how a subgame perfect strategy that punishes greedy offers in the ultimatum game can persist in a population. Hamilton (1964) showed how population viscosity and other forms of correlated interaction can explain the evolution of altruism. The naturalistic investigation of the social contract has valuable things to teach social philosophy, if social philosophers are willing to investigate a new point of view.

Postscript July 2006

Naturalists describe man as he is, not as they would wish him to be. Naturalists use dynamic models of cultural evolution to explain how men may have come to be the way they are. Any credible naturalistic treatment must be compatible with a variety of outcomes because that is what we observe in the world.

Sometimes Justice lacks a blindfold, as in the depiction of royal justice that is chosen as the frontispiece of my *Evolution of the Social Contract.* She has the scales and the sword, but she knows with whom she is dealing. I believe that this is often a more accurate depiction of the real world than the conventional idealization. If so, naturalists might well eschew the veil of ignorance except in circumstances where they can show that it is operative.

That is not to criticize Harsanyi or Rawls. They are engaged in a quite different enterprise – that of developing the consequences of a certain idealized conception of justice. Evolutionary models may connect with this enterprise by showing how distributive justice according to Harsanyi or according to Rawls may evolve but also by showing how such cultural patterns may fail to evolve.

The least controversial case of bargaining is that where all bargaining solutions can agree on the equal split – that is, the case in which the situation of the bargainers is completely symmetrical. The existence of a multiplicity of polymorphic evolutionarily stable equilibria (first seen by Sugden) raises questions for evolution of the equal split in even this most favorable case, and I believe that these questions need to be taken seriously.

Evolution of the equal split is promoted by positive correlation of encounters, either when simply added to a replicator dynamics model, or when endogenously generated by local interaction, signaling, or dynamic interaction networks. But positive correlation also alters the evolution of bargaining behavior in situations where the positions of the agents involved are not symmetric, as pointed out in the foregoing chapter and elsewhere.

Although correlation mechanisms have been widely discussed in connection with the evolution of altruism, their effect on evolution of bargaining largely remains to be explored.

We should not expect some monolithic result, saying that one bargaining solution must always evolve but rather a growing analysis that lays out a variety of possibilities.

References

Binmore, K. 1987. Nash bargaining theory II. In *The Economics of Bargaining,* ed. K. Binmore and P. Dasgupta. Blackwell, Oxford, pp. 61–76.

Binmore, K. 1994. *Game Theory and the Social Contract,* Vol. 1: *Playing Fair.* MIT Press, Cambridge, MA.

Binmore, K. 1998. *Game Theory and the Social Contract,* Vol. 2: *Just Playing.* MIT Press, Cambridge, MA.

Binmore, K. 2008. Naturalizing Harsanyi and Rawls. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 303–333.

Binmore, K., Gale, J., and Samuelson, L. 1995. Learning to be imperfect: The ultimatum game. *Games and Economic Behavior* 8, 56–90.

Foster, D., and Young, P. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38, 219–232.

Gauthier, D. 1986. *Morals by Agreement.* Clarendon Press, Oxford.

Hamilton, W. D. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7, 1–52.

Harsanyi, J. 1953. Cardinal utility in welfare economics and the theory of risk-taking. *Journal of Political Economy* 61, 434–435.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321.

Harsanyi, J. 1975. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review* 69, 594–606.

Harsanyi, J. 1976. *Essays in Ethics, Social Behavior, and Scientific Explanation.* D. Reidel, Dordrecht.

Hobbes, T. 1651. *Leviathan.* Andrew Crooke, London.

Hofbauer, J, and Sigmund, K. 1988. *The Theory of Evolution and Dynamical Systems.* Cambridge University Press, Cambridge.

Hume, D. 1739. *A Treatise of Human Nature.* John Noon, London; quoted from the edition of L. Selby-Bigge, Clarendon Press, Oxford, 1888.

Kalai, E., and Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43, 513–518.

Kandori, M., Mailath, G., and Rob, R. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61, 29–56.

Maynard Smith, J., and Price, G. R. 1973. The logic of animal conflict. *Nature* 246, 15–18.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–162.

Nash, J. 1953. Two-person cooperative games. *Econometrica* 21, 128–140.

Nydegger, R. V., and Owen, G. 1974. Two-person bargaining, an experimental test of the Nash axioms. *International Journal of Game Theory* 3, 239–250.

Raiffa, H. 1953. Arbitration schemes for generalized two-person games. In *Contributions to the Theory of Games,* Vol. 2, ed. H. Kuhn and A. W. Tucker. Annals of Mathematics Studies, No. 28. Princeton University Press, Princeton, N J, pp. 361–387.

Rawls, J. 1957. Justice as fairness. *Philosophical Review* 67, 164–194.

Rawls, J. 1971. *A Theory of Justice.* Harvard University Press, Cambridge, MA.

Rawls, J. 1974. Some reasons for the maximin criterion. *American Economic Review, Papers and Proceedings* 64, 141–146.

Skyrms, B. 1994. Sex and justice. *Journal of Philosophy* 91, 305–320.

Skyrms, B. 1996. *Evolution of the Social Contract.* Cambridge University Press, Cambridge.

Sugden, R. 1986. *The Economics of Rights, Cooperation and Welfare.* Blackwell, Oxford.

Van Damme, E. 1987. Stability and Perfection of Nash Equilibria. Springer, Berlin.

Van Huyck, J., Battalio, R., Mathur, S., and Van Huyck, P. 1995. On the origin of convention: Evidence from symmetric bargaining games. *International Journal of Game Theory* 24, 187–212.

Young, H. P. 1993. An evolutionary model of bargaining. *Journal of Economic Theory* 59, 145–168.

15

An Alternative Model of Rational Cooperation

Edward F. McClennen

15.1 Introduction

I want to extend here a line of reasoning that I pursued in *Rationality and Dynamic Choice* (1990). In that book I argued that the standard Bayesian model of expected-utility reasoning needs to be revised to accommodate a capacity, on the part of rational decision makers, to effectively coordinate with their own future selves – to be guided by plans that they have deliberately adopted. I also suggested that an analogous line of reasoning might be employed to show that rational agents could engage in a coordination of their choices with others to a greater extent than the standard theory would seem to admit and, in particular, that they could achieve coordination by following mutually accepted rules. It is this suggestion that I now want to explore more fully. It is not that the standard theory altogether denies the possibility of such coordination. Rather, on its view, rational agents will be disposed to free-ride on the cooperative efforts of others and thus effective cooperation will require the adoption of a system of surveillance and sanctions. In addition, it views the terms of the agreements that rational agents reach as driven by essentially noncooperative considerations, such as the relative bargaining power of the participants. But enforcement schemes require the expenditure of scarce resources, and bargaining based on the principle of to each according to threat advantage tends to generate destabilizing and mutually disadvantageous conflict. The two problems, moreover, appear to be connected in an important way. The sense that one's relationship to others is defined by relative threat advantage is likely to contribute to one's disposition to free-ride whenever one can. Thus, on a number of counts, the standard model of cooperation seems to imply that rational agents will do less well than they might ideally do. It is worth considering, then, whether there is not a more satisfactory account to be offered of the nature and conditions of rational cooperation.

The matters to be addressed here relate in a number of different ways to the work of both Rawls and Harsanyi. Both Rawls (1955) and Harsanyi (1977b) have given an account of how to think about rule-guided behavior, particularly within the context of moral theory. Although I share their conviction that the concept of a rule plays a pivotal role there, I want to follow up on something implicit in Rawls's account of rules and make the case for rule-governed choice at a deeper level yet, as part of a more general theory of practical reason.[1] The stress I shall place on rational, prudential, or practical choice also connects with long-standing research programs of both Rawls (1971) and Harsanyi (1953, 1955) – albeit one from which Rawls appears to have retreated in his later years (see Rawls, 1993). This is the idea of trying to justify various social or political principles by showing that they would be agreed on by rational persons who have to choose under conditions of radical uncertainty. Harsanyi (1955, 1978) has also argued that a basic utilitarian principle can be validated by axiomatic social choice theory. The approach I shall take here is very different. I shall not rely on the deep uncertainty to which Rawls appealed in his construction; nor shall I make appeal to the axioms of social choice theory. What emerges, however, is a principle governing political and social structures that is very similar to Rawls' difference principle.[2]

15.2 Mutual Gains and Losses

Starting with Adam Smith, and running like a bright thread throughout virtually all the subsequent theoretical literature on political economy, one can mark a preoccupation with the conditions under which individuals can transact with one another to their mutual advantage. In the more formal literature, this concern culminates, in the middle of the twentieth century, in a fundamental theorem of welfare economics, according to which individuals can, under conditions of perfect competition, achieve

[1] In Harsanyi (1977b) a consequentialist case is made for why the *moral* decisions of individuals should be guided by rules, by appeal to what he characterizes as considerations of "coordination effects" and "expectation and incentive effects." The arguments presented there, however, seem perfectly adaptable to models of interaction governed by the merely personal (as distinct from the moral) preferences of each participant. That is, by parity of reasoning, coordination effects and expectation and incentive effects seem fully relevant to interest-driven forms of interaction as well.

[2] At this more substantive level, then, I find myself in agreement with Rawls and opposed to the utilitarian view that Harsanyi has put forward.

an outcome that is Pareto optimal and strictly Pareto efficient relative to the outcome of no transactions (that is, each does better as a result of such transactions).[3]

The theme of Pareto-efficient changes in institutional structures is also central to Coase's important work, both on the theory of the firm (1937) and the problem of social cost (1960), to Posner's economic analysis of law (1986), the public choice tradition initiated by Buchanan and Tullock (1962), Axelrod's work on iterated prisoner's dilemma games (1984), and Ullman-Margalit's study of the emergence of norms (1978). The Pareto conditions also figure centrally in virtually all axiomatic bargaining and social choice models.

To be sure, these explorations have often been accompanied by the suggestion that arrangements satisfying the Pareto conditions are perhaps best secured indirectly rather than by any deliberate attempt by participating agents. This is, of course, the idea that is so strikingly captured by Smith's "invisible hand" metaphor: The wealth of nations is an unintended by-product of each person's pursuing his or her own personal interests. It also finds powerful reincarnation in some evolutionary accounts of the emergence of institutional structures (Alchian, 1950; Axelrod and Dion, 1988; Hayek, 1967; Nelson, 1994; Sugden, 1989).

Finally, where satisfaction of the Pareto conditions is not assured, much effort is invested in the search for mechanisms that will overcome this problem. The recent "folk" theorems on indefinitely repeated games are a case in point.[4] The objective is to show that Pareto-efficient outcomes can be secured by backing up coordination schemes with appropriate sets of surveillance and enforcement devices.

The good news about the emergence of structures that are Pareto-efficient has to be tempered, however, by the bad news that historical processes of institutional development tend to be path dependent in ways that work against adaptive efficiency and that virtually all institutional arrangements are subject to manipulation by special interests, to the short-term advantage of some but often to the long-term disadvantage of all. The former theme is central to the work of Arthur (1994) and also to the new institutional theory associated with North (1990). The latter theme is central to the literature on "rent-seeking" (see Buchanan, Tollison, and Tullock, 1981), to North (1990), Olson (1982), Knight (1992), and Hardin (1995).

[3] See, for example, Arrow (1969).

[4] For a survey of this work, see Fudenberg and Tirole (1992, ch. 5), and the many citations therein.

Recent work in political economy has in fact identified deep pressures on human interaction that tend to prevent the satisfaction of the Pareto conditions. These conditions, then, seem to characterize merely an ideal that is remote from social reality. In one sense, there should be nothing surprising about this. After all, the historical record hardly supports any other picture. What is surprising, however, is that many of these conclusions are based on models of *ideally* rational beings who have substantial and common knowledge of each other's rationality, preferences, and the strategic structure of their interactions. On the accounts offered, Pareto suboptimality does not flow just from assuming that some are less than fully rational or possess limited or asymmetrical knowledge. Mutually disadvantageous free-riding and conflict over the distribution of goods are taken to be natural to the way in which even hyperrational and fully informed individuals interact. This poses the question that I want to address in this chapter. How does it happen that even thoroughly rational and knowledgeable agents manage to do so poorly?

15.3 Modeling Cooperative Interaction

My ultimate concern will be with dynamic models of ongoing interaction that involve a mixture of conflict and cooperation and which provide us with a reasonable approximation to real-life situations. It will help to begin, however, by considering how the theory of rational choice has been articulated for simple one-stage cooperative games, and, in particular, for a special class of such games, namely, *pure-coordination* games. These are games in which (1) outcomes are ranked identically by all participants, so that there is a complete convergence of interests, (2) the parties are not able to freely communicate with one another, and (3) there is more than one combination of choices that yield outcomes that are both Pareto optimal and Pareto efficient (relative to what uncoordinated choice can achieve).

As Schelling (1960) and Lewis (1969) clearly recognize in their original work on the nature of coordination games, it is the combination of (2) and (3) that generates a coordination problem. If there is common knowledge of the structure of the game and each participant's preferences with regard to outcomes, and there is a unique outcome that satisfies the Pareto conditions, then there is an obvious choice for each: do one's part to ensure the realization of that outcome.[5] Correspondingly, if the parties can communicate with one another, then presumably they can agree on what combination of choices

[5] Schelling (1960, Appendix C) expresses reservations about whether this is sufficient for one to regard the situation as trivially resolvable.

is to be played and proceed to execute that agreement.[6] This suggests, then, that for persons who deliberately attempt to coordinate their choices with one another, Pareto considerations will play a central role.

As it turns out, however, both Schelling and Lewis start with a much more general framework, which is then particularized in a manner that ends up marginalizing the Pareto conditions. They begin by modeling the way in which a player would deliberate, given a belief that there is common knowledge of the rationality, preferences, and choices of each participating agent, and perhaps also, at least in some cases, certain psychological "cues." Both make the assumption that such a situation calls for strategic deliberation – that the "best" choice for a player depends on what that player expects the other player to do, and where the choice of the other is guided by similar considerations. Successful coordination, on this account, will involve a convergence of choice and expectation: What each expects the other to choose is what the other does choose. However, this theme of convergence of choice and expectation is initially laid out by each in abstraction from any specific criterion of rational choice. Schelling speaks simply of "wise" choice and Lewis of choosing "appropriately."[7]

[6] Notice that within the context of games of pure-coordination, every Pareto-optimal outcome is also Pareto efficient. The plot of all outcome points in the graphic representation of a two-person game of pure coordination, with payoffs for the two players measured along the horizontal and the vertical axes, is simply a straight line with positive slope, that is, a line moving from inferior outcomes in the southwest, to more superior outcomes in the northeast. Whatever is taken as the benchmark point that represents what each can achieve without coordination, every Pareto-optimal outcome is also Pareto efficient relative to that point.

[7] Here is how Schelling (1960, p. 86) begins his formal characterization of a game of pure coordination:

> The pure-coordination game is a game of strategy in the strict technical sense. It is a behavior situation in which each player's best *choice of action* depends on the action he *expects* the other to take, which he knows depends, in turn, on *the other's expectations of his own*. This interdependence of expectations is precisely what distinguishes a game of strategy from a game of chance or a game of skill. In the pure-coordination game the interests are convergent; in the pure conflict game the interests are divergent; but in neither case can a choice of action be made wisely without regard to the dependence of the outcome on the mutual expectations of the players.

The parallel passage in Lewis (1969) is the following:

> We may achieve coordination by acting on our concordant expectations about each other's actions. And we may acquire those expectations, or correct or corroborate whatever expectations we already have, by putting ourselves in the other fellow's shoes, to the best of our ability. If I know what you believe about the matters of fact that determine the likely effects of your alternative actions, and if I know your preferences among possible outcomes *and I know that you possess a modicum of practical rationality*, then I can replicate your practical reasoning to figure out what you will probably do, so that I can act *appropriately*. (p. 27, emphasis added)

Immediately after, however, and without any comment on this move, both proceed to give content to the notion of a rational choice by appealing to a criterion that is central to almost all work in decision and game theory: the maximization of (expected) utility. More specifically, both simply assume that each will choose a utility-maximizing response to what each expects the other player to choose. It follows from this that in so far as they succeed in coordinating their choices, each choice will be an expected-utility-maximizing response to the other; that is, their choices will be in equilibrium.[8] For Schelling and Lewis, then, and for the many who have followed them, the equilibrium condition is essentially taken as a "given" and because characteristically there will be more than one equilibrium combination of strategies, what is sought is a model of the reasoning and deliberative processes of the players that explains how, despite their choosing independently and simultaneously, they may still converge on some particular equilibrium pair.[9]

Of course, preoccupation with equilibrium pairs does not have to mean abandonment of concern for outcomes satisfying the Pareto conditions. The most serious form of conflict between the Pareto requirements and the equilibrium requirement cannot arise in the case of games of pure coordination. In such games, all combinations of strategies satisfying the Pareto conditions are also equilibrium combinations.[10] The converse, however, is

[8] The first explorations of the implications of equilibrium reasoning addressed the special case of perfectly competitive games. But Nash (1950) explicitly extended the argument to nonstrictly competitive games, and shortly thereafter Schelling (1960) and Lewis (1969) extended it to the theory of games of *pure coordination*. This was followed by an extension of the equilibrium condition to dynamic games, and most of the subsequent work in dynamic game theory has been preoccupied with variations on, or refinements of, the notion of equilibrium. This is especially true of the literature on indefinitely iterated games, and the "folk theorems" that were developed in that context. See Fudenberg and Tirole (1992, ch. 5).

[9] It needs to be emphasized that I am here concerned with simple one-shot, simultaneous choice games in which we have abstracted from any observations concerning how these or similar games have been played in the past, and thus also from the kinds of learning experiences that would characterize iterated play. Such convergence as takes place, then, must be understood as the result of the capacity of each to imaginatively replicate the reasoning of the other.

[10] To see this, suppose that the combination is optimal but not in equilibrium. Then there is some participant who could, by choosing differently, increase the utility of the outcome on the assumption that everyone else continues to play for the original combination. Since, however, the parties agree on the utilities of the outcomes, this must mean that *everyone* would prefer to coordinate on this combination rather than the original one. Thus, contrary to the hypothesis, the original combination did not yield an optimal outcome. The converse, of course, is not necessarily true: A strategy pair can be in equilibrium, but its outcome

not true. Given the commitment to equilibrium requirements, then, what has predictably emerged is an account in which the Pareto conditions serve as a secondary requirement on a solution, that is, as one basis for selecting from among multiple *equilibria*.[11]

Yet Schelling suggests that the game of pure coordination introduces something distinct:

> The intellectual process of choosing a strategy in pure conflict and choosing a strategy of coordination are of wholly different sorts. At least this is so if one admits the "minimax" solution, randomized if necessary, in the zero-sum game. In the pure-coordination game, the player's objective is to make contact with the other player through some imaginative process of introspection, of searching for shared clues; in the minimax strategy of a zero-sum game – most strikingly so with randomized choice – one's whole objective is to avoid any meeting of minds, even an inadvertent one. (1960, p. 96)

need not be optimal. But, again, this poses no problem: Within the perspective of the standard theory, it is plausible to think that, with perhaps certain special exceptions, a rational solution to a pure coordination game will be a combination of choices that is both in equilibrium and optimal.

[11] Moreover, that games of pure coordination involve limited communication suggests that there will even be circumstances under which the Pareto conditions will fail to be satisfied. Consider, for example, the following game:

		c_1	c_2
Game 1	r_1	0, 0	1,1
	r_2	1, 1	0, 0

The choice combinations (r_1, c_2) and (r_2, c_1) both yield the same, maximally preferred or valued outcome, but the players cannot be assured of this outcome without coordination of choice. But how is choice to be coordinated? In the absence of communication or any additional clues, the "default" solution may be that the players end up choosing equiprobably between their two strategies. The outcome, then, has a probabilistically defined value of 1/2, which is Pareto suboptimal (although it is still reached through a strategy pair in equilibrium).

A somewhat different problem arises within Schelling's theory of saliency. On his account, Pareto optimality enters into solution theory in the form of a determinate of the (psychological) saliency of the corresponding combination of choices. And saliency and Pareto optimality *can* conflict. Consider, for example, the following:

		c_1	c_2	c_3
Game 2	r_1	2, 2	0, 0	0, 0
	r_2	0, 0	1, 1	0, 0
	r_3	0, 0	0, 0	2, 2

Here the combination (r_2, c_2) yields an outcome which is suboptimal, but the associated pair of strategies has a saliency that is lacking in the case of the pairs associated with the two optimal outcomes. On Schelling's account (r_2, c_2) may, in virtue of its saliency, emerge as the solution to this game.

In a game of pure conflict of interest, what one hopes for is precisely that the other player does not correctly anticipate one's choice. The appropriate model is hide and seek, avoidance and pursuit. Any anticipation of the choice of the other that works to one's own advantage must work to the disadvantage of the other. But in the pure coordination game, the task is to meet up in a manner that is maximally advantageous to both agents. As noted, of course, pairs of strategies whose outcomes meet the Pareto conditions will also be in equilibrium; but that they are in equilibrium seems quite irrelevant to the concept of successful coordination. Why, then, did Schelling and others choose to develop an account of pure coordination games that pivots on the standard equilibrium condition?

15.4 What Drives the Equilibrium Analysis?

The answer is to be found in the way in which interactive choice was implicitly conceptualized from the outset. Game theory began with a study of the special case of pure conflict of interest. And a solution to such games was first devised for an ideal version of such games, characterized by the following conditions:

Condition 1. (*Mutual rationality*)**:** Both players are rational.

Condition 2. (*Common knowledge*)**:** There is common knowledge of (a) the rationality of both players, (b) the strategy structure of the game for each player, and (c) the preferences that each has with respect to outcomes.

Condition 3. (*Maximization*)**:** A rational player is one who chooses so as to realize the most preferred outcome possible, that is, to maximize a Von Neumann–Morgenstern utility defined over the space of possible outcomes.

The solution proposed was based, in turn, on a thought experiment or "indirect" argument. First, one imagines that a theory of games has been worked out and both players know the theory. Then one traces out the implications of this intellectual situation for what the details of this theory would have to look like. That is, one tries to conclude something about the content of the theory, given the assumptions and an additional assumption that there is a theory that all the players know.[12]

[12] Here is how Von Neumann and Morgenstern (1944), for example, proceed with regard to the search for a theory of two-person, zero-sum games:

> We are trying to find a satisfactory theory – at this stage for the zero-sum two-person game. Consequently we are not arguing deductively from the firm basis of an existing theory – which has already stood all reasonable tests – but we are searching for such a theory. Now in doing this, it is perfectly legitimate for us to use ... [the method] of the indirect proof.

The most frequently cited version of such an indirect argument is the one presented by Luce and Raiffa (1957, p. 63). They begin by assuming that if there exists an absolutely convincing theory of the zero-sum, two-person game, then, under Conditions 1 and 2, mutual rationality and common knowledge, each player will know what the theory tells each of them to do. But given this, they argue, it is plausible to suppose that if the theory picks out specific strategies for the two players, mere knowledge that this is what the theory prescribes should not cause either to choose a different strategy. How could knowledge of this sort cause a given player to choose differently? Implicitly invoking Condition 3, they suppose that a rational player would, given an expectation as to how the other player will choose, select a strategy different from the one the theory prescribes, if the *outcome* associated with that other strategy had a higher utility than the *outcome* associated with the strategy prescribed by the theory, holding the choice of the other player fixed. That is, there would be no reconsideration only if what the theory prescribes is an expected-utility-maximizing response to what the theory prescribes for the other player.[13]

The argument, then, pivots on the assumption that if Conditions 1 through 3 hold, and there is a well-defined theory about how rational players should choose; this (perhaps together with certain additional assumptions) will enable each player to frame expectations about the behavior of the other (also rational) player, expectations that are sufficiently determinate to enable the player in question to treat his or her own decision problem as a simple maximization problem. On this way of thinking, each player is presumed to be in a position to take the behavior of the other (rational) player as a given, as a parameter whose value can be specified.[14]

The notion that one can conceptualize an ideal interactive situation between two rational players as posing a problem of parametric choice for each

> This consists of imagining that we have a satisfactory theory of a certain desired type, trying to picture the consequences of this imaginary intellectual situation, and then in drawing conclusions from this as to what the hypothetical theory must be like in detail. (p. 147)

[13] When r_i and c_j are such that knowledge of the theory would not lead either player to make such a change, then $(r_{i,} c_j)$ will be in *equilibrium.* The derivation of the equilibrium requirement from "best-reply" reasoning, in this manner, via the indirect argument, requires what is now recognized to be a problematic assumption; namely, that there exists for each game a uniquely rational solution. And at best, of course, all the best-reply argument establishes is that being in equilibrium is a necessary condition of a solution.

[14] Notice, moreover, that if one drops the assumption that one's counterpart player is thoroughly rational, one need not abandon thinking of one's problem as calling for parametric reasoning. One can still think of the choice behavior of the other player as simply a variable whose value you must estimate (either subjectively or by appeal to observational data).

player has a rather curious history. At the very outset of the formal study of games, von Neumann and Morgenstern (1944) argue that one encounters a conceptual, as distinct from a technical, difficulty in moving from the study of the isolated individual (the proverbial Robinson Crusoe) to the study of interacting persons. Crusoe's task, given his wants and resources, amounts to a simple maximization problem, complicated at most by the need to incorporate probabilistically defined outcomes. In the case of social interaction, however,

> the result for each will depend in general not merely upon his own action but on those of the others as well. Thus each participant attempts to maximize a function. . . of which he does not control all the variables. (p. 11)

This poses a problem that cannot be overcome simply by appeal to probabilities and expectations:

> Every participant can determine the variables which describe his own actions but not those of the others. Nevertheless these "alien" variables cannot, from his point of view, be described by statistical assumptions. This is because the others are guided, just as himself, by rational principles – whatever that may mean – and no *modus procedendi* can be correct which does not attempt to understand those principles and the interactions of the conflicting interests of all participants. (p. 11)

But the "indirect argument" approach to which not only Luce and Raiffa but von Neumann and Morgenstern themselves appeal involves solving the problem of choice in the context of a game by assuming, in effect, that under the stated conditions, contrary to Von Neumann and Morgenstern's suggestion, parametrization is possible.[15]

Notice carefully what this entails. With the reduction of the problem of interdependent choice to a special case of parametric choice, the choice behavior of the other player is conceptualized as simply another variable whose value the agent must estimate (and then maximize against). This means that the individual agent is presumed, in effect, to be an autarkic chooser – a Crusoe. Thus what characterizes the standard reasoning about interdependent choice is an implicit assumption that in deliberating about the choice of means to preferred outcomes, one can abstract from the context of the interactive problem itself, and simply consider how one would be prepared to evaluate such options, were it the case that one faced a problem

[15] Kaysen (1946–1947), in a very early review of Von Neumann and Morgenstern's work, argues that in the case of games,

> there is no possibility of what we have called parametrization that would enable each agent (player) to behave as if the actions of the others were given. In fact, *it is this very lack of parametrization which is the essence of a game.* (p. 2, emphasis added)

of individual decision making against nature. The intuitive notion, in effect, is that it is one and the same whether the outcome of a choice of an action is conditioned by choices that another agent makes or by natural events. Stated somewhat more formally, what is implicit in the standard argument is an appeal to something like the following condition:

Condition 4. (*Parametrization)***:** Let *G* be any game, and let *D* be the problem that a player in *G* would face, were the outcomes of the strategies available to one in *G* conditioned, not by the choices of another player, but by some "natural" turn of events in the world, so that one faces in *D* a classic problem of individual decision making under conditions of risk or uncertainty; and suppose that one's expectation with regard to the conditioning events corresponds to one's expectations with regard to how the other player will choose in *G*: then one's preference ordering over the strategies in G must correspond to the preference ordering one would have over the corresponding strategies in *D*.

What the argument turns on, then, is not just Condition 3, maximization, but an assumption about how maximization is to be conceptually *anchored*, that is, about what standpoint or perspective it is from which it is to be applied; and it is Condition 4 that provides that specification. Condition 4 assures, in effect, that the *rational* standpoint is an *autistic* standpoint from which the choice of the other player is just another event in the world, one that a rational player must try to estimate (if only probabilistically) and then maximize against that estimate.

Notice, finally, that nothing in the argument sketched above presupposes that agents have conflicting preferences with regard to outcomes. The crucial presumption, then, is that this way of thinking about rational interactive choice carries over to pure coordination problems, as well as to mixed games (involving both conflictual and cooperative dimensions). In this way, the equilibrium condition came to be viewed as a central feature of a general theory of games.

15.5 Rethinking Rational Cooperation

Is it really so clear that a rational player should always view an interdependent decision situation involving another rational agent from the perspective of Condition 4, and thus as posing a problem of maximizing under conditions of parametric choice? Without denying the intuitive plausibility of Condition 3, what I want to do is concentrate on the question of the appropriateness of Condition 4 as a restriction on rational interactive choice.

One can begin by noting that there is something odd about conceptualizing a pure-coordination game as one in which each agent is, in effect, autistically preoccupied with ensuring that the choice he or she makes is a utility-maximizing response to the choice that the other player makes. To be sure, the intelligibility of that approach cannot be denied in the case of zero-sum games of pursuit and avoidance, where the gains to one are matched by losses to the other, so that there are no mutual gains to be realized, gains that require coordination. But that is not the situation in the game of pure cooperation. There the players face a problem of how to *coordinate* their choices – how to "meet up" at an outcome that is maximally preferred by each.

I am concerned here with much more than merely a conceptual point, however. Suppose that the strategies chosen are in equilibrium, but the associated outcome is Pareto suboptimal. It is simply not clear why this constitutes a successful resolution of the coordination problem. It is true that if the outcome results from an equilibrium combination, then each has done the best they can, given what the other has done. And perhaps their expectations are concordant. But what does this have to do with successful coordination?[16]

The problem with the standard theory, however, goes much deeper than just its embracing of the equilibrium condition. Consider the following, very simple game of pure coordination:

Game 3

	c_1	c_2
r_1	1, 1	1,1
r_2	1, 1	0, 0

Suppose that row player expects column player to choose c_1 and column-player expects row-player to choose r_1. We can either suppose that they have been able to communicate to one another and have "agreed" to coordinate on this combination, or that each has been able to imaginatively second guess that the other will choose the "first" strategy. Given an expectation that the other will do his or her part, each has no reason to play in the manner

[16] Lewis (1969, p. 8) makes the following interesting and relevant remark about equilibrium pairs:

> This is not to say that an equilibrium combination must produce an outcome that is best for even one of the agents ... In an equilibrium, it is entirely possible that some of or all the agents would have been better off if some or all had acted differently.

Although this is not his intention, this can be read as a brief against applying the equilibrium condition within the context of games calling for coordination.

specified. Each has just as much reason, from an expected utility perspective, to choose the "second" strategy. On the assumption that each agent is just as likely to select one as another of two utility-maximizing responses, the ex ante return to each in this game is .75.[17] By way of contrast, two players who are capable of appreciating the logic of coordination problems will be able to achieve an ex ante payoff of 1. What this example demonstrates is not simply how constrained an autistic approach is, but the *consequential* costs of deliberating in this manner. Autistic deliberators do less well than those who are capable of coordinating their choices. That is, the objection to be leveled against the standard view is a consequentialist objection. Agents who reason and choose in the standard fashion end up doing less well, in terms of preferred consequences, than those who are able to reason in an alternative manner. But what is an appropriate criterion of rational choice for such games, if it is not the equilibrium condition? What makes sense in this context, I suggest, is a criterion based on the Pareto conditions:

Condition 5. (*The Pareto principle*)**:** Rational agents who know each other to be such will, *ceteris paribus*, confine their choice to strategies that can, in combination with the choices of the other agent(s), generate outcomes that are Pareto optimal and Pareto efficient relative to what each could expect to be the outcome were neither player to attempt to coordinate with the other.[18]

In a game of pure coordination, every outcome that satisfies the Pareto conditions is reached by a set of strategies that are in equilibrium. But it does not follow that the process of what Schelling calls "drawing expectations to a focus" is appropriately modeled in terms of the standard equilibrium theory. The point is simply that the concept of trying to coordinate upon a combination of strategies that satisfies the Pareto conditions can be understood

[17] Notice, of course, that each playing equiprobably between the first and second strategies does not generate an equilibrium pair; moreover, each player's first strategy weakly dominates their second. But the issue remains as to how to get, within the context of this game, from expected-utility reasoning to the equilibrium requirement. The problem is that while (r_1, c_1) is the only equilibrium pair, it is a weak equilibrium. If one has no motive for departing from playing the first strategy, still one has no motive to remain with the first strategy, as long as one assumes that the other will play for the equilibrium. In McClennen (1992), I offer an argument about why I think that a weak-equilibrium pair typically cannot qualify, within the standard theory, as a solution to such a game, given Conditions 3, 4, and 5.

[18] The proviso recognizes the force of Schelling's suggestion that in a game in which there is a unique "second-best" combination of choices, and many "first-best" combinations, one can expect players who cannot communicate with one another to converge on the second-best combination. For an example, see Game 2, note 11.

to furnish an alternative account of the process of convergence. Intuitively, the problem that the two agents face in a game of pure coordination is that each is able to identify a number of combinations of strategies as generating a maximally-preferred outcome, but, by definition, neither agent is in a position to secure unilaterally the outcome that is maximally preferred by that agent. The preferred outcome can be achieved only by that agent *coordinating* his or her choice with the other agent.

Focusing on the set of outcomes satisfying the Pareto conditions leaves in place all of what Schelling and Lewis have to say about the role of psychological salience. Within the framework of the pure-coordination game, where there is a barrier to communication, there is a need for some way of sorting through the alternative combinations that satisfy the Pareto conditions. This model also leaves in place their notion of a process by which expectations are drawn to a focus. Suppose that there is common knowledge of each other's rationality, as well as the strategy and payoff structure of the game, and that both players perceive their task as effectively coordinating their choices so as to achieve what is mutually perceived to be a best outcome. Suppose, in addition, that there is a strategy combination (r_i, c_j) that is salient among the multiple combinations that satisfy the Pareto conditions. To say that (r_i, c_j) is salient is to say that Row expects Column to choose c_j, and that Column expects Row to choose r_i. And given such an expectation on the part of each, and an understanding that the task is to coordinate on a combination that satisfies the Pareto conditions, Row now has a reason for choosing r_i, and Column has a reason for choosing c_j. Moreover, each can replicate the reasoning of the other. It is because (r_i, c_j) is salient that each expects the other to play for this combination. But then each also expects that the other expects that they will play for this combination, and so on. Finally, just as on the standard account, coordinated choice involves conditional strategies: one's best choice depends on what one expects the other player to do. There can be no *coordination* of choice unless each party seeks to anticipate what the other party will do. But it is not part of the logic of this alternative model that one expects that the other player will unilaterally choose to maximize expected utility in response to what he or she expects one to do, any more than that one is required to always choose so as to maximize expected utility in response to what one expects the other to do.

Such an alternative account of cooperation between ideally rational beings carries with it one very important implication. On this account, rational beings will not approach pure coordination problems as if they call for reasoning in accordance with principles of parametric choice. Thus, the model brackets Condition 4, parametrization, and its implications. To be sure, since in pure coordination games outcomes satisfying the Pareto conditions are invariably also equilibrium outcomes, their choice behavior will not, in fact, be inconsistent with such principles; but those principles, in effect, will play no role in the reasoning of such persons.

15.6 Mixed Games

However satisfactory the revised account of pure-coordination games might prove, we face a substantial issue in any attempt to extend the idea of such coordination to mixed games. Here we are forced to choose between the equilibrium reasoning to which we are led by invoking Condition 4 and the optimality condition embodied in Condition 5. Unlike the case of games of pure cooperation, in the case of mixed games, optimal outcomes are characteristically not in equilibrium. Moreover, mixed games raise an issue that posed no problem in the case of games of pure cooperation. In the case of games of pure cooperation, it makes no difference which optimal outcome is selected – all are equally good. In the case of mixed games, there may be many different optimal outcomes between which the players will not be indifferent. That is, there is an issue regarding how the gains from cooperation are to be distributed. On the standard way of thinking one has no choice but to fall back on taking the equilibrium requirement as essential and be content with appealing to the optimality condition as only a secondary criterion that can be invoked in cases in which one or more of the equilibria are also Pareto optimal. In such cases, it is presumed that the solution will be one of the equilibrium outcomes that is also optimal. However, in cases in which optimal outcomes are not in equilibrium, coordination can play no role, even though it would enable the two players to do even better.

15.7 Repeated Interaction over Time

Matters prove to be different if it is supposed that a game is to be played an indefinite number of times. This is the focus of what are known as the "folk theorems" concerning indefinitely repeated games. Here the question of assurance is bound to loom large, even for agents who are otherwise predisposed to cooperate. What is characteristic of indefinitely iterated games is that multiple (subgame perfect) equilibria can be identified, equilibria, which are typically Pareto suboptimal, while combinations of strategies that do satisfy the Pareto conditions are not subgame perfect equilibria. Now, intuitively, the fact that any given player will repeatedly encounter other members of the group suggests that it should be possible for them to work out, tacitly or explicitly, some sort of coordination scheme, which will ensure that they do not have to settle for a mutually suboptimal outcome. However, since the outcomes of such coordination schemes typically do not satisfy the (subgame perfect) equilibrium condition, their stability will have to be secured artificially by the introduction of an appropriate system of rewards and punishments. The "folk theorems," then, explore various sets of conditions under which interactive games over time lend themselves to resolution in terms of Pareto-optimal and Pareto-efficient coordination schemes, backed up by appropriate sanctions.

The work in this area is quite technical and not easy to summarize; however, a number of different types of sanction systems can be distinguished. First, an *informal* arrangement (expressing a norm of reciprocity) may emerge, in which each participant is motivated to conform to a cooperative agreement governing pair-wise interactions, by an expectation that defection will be met by retaliations whenever that same partner is encountered again, retaliation whose expected cost outweighs the immediate gains to be secured by defection. These results are, of course, sensitive to the probability that one will encounter the same player again, one's discount for the future, and the severity of the loss involved in having that other player refuse to cooperate on future encounters. Second, some of these limitations can be overcome if there is more widespread reporting and punishment of defectors, specifically if there is an *informal institutional* arrangement under which others in the community will also retaliate against anyone identified as a defector. This presupposes, in turn, some sort of communication system between the participants, so that all members of the group can identify the defectors. Third, because the costs of such a communication system are nonnegligible, efficient community (as distinct from individual) enforcement may require the centralization of the reporting system (as, for example, takes place in the case of a centralized credit bureau reporting system). Finally, of course, one may sustain cooperation by employing *formal institutional* arrangements, that is, an enforcement mechanism, involving third-party surveillance and apprehension and a legal system dispensing appropriate punishments. What has become increasingly clear, however, is that whichever method of enforcement is employed, there are significant associated costs of surveillance and enforcement. A central concern, then, becomes that of comparative costs of alternative schemes.[19]

Notice that within the context of indefinite iteration, it is almost inevitable that there will be any number of different arrangements that could

[19] For an excellent survey of work in this area, see Calvert (1995).

be reached, differing from one another in terms of their distributive implications. The formal work in this area, however, does not address this problem at all. The thrust of the theorems is simply to establish that if there are outcomes that are Pareto superior to what could be expected to be the outcome of purely noncooperative interaction, there will exist ways to arrange sanctions such that any one of those Pareto-superior outcomes can be achieved as a result of a strategic plan of surveillance and enforcement that ensures that the equilibrium condition is satisfied.

15.8 Bargaining Theory

What is needed to complete the standard account, then, is a theory of bargaining, which can settle the issue of the distribution of gains from cooperation and thus provide a specific outcome that will be accepted as the solution for each particular iteration of the game. The most widely accepted theory of bargaining is one due to a generalization by Harsanyi of Nash's original two-person bargaining theory. The essential features of this theory are as follows:

i. Rational players who know each other to be such will not fail to achieve a Pareto-optimal outcome that is Pareto superior to what they could achieve if they were not to cooperate. That is, faced with the recognition that agreements can generate mutual gains, persons will be disposed to continue to negotiate over that game in the series, as long as there are mutual gains to be realized.

ii. The particular point on the Pareto-optimal frontier that is finally settled on will be determined by a generalization, introduced by Harsanyi, of the two-person bargaining model developed by Nash.[20] The Nash theory views the outcome of a two-person bargaining game as a function of the relative threat and bargaining advantages of the two players, and the *n*-person version involves decomposing the *n*-person game into all the possible two-person games embedded in it. The payoff for each player in the *n*-person game is an additive function of the payoffs from the two-person component games. Relative threat advantage for

[20] The seminal work in this area consists of two papers by Nash (1950, 1953). There is an alternative model of bargaining that has received some attention. See Kalai and Smorodinsky (1975). For a discussion of the differences between them, see, for example Roth (1979). Both the Nash and the Kalai and Smorodinsky models, however, rely on the concept of an equilibrium of forces. For the equilibrium of forces interpretation of the Kalai and Smorodinsky model, see Gauthier (1986, ch. 5). Harsanyi's generalization to the *n*-person case is to be found in Harsanyi (1963).

each of the component two-person games is determined by the underlying *noncooperative* game that establishes a background against which two-person negotiation takes place.[21] And relative bargaining advantage is a function of the shape of the set of outcomes that are Pareto optimal and Pareto efficient relative to the outcome of no agreement. In effect, each can be expected to hold out for the best that he or she can do, given his or her relative threat power and bargaining advantage and also be willing to settle for that amount. In all of this, one can mark a deep indebtedness to concepts borrowed from physics: The final point on the optimal frontier is essentially determined by a complex equilibrium of forces.[22]

iii. In accordance with the previous discussion, it is assumed that some mechanism can be put into place to bind the parties to the terms of whatever agreement is reached, that is, some mechanism analogous to the sort that is presupposed within the context of the "folk" theorems of indefinitely iterated games. Since the game is an *n*-person game, the enforcement mechanisms will have to be more complicated. We are no longer in a world in which each player can be assumed to have full knowledge of what some particular other player did on previous rounds. That is, retaliation will have to be organized by the group as a whole. In terms of the arguments previously explored, the supposition is that this can be accomplished by some system of social or third-party enforcement mechanisms.

This model of bargaining marks the triumph of what is known as noncooperative game theory, in which even problems of cooperation are to be solved by appeal – in the case of the crucial assumption (ii) – to a theory of purely competitive interaction. To be sure, the need for (iii) points to one shortcoming of this model. In any realistic case, the surveillance and enforcement devices employed will prove, over time, to be costly, and only partially effective. Resources must be continually expended to maintain even

[21] Relative threat advantage sets the "status quo" point from which bargaining takes place, and relative bargaining power determines, given the status quo point, where on the optimal frontier the solution will lie.

[22] There is a very thorough discussion of the relation between the two models in Harsanyi (1977a, pp. 149–166). There is a purely formal route to the derivation of one dimension of this result (what is known as the "simple" bargaining model for two players with an exogenously specified status quo point), namely, the Nash bargaining model, but these axioms give exactly the same result as a theory put forward by the economist Zeuthen (1930), in which the solution is conceptualized as emerging from an idealized seesaw process in which each side makes just enough of a concession to place the other player in a position where he or she now has to make the concession.

a partial level of compliance; and full compliance could be secured only with an unreasonable expenditure of resources.

But the problem goes even deeper. Presumably, bargaining can be expected to take place just once – the first time the game is played, and, at the same time, an agreement will have to be reached about enforcing the arrangement. Because the game does not change, all the parties will then expect that the same solution will emerge for each successive game, so negotiation costs will be at a minimum. Moreover, each participant will expect that each successive iteration will be enforced in the same manner. However, given all this, why would rational players agree to a system of surveillance and enforcement that will bind them to indefinitely accept bargaining terms that are determined by the distribution of threat advantage and bargaining power? Such participants are, in effect, being coerced twice. The terms of the bargain for an indefinite iteration of the game are settled coercively, and then they are expected to agree to coercive measures to ensure that the agreement is maintained over an indefinite sequence of iterations. To suppose that this whole process will unfold smoothly seems doubtful.

15.9 A More Realistic Setting

What happens if we move to a more realistic setting? Imagine, for example, that a sequence of interactive situations takes place, where once again (1) each such agreement is one that persons enter into from a sense of what will serve their own interests and (2) these are interests that are not shared in common with the other participants, that is, the agreement is between persons whose own interests are disparate. Unlike the indefinitely iterated situation, let us also assume that (3) each agreement in the series is expected to hold over an indefinite period of time, but where (4) each subsequent agreement in the series represents a renegotiation of the previous agreement that took place. A situation conforming to this description would be one in which a contractual arrangement is perpetually renegotiated between roughly the same group of interested persons, and while there are mutual gains to be realized by cooperation between them, their interests are not coincident. That the arrangement on any given round needs to be renegotiated reflects the fact that, over time, circumstances change, including the relative bargaining and threat advantages of each player, thereby creating a series of new bargaining situations. Finally, we suppose that each bargain reached is reinforced by some sort of surveillance and enforcement system, so as to ensure (at least for the time the bargain remains in force) that most persons will comply with the bargain.

If we make projections regarding the consequences of ongoing interaction between rational individuals in this sort of setting, a number of things become clear. Once again, the surveillance and enforcement devices employed will prove, over time, to be costly, and only partially effective. Second, perpetual renegotiation will be unavoidable, since under changing circumstances this or that party to the agreement will anticipate that they are now in a position to do better than they had the previous time; that is, changing circumstances will perpetually alter the distribution of threat and bargaining advantage. Participants must recognize that negotiated (and renegotiated) distributive terms will reflect the distribution of bargaining advantage and that they must, then, reckon with the real possibility of reversals of their own fortune over time. This poses a serious problem for each participant. The surveillance and enforcement mechanisms will, presumably, be fixed; they will not vary over time according to the distribution of advantage. But this means they will have to accept surveillance and enforcement mechanisms that will force them to comply with terms in the future that they may find onerous.

Can we assume that the outcome of each round will be Pareto optimal? The standard model takes it as given that rational players will continue to bargain until an optimal outcome is reached. There is, however, considerable tension between the idea that all the players will cooperatively arrive at an optimal outcome, on each round of interaction, but that which optimal outcome is selected is a matter of relative threat and bargaining advantage. What we have, in effect, is a theory of noncooperative interaction – a theory of strategic interaction – onto which has been grafted, quite incongruously, the idea that the outcome will be optimal.

The potential incoherence of the standard model becomes all the clearer when one realizes that it assumes that perceptions of fairness play no role in such an ongoing form of interaction. The evidence, however, suggests this is not the case. In particular, the findings of Fehr and others on strong reciprocity are of considerable relevance here.[23] Their work focuses on the case in which the disposition is operative even in the one-off situation. That is, their hypothesis is that persons will be disposed "to sacrifice resources for rewarding fair and punishing unfair behavior even if this is costly and provides neither present nor future material rewards for the reciprocator."[24] But in the context of ongoing, repeated interactions, one would expect that considerations of fairness will play a significant role – that persons will resist what they

[23] See Fehr, Fischbacher, and Gachter (2002).

[24] Fehr, Fischbacher, and Gachter (2002, p. 2).

think to be a distribution judged unfair to themselves, with a view to setting a precedent for future negotiations. The folk theorems are designed to show that the shadow of the future can play a role in ensuring compliance, but it could also play a role in how negotiation is shaped over time. Given ongoing negotiations, one would expect that rational participants will resist the efforts of others to settle distributive terms solely by reference to relative power considerations, and that the use of such power relations by some will occasion continuing conflict that will inevitably work to the disadvantage of all.

These various considerations suggest that, at least in general, strategic interaction in terms of threat advantage and bargaining power will not yield an optimal outcome. Such interaction will take place in terms of the kinds of considerations that define noncooperative games, typically at the cost of realizing an optimal outcome. I say "in general" because there will be situations in which strategic considerations and optimality conditions can be reconciled. An obvious case will be competitive sports, in which strategic interaction is the rule but where considerations about what would be mutually beneficial to competitors clearly have a role to play. Think about competitive members of a sport's league who agree to a draft system for the selection of new players, with the teams that did least well the previous year having first pick. A parallel case can be made for the importance of strategic competition in the marketplace, where there are very well-defined rules governing what competitive units can and cannot do.

15.10 An Alternative Model of Cooperation

Given the previous considerations, I propose to explore an alternative account of how a particular optimal outcome could emerge as solution to a given round of such an ongoing game. I shall then go on to show that this alternative view of cooperative interaction does not have the defects I have attributed to the standard view, namely, that compliance with its terms would require costly surveillance and enforcement devices, extensive negotiation costs, and would likely occasion the kinds of conflict that would constantly undermine the optimality of the solution. We are looking, then, for an account in which the solution is picked out in some other way than by application of *noncooperative* game theory, where compliance with it is ensured in some other way than by costly surveillance and enforcement mechanisms, and where negotiation engenders cooperation rather than costly conflict.

Let us start by supposing, in rational bargaining, as the standard theory does, that Condition 5, the Pareto principle, is satisfied. Our rational individuals, in virtue of accepting Condition 5, have an interest in gaining as much as possible consistent with others also gaining. But nothing has been said yet about their attitude toward how the benefits are to be distributed. Let us also suppose, however, that they will not demand a larger share than others, unless and only to the extent that unequal shares can be shown to work to the advantage of all and not just to themselves. Correspondingly, we shall suppose that they will not accept a smaller share than the others unless and only to the extent that such an unequal distribution can be shown to work to their own advantage. Thus, each accepts that the criterion for a justifiable unequal distribution is that such an unequal distribution works to the advantage of all. This yields the following additional condition:

Condition 6: There is a presumption in favor of equal sharing of benefits, but inequalities can be justified to the extent to which they work to the advantage of each participant.

When the arrangement meets both Conditions 5 and 6, I will speak of it as a *fully cooperative* arrangement. When an arrangement satisfies both Conditions 5 and 6, one can also say of those who negotiate it that each thereby signals to each other participant his or her intention to accept the principle of *mutual* gain, both in respect to the arrangement and to its distributive implications. The reference to intentions is important. I am supposing that the arrangements to be settled on are understood to extend over time and to be adjusted, from time to time, to reflect changes in the background environment. A full cooperator is disposed to refrain from acting strategically and limits himself or herself to strategies that work to the mutual benefit of all participants. But if one is to refrain from pressing one's own strategic advantage, not just now but also in the future, one needs assurance that others will also refrain, just as persons who are disposed to be cooperative need assurances that others will respond in kind and not take advantage of them by free-riding on the agreement. How persons negotiate the various features of the arrangement, at any given point in time, then, is crucial. It helps to provide each participant with some sort of assurance about what they can expect others to do in the future.

Under conditions of full cooperation, each participant will be aware that their own interests, as well as the interests of others, have been taken into account systematically in the design of the arrangement, and this will again signal something important about their intentions toward one another. If some were to seek to alter the arrangement in a way that worked to their own advantage but to the disadvantage of other participants, what they would signal thereby is that they are not prepared to look for ways in which mutual gains can be realized, that they are willing to press for unilateral advantages. However, they would also send the wrong signal if they were to press for a change that benefited everyone, but in a differential manner that could not be defended to everyone, that is, that did not satisfy Condition 6. Again, it is intentions rather than the actual terms they settle on at any given point in time that are important. After all, the actual terms (unequal or equal) could have been the outcome of the distribution of threat and bargaining advantage and thus be expressive of very different intentions on the part of participants.

An agreement that satisfies not just Condition 5 but also Condition 6 will possess, I suggest, two important features. First, full cooperators will not free-ride. Full cooperators will, by definition, be committed to a particular distributive formula, one that meets Condition 6. However, to reach such an agreement, and then choose to free-ride would be, in effect, to alter that distributive formula in a way that increases the payoff to the free-rider (he or she avoids the cost of compliance). But that means that the distributive formula has been altered in a way that cannot be shown to work to the mutual advantage of all those involved. That is, the distributive formula that results from free-riding will not satisfy Condition 6. It is thus the mark of a fully cooperative arrangement between persons that a participant will not be disposed to free ride while others comply, and insofar as a player is aware that others understand the arrangement, he or she will expect that those others will not free-ride either. Among full cooperators, then, there will be no need to make a costly investment in elaborate enforcement mechanisms. It also means that full cooperators will, in effect, be able to treat any optimal outcome on which they agree as also a natural equilibrium outcome, given their cooperative dispositions.

There is another important feature of full cooperation. All can expect that over time, changing circumstances will call repeatedly for changes in the arrangement. New circumstances will mean, in some cases, that the arrangement ceases to be one from which every person benefits, that is, either Condition 5 or 6 will now fail to hold. This means that full cooperators will be disposed to renegotiate the agreement. However, in negotiating either kind of needed change, the parties involved will expect that the same considerations that shaped the original agreement will shape the alterations. That is, those who negotiate in a fully cooperative manner will not be disposed to renegotiate in a manner that simply secures more for themselves at the expense of others. In contrast to strategic negotiation, there is no reason to suppose that such negotiations and renegotiations that take place will occasion mutually disadvantageous conflict. Participants will not have to worry that changing circumstances can leave them in a greatly weakened bargaining position. Thus one can expect that full cooperators will settle on an arrangement, and a distributive rule, that satisfies the optimality condition. To be sure, under conditions of declining prospects all may have to settle for less, but the new distribution will still apportion the losses to be incurred in an equal manner or apportions losses unequally, if this would work to the relative advantage of all. This is not to say, of course, that the adjustment must work to the advantage of all, compared with how things were arranged previously. The relevant question is whether the resulting distribution meets the conditions of full cooperation.

15.11 The Case for the Revised Model of Cooperation

The defender of the standard theory will, no doubt, want to insist that the presuppositions of their theory capture in a particularly compelling and deep way features of what it means to be rational. What I have suggested in reply is that they have a plausible case to make with regard to maximization but that their argument goes through only by appeal to the additional and much more questionable Condition 4, parametrization, and to an account of bargaining that incoherently tries to develop a theory of cooperation based on a theory of noncooperation.

I began the critique of Condition 4, parametrization, in Section 15.5 with what is essentially a conceptual point. Parametrization seems quite out of place as a condition on rational choice within the context of pure coordination games. The problem of coordination calls for one to think about how to meet up with another and not simply maximize against one's best estimate of what the other is likely to choose.[25] That point carries over, I suggested, to the problem of coordination that is posed by mixed games, but in that context, a quite distinct consideration emerges to militate against parametrization. Agents who deliberate in that manner do less well than

[25] The problem isolated here is closely related to characterizing the conditions of trustworthiness. Despite the way in which most decision theorists proceed (see, e.g. Hardin, 1993), I think we should insist, following Morris (1999), that trustworthiness among agents is not something that is secured by establishing conditions under which conformity to some rule is utility maximizing for each agent, given how each other agent chooses, that is, it is not just a matter of ensuring that the coordination scheme is in equilibrium. On the contrary, trustworthiness involves being rule-governed in one's choice behavior.

those who can interact with one another in the manner described in the alternative model. Under the ideal conditions discussed in Section 15.7, and the more dynamic and realistic settings discussed in Sections 15.9, the standard theory implies that compliance will typically require a comparatively costly system of surveillance and sanctions to ensure that participants will abide by the terms of a mutually advantageous arrangement. Those who can act voluntarily in accordance with such arrangements and rules can thus expect to realize savings with respect to enforcement and surveillance systems. Such a way of organizing their joint activities will be more efficient. This is a *consequentialist* argument for not approaching every interdependent choice situation from the parametrization perspective developed by the standard theory. A fully cooperative approach to bargaining is one that will better serve the interests of the participants than one that is not so governed. This is not simply a conceptual point. On the contrary, the argument is that those who are capable of full cooperation do better in terms of furthering whatever interests they have than those who do not and that this has direct bearing on the credibility of claims about what counts as a rational approach to various forms of interaction.

The case for the revised account of cooperation, however, extends well beyond the point about savings in surveillance and enforcement costs. The model of full cooperation imagines participants who are committed to advancing their own interests, but only insofar as that can be done in a manner that also advances the interests of the other participants. There is no reason for participants to be concerned, then, that subsequent negotiations will take place in a manner that works to their own disadvantage, as a result of others pursuing their own interests unilaterally. What will not take place is the kind of competition over benefits that can lead to deep and continuing conflict between persons – conflict that can result in mutually disadvantageous outcomes.

I have argued that strategic interaction is virtually assured to be suboptimal and that fully cooperative interaction will, in contrast, result in an optimal outcome. It does not follow, of course, that the latter form of interaction is Pareto superior to the former. To establish that would require a much more detailed exploration of the causal effects of both forms of interaction than I can present here. I think, however, that this stronger thesis is a plausible conclusion, at least insofar as we are considering the long-term effects of both kinds of interaction: strategic and fully cooperative. The skeptical reader may insist, of course, that even were this stronger thesis shown to be true, I would still be begging the central issue if I were to say that this makes the case for the *rationality* of full cooperation. The point, presumably, is that I have abandoned the equilibrium condition as a requirement of rational interaction. On the standard way of thinking, I have described a model (of full cooperation) that has no application to rational interaction in virtue of its failure to meet the equilibrium condition. What I have argued, however, is that the equilibrium theory gives us no account of coordination and hence offers a questionable account of the conditions for an effective pursuit of the gains that coordination can realize.

15.12 Rule-Governed Choice

If decision and game theorists have tended to think about coordination problems in the wrong way, philosophers for their part have thought about rule-guided behavior in the wrong way. It is often argued that rule-governed behavior is rule-*bound* behavior and, as such, is consequentially indefensible. The suggestion seems to be that what is defensible from the point of view of consequences is that a flexible policy of making exceptions to following the rules be adopted, whenever allowing for such exceptions contributes to the realization of the very ends for which the rules themselves were adopted. On this account, rules are to be understood as maxims: "rules of thumb." With the extension of the proposed alternative model of coordination to cases in which there is ongoing interaction between individuals, who can tacitly or explicitly agree on various coordination schemes, however, we arrive at a model in which sense can be made of a quite different way to think about rule-governed choice. Within that context we can imagine persons who are capable of tacitly or explicitly agreeing to a set of rules that define a coordination scheme and then choosing in a rule-governed manner, even when that choice is not supported by standard (expected) utility-maximizing considerations. For such persons, the central question is *not*, What available alternative is mandated by expected utility reasoning? but rather, What course of action is called for here, given some coordination scheme on which we have tacitly or explicitly agreed, and which we have each judged to be to our own advantage? This is not to say that such persons are slavishly committed to rules. The commitment to a given rule will characteristically be a conditional commitment. That is, it will be a defeasible commitment, one that can be set aside by reference to any one of a number of considerations. All that we need to suppose here is that such a person will not be disposed to free-ride on the cooperative dispositions of others.

It is precisely this sort of choice behavior that the standard account of rationality cannot accommodate. Parametric reasoners cannot make even this very modest sort of "commitment" to act in accordance with a scheme or rule. For such a reasoner, what was agreed upon at some point in the past will characteristically be irrelevant. The only relevant question will be whether the expected utility of acting in conformity with that plan or rule is greater than the benefits to be secured by deviating. Such reasoners will, of course, be able to implement coordination schemes. But because each will be disposed to free-ride on the coordination commitments of the others, they will typically have to arrange to have the coordination scheme backed up by some sort of enforcement device.[26]

15.13 The Efficiency-Egalitarian Principle

Full cooperators are committed to the idea that inequalities in the distribution of the benefits of cooperation are acceptable only to the extent that they work to the mutual advantage of all participants. Let us call this the *egalitarian-efficiency principle.* This is a class of distributions that are significantly less likely to engender noncooperation or conflict. These are less likely to engender noncooperative or competitive responses from those who receive lesser shares because those individuals would be even less well off were the inequality to be reduced as a result of some redistributive measure. Such, at any rate, is what can be expected insofar as there is a shared or public recognition of the efficiency implications of the inequality in question. Correspondingly, then, this is one structuring principle that is specifically responsive to the problem of noncooperation and competition that inequalities can generate.

Now, a version of this principle figures centrally in Rawls's seminal work, *A Theory of Justice* (1971). There the reader is invited to contemplate a modified egalitarian principle for evaluating the social and political structure, according to which features of the basic social structure of a given society that gives rise to inequalities in life prospects can be justified if, but only if, it is the case that, were that feature to be eliminated, the expected benefits of social cooperation would be reduced for each and every representative participant. By way of marking out an important conceptual point here, let us say of a nonegalitarian feature of a social structure that works to the mutual

[26] To be sure, an obvious functional substitute for social sanctions or formal enforcement system is a set of shared cultural beliefs regarding the "appropriateness" or "correctness" of acting in accordance with the rules in question. Careful accounts usually acknowledge this point. But the appeal to shared cultural values is an appeal to what is not a matter of a deliberate, and voluntary, agreement between persons. It involves an appeal to a "nonrational" or ideological factor. More about this shortly.

advantage of all that it is an *efficient* inequality. Rawls's principle, then, proscribes all but efficient inequalities. More precisely, it requires the social arrangement to be as egalitarian as possible, consistent with the demands of Pareto efficiency.[27] This is, and is intended by Rawls to be, a very stringent condition. The test is not whether, were all nonegalitarian features eliminated, each would be worse off; rather, the test is to be applied at the margin. Specifically, consider any arrangement that involves an unequal distribution of benefits, and consider now any modification of that arrangement in the direction of a more egalitarian distribution. If that change would *not* work to the disadvantage of all who participate, then the principle in question justifies that change. Alternatively put, the principle requires that one select the most egalitarian of the efficient arrangements.

Just what kinds of nonegalitarian arrangements could be defended in this manner is, presumably, dependent in part on historical and other background conditions, but most importantly on considerations about what would in fact psychologically motivate or provide incentives for persons in this or that setting.[28] It is plausible to suppose, however, that there are significant classes of inequalities that will prove to satisfy the principle and thus that will not (at least under conditions of public information) engender competitive, conflictual interactions.

Any civil order presupposes a certain amount of hierarchical differentiation of roles. Thus, for example, some must be granted authority over others, in the interests of both defense against external threats and internal order, as well as for the efficient organization of other kinds of coordinated activities. Similarly, it can be argued that a competitively organized system of private exchange markets based on the sanctity of private property rights is absolutely essential for mutually advantageous economic growth and development, even though it inevitably leads to significant economic inequalities. In this case, the argument is that differential rewards are needed in order

[27] In his original discussion, Rawls (1971, section 13) considers a set of conditions under which his difference principle (which is essentially a leximin principle) is equivalent to the egalitarian-efficiency principle. Moreover, he appears to take the leximin formulation as more basic. My approach here is to take the egalitarian-efficiency formulation as more basic, but space limitations preclude my doing full justice to this matter. On the egalitarian-efficiency formulation, the levels of well-being associated with a strictly egalitarian arrangement serves as a natural benchmark. For the reason explained in the text, however, one cannot thereby justify any nonegalitarian arrangement that raises everyone's level of (expected) well-being. On the principle in question, one must seek the most egalitarian of the efficient arrangements.

[28] One wishes here that economists, who make so much of the problem of incentives, had more to offer us by way of a carefully worked out theory of the same.

to provide appropriate incentives to persons to contribute their fair share of effort or develop their talents and abilities in ways that prove to be mutually advantageous. In keeping with the spirit of the efficiency-egalitarian principle, it is plausible to suppose, however, that they will insist on certain substantive and procedural constraints that will keep inequalities within definite limits. In the case of the alleged need for hierarchical structures, we may follow Buchanan and Tullock's (1962) lead and suppose that such individuals will nonetheless insist that certain fundamental rights are to be equally distributed, that there is equal citizenship in a political sense, and equal participation in majoritarian voting rules.[29] In somewhat parallel fashion, we may suppose that such individuals will also be reluctant to let their economic prospects be determined by a competitive market system unless they are assured that, in the event that they are disadvantaged by such a system, through no lack of effort on their part, their basic needs will be met.

15.14 The Problem of Ideology

The efficiency-egalitarian principle is, of course, very demanding. It is predicated on the hypothesis that by allowing only those inequalities that meet the test of mutual advantage one can minimize surveillance costs and the costs of mutually disadvantageous conflict. But this hypothesis is surely subject to an important qualification. I have argued the case for full cooperation from the standpoint of a citizenry composed of people who are prepared to view their arrangements with others from an instrumental perspective and who are primed to ask the question whether their arrangements can be shown to serve their interests. In reality, the tendencies of arrangements to modulate what would otherwise turn out to be conflictual and mutually disadvantageous interaction will be conditioned by the absence of a virtually open-ended set of possible ideological beliefs. Even the most egalitarian of arrangements can engender conflict if some of those involved share a deep conviction that they have a special entitlement to a larger share. And correspondingly, a deeply nonegalitarian social and political order may yet be relatively efficient, and stable, if it is underpinned by widespread acceptance of a belief system that stresses the value of a nonegalitarian, hierarchical order. Plainly and simply, ideology may lend support to the subordination of some to others in respects that cannot be ratified by reference to ordinary

[29] Buchanan and Tullock (1962) treat the majoritarian principle as rationalizable by reference to mutual advantage (in many settings it will involve lower net costs as compared, say, to any supermajority principle, including, at the limit, the principle of unanimity). But, historically, majoritarian principles have also tended to reduce socioeconomic inequalities.

conceptions of common interests. Notice also that such an ideology may make appeal to values that directly ratify the nonegalitarian features of the system of relations; but the ideology may also function as the result of propaganda efforts that lead to widespread acceptance of false beliefs regarding, for example, what serves this or that shared interest.

There may even be other grounds on which inequalities could be justified in such a way that their presence would not engender mutually disadvantageous conflict. After all, the egalitarian-efficiency principle itself sanctions a whole class of inequalities. My own sense, however, is that aside from the inequalities sanctioned by that principle, there could not be any other nonideological grounds of this sort – that arguments for inequalities that are not grounded in considerations of mutual advantage are likely to be ideological in the classic sense: They are designed to serve the interests of some at the expense of others. Such at any case seems likely from a liberal, secular perspective. Whichever way that issue is resolved, however, this much seems clear, in the presence of certain ideological beliefs, a much larger class of inequalities can presumably be reconciled with efficient interactions between persons.

Moreover, differential bargaining power and ideology often go hand in hand. Agreements that are predicated on nothing more than the distribution of power can, of course, prove very unstable. Those who are advantaged by a given informal arrangement will have a motive, then, to encourage beliefs about the legitimacy of the particular distributive formula, and to this end, they will invest resources in disseminating and inculcating (through education, indoctrination, and other means of socialization) viewpoints that serve to rationalize their interests.[30] Given what we know about the transmission of cultural values, such nonegalitarian systems of social and political order can prove remarkably stable.[31]

It seems clear, then, then, that we can expect, given chance events and path-dependency, that both "permissible" and "impermissible" nonegalitarian forms of social organization are likely to emerge, under a variety of conditions, in response to economic needs and interests. Indeed, if we suppose that arrangements between persons may be articulated in ways that permit even a very significant role differentiation and distribution of benefits, it is plausible to suppose that, in a wide range of cases, permissible

[30] See Ullmann-Margalit (1978, part 3), for a sustained discussion of how this sort of reinforcement can work. It is a remarkable fact that while her discussion, in parts 1 and 2, of how norms arise to solve coordination and prisoner's dilemma problems has received extensive discussion and citation, virtually no mention is made of part 3!

[31] See Knight (1992, pp. 80–82, 185–188) for a discussion of the role of ideology in stabilizing expectations.

and impermissible arrangements will constitute functional substitutes, at least from the perspective of securing many of the gains of cooperation.

One can try to argue, of course, that ideologically based nonegalitarian structures may well prove, at least in certain contexts, to be inferior substitutes. The point would be that opportunism is likely to be encouraged to a much greater extent in nonegalitarian as opposed to egalitarian structures. What is missing from nonegalitarian structures that cannot be ratified by the requirements of full cooperation as expressed in the efficiency-egalitarian principle is an explicit norm of reciprocity. The suggestion, then, is that this is likely to function, in the case of egalitarian networks, to modulate opportunism and rent-seeking. Some have even suggested that one can mark here the existence of two equilibria, based, respectively on trust and distrust, each of which is self-reinforcing, and one of which – trust – is Pareto superior to the other.[32] Generalizing this, it seems plausible to suppose that interactions organized on a principle of reciprocity will tend toward a self-reinforcing equilibrium that is Pareto superior to the self-reinforcing one to which opportunistic, rent-seeking behavior tends. But any optimistic story here will have to be substantially tempered by a consideration of the role that ideological factors can, and do, play.

15.15 The Question of Adaptive Efficiency

A more promising line of argument, I suggest, is that one will typically find both types of organization, those based on special interests strategically using ideological measures to stabilize their control of the distribution of benefits and those based on full cooperation, each faced with adapting to changing technological, natural, and social conditions. This, in turn, suggests that the continued success of a particular system of relations may be a function of whether it can effectively adapt to changing conditions. North (1990, pp. 80–81) characterizes such adaptive efficiency in the following way:

> Adaptive efficiency is a dynamic concept that is concerned with the willingness to acquire knowledge and learning, to induce innovation, to undertake risk and creative activity of all sorts, as well as to resolve problems and bottlenecks of the society through time. . . . Institutions should encourage trials and eliminate errors . . . to explore many alternative ways to solve problems. It is equally important to learn from and eliminate failures.

This kind of efficiency is very different from allocative efficiency. It is not concerned with how efficient a particular arrangement is at a given

[32] For an argument of this sort, see the final chapter of Putnam (1993).

moment in time but with the performance of the arrangement over time and especially with how effective it is at dealing with changing conditions. The question, then, is whether a case could be made against ideologically based arrangements by appeal to adaptive efficiency.

What are the prospects for adaptive efficiency when a considerable intellectual investment has been made in *masking* certain of the true consequences of the institutional arrangements then in place, in presenting a biased distribution as if it were justifiable in terms of the principle of mutual gain? The imperatives of adaptive efficiency call for (1) close and careful attention to social facts, (2) a commitment to empirical experiment, and (3) the systematic dissemination of the information to be gained thereby. The reliance on ideology is antithetical to the first of these requirements. Ideological justifications, if they are to be successful, require that a close and careful attention to the relevant social, economic, and political facts be replaced by a systematic misrepresentation of those facts. In the face of such a misrepresentation, there will be no reason to suppose that persons will be in a position to judge just how effective various adaptive measures will be. Even if those in power can determine how to adapt to changing circumstances in a manner that works to their own advantage, their preoccupation with their own interests, and especially with maintaining control, may effectively blind them to the possibilities for even greater gain through cooperation.

Second, what is needed to deal effectively with changes, both small and large, is a special kind of learning that comes from direct and multiple experimentation. This calls, among other things, for decentralized decision making that will permit a maximum exploration of alternative ways of doing things. But those who have successfully structured civil institutions to their own advantage will likely resist such decentralized experimentation.

Moreover, such experimentation will have little impact unless it is coupled with the reliable (accurate) and open exchange of information derived from the experimentation, information not only about what works and works effectively but also about what does not. However, the generation and circulation of reliable information of this sort creates a climate that is likely to be viewed as not in the interests of the advantaged faction, for it encourages people to continually challenge claims about what works well, especially when what is being circulated includes important information about what changes work and what changes do not.

The degree of success that can be achieved by an ideological justification of the main institutions of society will surely be a function, in part, of the degree of control that those who are advantaged have over the dissemination of information. It must be granted that where the advantaged can completely control who is allowed to know what, a relatively stable order may be achieved. But the requisite factional control of information is still likely to be upset for at least three distinct reasons.

First, in the modern age at least, the worldwide dissemination of information makes it increasingly harder to maintain the kind of "closed" conditions under which control of relevant information by the advantaged would be possible. This will be especially true where other states or societies, that have moved toward more democratic and egalitarian forms of governance, are prepared to use the means available to reach the disadvantaged members of the less democratic societies with relevant information. Second, what is required for a society that seeks to adapt effectively to changing circumstances is that a significantly large number of those who form the controlling part – the bureaucratic elite – must be made aware of the true ends of policy, namely, to promote the interests of the advantaged. The promulgation of an ideology alone is not enough to deal effectively with situations that, in principle, cannot be anticipated. If the bureaucratic elite are not informed, there is no reason to suppose that there will be an effective response to changing conditions. In particular, it will be an open question whether any such arrangement will be stable – for many decisions will be reached that impact negatively on the capacity of the ruling group to manage affairs effectively. However, if they are informed, this is likely to breed cynicism on the part of members of the privileged group, which will, in turn, engender forms of opportunism that are not directed at enhancing the position of the advantaged as a whole, but only, say, some group within the advantaged faction. In this way, privileged factions themselves are likely to break apart, and this will pose a real problem for such societies.

Third, one must be careful not to overestimate the power of those who are advantaged to control the dissemination of reliable information successfully and thus effectively mask their own objectives. Those who are disadvantaged do, after all, confront in their daily lives powerful reminders of the effects of discrimination against them, reminders that are likely to fuel deep resentment. One is bound to wonder, then, just how such an ideology could be successful unless it is substantially reinforced by the use of coercive power and at a level and scope that must prove enormously expensive in terms of basic resources.

In conclusion, these various considerations regarding ideology offer a substantial challenge to the notion that ideologically based systems of beliefs and values can function well from the perspective of the imperatives of adaptive efficiency. If ideological appeals can serve to reduce the conflict that might otherwise attend an indefensible distribution of benefits, and thereby resolve some of the problems that arise regarding allocative efficiency, such appeals will only exacerbate the problems of adaptive efficiency. What the advantaged hope to gain in the one regard is offset by the opportunity costs it involves.

By way of contrast, none of the considerations discussed pose any problem for a social order predicated on full cooperation. Those who are prepared to embrace the concept of a society based on full cooperation will respond positively to the three imperatives of adaptive efficiency. One can expect, then, that such a society will be much more adaptively efficient than one based on ideology.

15.16 A Final Observation

It is unfashionable in this postenlightenment age to speak of the ideal of progress toward a more rational ordering of relations between persons. In fact, however, economic approaches to the setting of priorities – the rational ordering of scarce resources to meet ordinary needs and desires, and the organization of human energies to pursue such mundane ends – has had an enormous impact on social and political relations in the modern age, first in the North Atlantic community and more recently in other parts of the world. This has led to the increasing erosion of the staying power of traditional ideologies. But with that erosion, the issue of justice as fairness becomes all the more pressing: Absent various rationalizations for inequalities, those who are disadvantaged are bound to ask the question why some do so very well while they do poorly. It is ironic (but then in its own way quite predictable) that the economic rationalization of relations has been accompanied by a sustained effort to marginalize considerations of fairness. The liberal credo in recent years has involved an increasing stress on the *deregulation* of economic activity and a call to reduce redistributive measures, but it does so in the face of the fact that the increasing acceptance of the "economic" approach creates the conditions under which these very issues cannot be so easily finessed.[33]

[33] Dasgupta (1993) provides a powerful argument along somewhat parallel lines about why the state that is associated with an economically rationalized society cannot avoid responsibility for a social minimum. The economic rationalization of relations between persons creates the very conditions under which traditional forms of social security, based on kinship, clan, and village organization of social relations cease to be operative.

References

Alchian, A. A. 1950. Uncertainty, evolution, and economic theory. *Journal of Political Economy* 58, 211–221.

Arrow, K. 1969. Political and economic evaluation of social effects and externalities. In *Frontiers of Quantitative Economics*, ed. M. D. Intriligator. North-Holland, Amsterdam, pp. 3–31.

Arthur, W. B. 1994. *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press, Ann Arbor.

Axelrod, R. 1984. *The Evolution of Cooperation*. Basic Books, New York.

Axelrod, R., and Dion, D. 1988. The further evolution of cooperation. *Science* 242, 1385–1390.

Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

Buchanan, J. M., Tollison, R. D., and Tullock, G. 1981. *Toward a Theory of the Rent-Seeking Society*. Texas A&M University Press, College Station.

Buchanan, J. M., and Tullock, G. 1962. *The Calculus of Consent*. University of Michigan Press, Ann Arbor.

Calvert, R. L. Rational actors, equilibrium, and social institutions. In *Explaining Social Institutions,* ed. J. Knight and I. Sened. University of Michigan Press, Ann Arbor.

Coase, R. H. 1937. The nature of the firm. *Economica* 4, 386–405.

Coase, R. H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Dasgupta, P. 1993.*The Origins of Well-Being and Destitution*. Clarendon Press, Oxford.

Fehr, E., Fischbacher, U., and Gachter, S. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* 13, 1–25.

Fudenberg, D., and Tirole, J. 1992. *Game Theory*. MIT Press, Cambridge, MA.

Gauthier, D. 1986. *Morals by Agreement*. Oxford University Press, Oxford.

Hardin, R. 1993. The street-level epistemology of trust. *Politics and Society* 21, 505–529.

Harsanyi, J. C. 1953. Cardinal utility in welfare economics and the theory of risk-taking. *Journal of Political Economy* 61, 434–435.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321.

Harsanyi, J. C. 1963. A simplified bargaining model for the *n*-person cooperative game. *International Economic Review* 4, 194–220.

Harsanyi, J. C. 1977a. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge.

Harsanyi, J. C. 1977b. Rule utilitarianism and decision theory. *Erkenntnis* 11, 25–53.

Harsanyi, J. C. 1978. Bayesian decision theory and utilitarian ethics. *American Economic Review, Papers and Proceedings* 68, 223–228.

Hayek, F. A. 1967. Notes on the evolution of systems of rules of conduct. In his *Studies in Philosophy, Politics, and Economics*. University of Chicago Press, Chicago.

Lewis, D. 1969. *Convention*. Harvard University Press, Cambridge, MA.

Kalai, E., and Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43, 513–518.

Kaysen, K. 1946–1947. A revolution in economic theory? *Review of Economic Studies* 14, 1–15.

Knight, J. 1992. *Institutions and Social Conflict*. Cambridge University Press, Cambridge.

Luce, R. D., and Raiffa, H. 1957. *Games and Decisions*. Wiley, New York.

McClennen, E. F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press, Cambridge.

McClennen, E. F. 1992. The theory of rationality for ideal games. *Philosophical Studies* 65, 193–215.

McClennen, E. F. 1997. Pragmatic rationality and rules. *Philosophy & Public Affairs* 26, 210–258.

Morris, C. 1999. What is this thing called reputation? *Business Ethics Quarterly* 9, 87–102.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–162.

Nash, J. 1953. Two-person cooperative games. *Econometrica* 21, 128–140.

Nelson, R. R. 1994. Evolutionary theorizing about economic change. In *Handbook of Economic Sociology*, ed. N. J. Smelzer, and R. Swedberg. Princeton University Press, Princeton, NJ, pp. 108–136.

North, D. C. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge.

Olson, M. 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. Yale University Press, New Haven.

Posner, R. A. 1986. *Economic Analysis of Law*. Little, Brown, Boston.

Putnam, R. D. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ.

Rawls, J. 1955. Two concepts of rules. *Philosophical Review* 64, 3–32.

Rawls, J. 1958. Justice as fairness. *Philosophical Review* 67, 164–194.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA.

Rawls, J. 1993. *Political Liberalism*. Columbia University Press, New York.

Roth, A. E. 1979. *Axiomatic Models of Bargaining*. Springer, Berlin.

Sabel, C. F. 1994. Learning by monitoring: The institutions of economic development. In *Handbook of Economic Sociology*, ed. by N. J. Smelzer, and R. Swedberg, Princeton: Princeton University Press, Princeton, NJ, pp. 137–165.

Schelling, T. C. 1960. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

Sugden, R. 1989. Spontaneous order. *Journal of Economic Perspectives* 3, 85–97.

Ullman-Margalit, E. 1978. *The Emergence of Norms*. Oxford University Press, Oxford.

Von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Zeuthen, F. 1930. *Problems of Monopoly and Economic Welfare*. Routledge and Kegan Paul, London.

PART FIVE

RIGHTS AND LIBERTIES

16

Republican Political Theory

Philip Pettit

Republican political theory takes its starting point from a long-established tradition of thinking about politics (Pocock, 1975). The republican tradition is associated with Cicero at the time of the Roman republic; with a number of writers, preeminently Machiavelli – "the divine Machiavel" of the *Discourses* – in the Renaissance Italian republics; with James Harrington, Algernon Sydney, and a host of lesser figures in and after the period of the English civil war and commonwealth; and with the many theorists of republic or commonwealth in eighteenth-century England, America, and France. These theorists – the commonwealthmen (Robbins, 1959) – were greatly influenced by John Locke and, later, the Baron de Montesquieu; indeed, they claimed Locke and Montesquieu, with good reason, as their own. They are well represented in documents like *Cato's Letters* (Trenchard and Gordon, 1971) and, on the American side of the Atlantic, the *Federalist Papers* (Madison, Hamilton, and Jay, 1987).

The commonwealthmen helped to shape habits of political reflex and thought that still survive today. Their distinctive refrain was that while the cause of freedom rests squarely with the law and the state – it is mainly thanks to the constitution under which they live that people enjoy freedom – still the authorities are also an inherent threat and people have to strive to "keep the bastards honest." The price of liberty is civic virtue, then, where that includes both a willingness to participate in government and a determination to exercise eternal vigilance in regard to the governors. The commonwealthmen tended to advocate the removal of the monarchy in America but in England most were content to see the king constitutionally fettered. England was "a nation," in Montesquieu's (1989, p. 70) unmistakeable reference, "where the republic hides under the form of monarchy" (Rahe, 1992, p. 524).

I find the republican tradition of thought a wonderful source of ideas and ideals, and in this essay, I hope to communicate why (see Pettit, 1997a).

I am not alone in finding this tradition inspirational. Historians like John Pocock (1975) and Quentin Skinner (1978, 1983, 1984) have not only made the republican way of thinking visible to us in the past couple of decades, they have also shown how it can give us a new perspective on contemporary politics. Skinner in particular has argued that it can give us a new understanding of freedom, and my own argument builds on this. Legal thinkers like Cass Sunstein (1990, 1993a, 1993b), however, have gone back to the republican tradition in its distinctively American incarnation in the late 1800s and have made a strong case for the claim that the tradition suggests a distinctive way of interpreting the U.S. Constitution and, more generally, that it gives us an insightful overview on the role of government. Criminologists and regulatory theorists like John Braithwaite, with whom I have actively collaborated (Braithwaite and Pettit, 1990), find in the republican tradition a set of compelling ideas for articulating both the demands that we should place on a regulatory system, say, the criminal justice system, and the expectations that we should hold out for how those demands can be best met (Ayres and Braithwaite, 1992). And these are just a few thinkers among many commentators who have begun to chart republican connections, and sometimes to draw actively on republican ideas, in recent years.[1]

My own approach to republican political theory is to give center stage to the notion of freedom that was shared among republican thinkers generally and to derive other republican claims from the commitment to this ideal. In this chapter, I will present the republican ideal of freedom in the first section and then try to illustrate, in the second, the way in which that ideal has significance for contemporary political thought.

16.1 The Republican Ideal of Freedom

16.1.1 The Constant Connection

Early in the last century Benjamin Constant (1988) delivered a famous lecture entitled "The Liberty of the Ancients and the Liberty of the Moderns." He depicted the liberty of the moderns, in the familiar negative or liberal fashion, as the absence of interference. I am free in this sense "to the degree to which no human being interferes with my activity" (Berlin, 1958, p. 7). He depicted the liberty of the ancients, however, as the liberty associated, ideally, with being a direct participant in a self-governing democracy. I am

[1] For example, Michelman (1986), Elkin (1987), Pagden (1987), Taylor (1989), Oldfield (1990), Fontana (1994), Hutton (1995), Blom (1995), Spitz (1995), Viroli (1995).

free in this sense, not through being uncontrolled by others, but through sharing with others the power to control all. The liberty of the ancients is the most prominent form of what Isaiah Berlin (1958) later called *positive freedom*.

The most important observation in introducing the republican conception of freedom is to recognize Constant's image of the liberty of the ancients as a caricature that served to hide the true republican way of thinking, only recently so prominent, from his contemporaries' eyes. Constant may not have been consciously propagandizing but what he achieved was to mesmerize later generations into thinking that the only feasible, perhaps the only sensible, notion of freedom was the liberal idea of freedom as noninterference. The liberty of the ancients is no match for freedom as noninterference – even if it is thought desirable, it must be judged to be unattainable – and the effect of setting up the two as the only relevant alternatives was to give victory, inevitably, to the liberal ideal.

The republican way of thinking about freedom, effectively suppressed by Constant, represents it as nondomination, not as direct democratic standing. And the difference between freedom as noninterference and freedom as nondomination is easily explained. Assume that one person dominates another to the extent that they have the capacity to interfere arbitrarily – to interfere on an arbitrary basis – in some or all of the other's choices (Pettit, 1996, 1997a). Where freedom as noninteference makes the absence of interference sufficient for freedom, freedom as nondomination requires the absence of a capacity on the part of anyone else – any individual or corporate agent – to interfere arbitrarily in their life or affairs. The difference between the two ways of conceiving of liberty may seem slight, but a little reflection will reveal hidden dimensions to the contrast.

16.1.2 Interference and Arbitrary Interference

The two conceptions of freedom both invoke the notion of interference, and we may begin our exploration of the contrast between the two ways of conceiving liberty with a comment on this. On almost all accounts, the intrusions that count as interference have to be intentional acts or at least acts for which the agent can be held responsible (Miller, 1990, p. 35). They have to be intentional or quasi-intentional. The reason for this stipulation is that freedom under most accounts is a condition defined in relation to other intentional agents, not a condition defined by reference to favors bestowed by nature, not a condition defined by how far a person escapes various brute, nonintentionally imposed limitations (see Spitz, 1995, pp. 382–383).

But the intrusions that constitute interference may be restricted to acts that make certain options impossible for the agent, or they may be extended to include acts that coerce or manipulate the agent in choosing between options. I shall assume that for both conceptions of freedom interference is to be understood in the broader fashion. Acts of interference will include any acts that worsen the agent's situation, or least worsen it significantly, either by reducing the alternatives available in choice or by raising the costs associated with some of the alternatives or by misleading the subject as to the options or costs in question, for example, through making a threat that one does not mean to carry out.

Freedom as nondomination differs from freedom as noninterference in invoking the notion, not just of interference, but of interference on an arbitrary basis. An act is perpetrated on an arbitrary basis, we can say, if it is subject just to the *arbitrium*, the decision, or judgment, of the agent; the agent was in a position to choose it or not choose it, at their pleasure. When we say that an act of interference is perpetrated on an arbitrary basis, then, we imply that like any arbitrary act it is chosen or not chosen at the agent's pleasure, and, in particular, because interference with others is involved, we imply that it is chosen or rejected without reference to the interests or the opinions of those affected. The choice is not forced to track what the interests of those others require according to their own judgments.[2]

Under this conception of arbitrariness, then, an act of interference will be nonarbitrary to the extent that it is the contrary of the arbitrary act, to the extent that it stands at the opposite extreme. The nonarbitrary act of interference is not subject, as the arbitrary act is subject, to the *arbitrium* of the interferer. On the contrary, it is subject, as we might put it, to the *arbitrium* of the interferee. The nonarbitrary act is forced to track the interests and ideas of the person suffering the interference.

Or if it is not forced to track all of the interests and ideas of the person involved, it is at least forced to track the relevant ones. I may have an interest in the state's imposing certain taxes or in punishing certain offenders, for example, and the state may pursue these ends according to procedures

[2] Notice that an act of interference can be arbitrary in the procedural sense intended here – it may occur on an arbitrary basis – without being arbitrary in the substantive sense of actually going against the interests or judgments of the persons affected. An act is arbitrary, in this usage, by virtue of the controls – specifically, the lack of controls – under which it materializes, not by virtue of the particular consequences to which it gives rise. The usage I follow means that there is no equivocation involved in speaking, as I do speak, either of a power of arbitrary interference or of an arbitrary power of interference. What is in question in each case is a power of interfering on an arbitrary, unchecked basis.

that conform to my ideas about appropriate means. But I may still not want the state to impose taxes on me – I may want to be an exception – or I may think that I ought not to be punished in the appropriate manner, even though I have been convicted of an offense. In such a case, my relevant interests and ideas will be those that are shared in common with others, not those that treat me as exceptional because the state is meant to serve others as well as me. They will be interests that I can avow and assert politically, consistently with wanting to live under a shared political arrangement with others. So in these cases, the interference of the state in taxing or punishing me will not be conducted arbitrarily and will not represent domination.

The republican tradition of thinking took a distinctive view of what is required for an act of interference, in particular, an act of legal or government interference, to be nonarbitrary, and I follow that tradition in giving this account of nonarbitrariness. Consider Tom Paine's (1989, p. 168) complaint against monarchy. "It means arbitrary power in an individual person; in the exercise of which, *himself*, and not the *res-publica*, is the object" (cf. Sydney, 1996, pp. 199–200). What is required for nonarbitrary state power, as this comment makes clear, is that the power be exercised in a way that tracks, not the power-holder's personal welfare or worldview but rather the welfare and worldview of the public. The acts of interference perpetrated by the state must be triggered by the shared interests of those affected under an interpretation of what those interests require that is shared, at least at the procedural level, by those affected.

Where freedom as noninterference opposes freedom directly to interference – freedom just is noninterference – the second varies this opposition in two ways. The antonym of freedom no longer involves interference as such, only interference on an arbitrary basis. The antonym of freedom does not require actual arbitrary interference, only vulnerability to someone with the capacity for such interference.

The first variation has the effect of making it harder for people to lose their freedom or to have their freedom reduced. For it means that, if an agent interferes nonarbitrarily in their choices, that does not offend as such against their freedom; whatever damage is done by the interference, the nonarbitrariness is enough to ensure that their freedom is not compromised. But the second variation has the contrary effect of making it easier, not harder, for someone to suffer a loss of freedom. For it means that, if an agent has the capacity to interfere in any of their choices, then that in itself compromises their freedom; they suffer a loss of freedom even if the other person does not actually exercise their capacity for interference.

16.1.3 The Harder-to-Lose-Freedom Effect

The harder-to-lose-freedom effect makes for a difference in how law bears on liberty under the two conceptions. Under freedom as noninterference, a regime of law, being necessarily coercive, systematically compromises people's freedom, even if the consequence of putting the regime into operation is that less interference takes place overall. Subjection to the law, in and of itself, represents a loss of liberty. Under the second conception, however, subjection to the law need not represent a loss of liberty for anyone who lives under it, provided – and of course it is a big proviso – that the making, interpretation, and implementation of the law is not arbitrary: provided that the legal coercion involved is constrained to track the interests and ideas of those affected. The proviso, intuitively expressed, is that the legal regime represents a fair rule of law.

Consistently with not itself constituting a compromise of liberty, of course, a regime of legal coercion and restraint may have the same effect as a natural obstacle in limiting the choices available to people or in making them more costly: in defining the range over which people enjoy undominated choice. Proponents of freedom as noninteference do not count natural obstacles as factors that compromise liberty – this, because they are in no way intentional – but they do admit that such obstacles affect the range of choice over which freedom as noninterference may be exercised; the obstacles condition freedom, as we might put the distinction, but they do not compromise it.[3] Proponents of freedom as nondomination move the locus of this boundary between compromising and conditioning factors so that the interference associated with a fair rule of law, like the natural obstacle, conditions people's liberty but does not compromise it: does not in itself count as infringing, violating, reducing, or offending against people's liberty.[4]

[3] When proponents of this ideal speak of making freedom as noninterference effective, not just leaving it as a formal freedom, I assume that they often have in mind removing or reducing the obstacles that condition the exercise of freedom as nondomination: extending the range of choice available to people. See Van Parijs (1995) on "real" or "effective" freedom.

[4] The extreme case of legal interference is punishment for an offence. Such punishment will always condition people's freedom as nondomination, removing the capacity for undominated choice (capital punishment), restricting the range over which such choice may be exercised (prison), or raising the costs of making certain undominated choices (fines). But it need not involve the person punished having their liberty compromised through subjection to the arbitrary will of another. This remark is not meant to make legal punishment seem any more tolerable, only to articulate a perhaps surprising corollary of the conception of freedom as nondomination.

Hobbes and Bentham are the great advocates of the idea that law represents a compromise of liberty. "As against the coercion applicable by individual to individual, no liberty can be given to one man but in proportion as it is taken away from another. All coercive laws, therefore, and in particular all laws creative of liberty, are as far as they go abrogative of liberty" (Bentham, 1843). Or as Hobbes had put it: "The Liberty of a Subject, lyeth therefore only in those things, which in regulating their actions, the Soveraign hath praetermitted" (Hobbes, 1968, p. 264).

But Hobbes and Bentham were consciously breaking with a longer tradition of thought – the republican or commonwealthman tradition – in taking this line (Skinner, 1983). That tradition was defended in the first instance by James Harrington (1992, p. 20), who argued that Hobbes was confusing freedom from the law with freedom proper: freedom by the law. John Locke took Harrington's side, embracing "freedom from Absolute, Arbitrary Power" as the essential thing (Locke, 1965, p. 325) and presenting law as essentially on liberty's side: "that ill deserves the Name of Confinement which serves to hedge us in only from Bogs and Precipices . . . the end of Law is not to abolish or restrain, but to preserve and enlarge Freedom" (Locke, 1965, p. 348). William Blackstone (1978, p. 126) represents the eighteenth-century orthodoxy when he follows the same line: "laws, when prudently framed, are by no means subversive but rather introductive of liberty; for (as Mr Locke has well observed) where there is no law there is no freedom."

The difference between the two conceptions of liberty in their attitude to the law was significant from the point of view of Hobbes and Bentham. The view that all law compromises people's liberty enabled Hobbes to withstand the criticism that he anticipated from republicans, that his Leviathan was utterly inimical to freedom, constituting an arbitrary rule as distinct from a rule of law: an arbitrary rule as distinct from the republican vision of an "empire of laws, and not of men" (Harrington, 1992, p. 8). And the same view enabled Bentham and those friends of his who opposed the American cause in the 1770s to argue against the complaint that since the British parliament was not constrained in the laws that it passed for the governance of the American colonies – because it was not constrained in the same way that it was constrained in Britain itself – those laws represented an arbitrary interference with Americans and compromised their liberty (Lind, 1776). Hobbes could argue that Leviathan did no worse than commonwealths in respect of the liberty of its subjects because all law compromises liberty. And Bentham and his friends could argue on the same grounds that in regard to liberty Americans fared no worse under the law imposed by the British parliament than those in Britain.

So much for the harder-to-lose-freedom effect of opposing freedom to nondomination, not noninterference. But what of the easier-to-lose-freedom effect of shifting the antonym?

16.1.4 The Easier-to-Lose-Freedom Effect

This effect occurs because someone loses freedom, not just to the extent that another person interferes on an arbitrary basis in their choices, but to the extent that another agent has the capacity to do this. With freedom as nondomination, someone loses freedom to the extent that they live under the thumb of another, even if that thumb is never used against them. Suppose that under the existing laws and mores a wife may be abused on an arbitrary basis by her husband, at least in certain areas and in a certain measure. Even if her husband is a loving and caring individual, such a wife cannot count as fully free under the construal of freedom as nondomination. And neither can the employee who lives under the thumb of an employer, nor the member of a minority who lives under the thumb of a majority coalition, nor the debtor who lives under the thumb of a creditor, nor anyone in such a subservient position.

Where the first effect of shifting antonym shows up particularly in the assessment of law and liberty, the second relates to the association between law and slavery. As it became a matter of common assumption after Bentham that law represents a compromise of liberty, albeit a compromise that may be for the good overall, so it became impossible to maintain that to be unfree is always, in some measure, to be enslaved (Patterson, 1991); no one was prepared to say that the law makes slaves of those who live under it. But before Bentham, when freedom was opposed first and foremost to domination, the association between unfreedom and slavery was complete. To be unfree was to live at the mercy of another, to live under a condition of enslavement to them.

Thus, Algernon Sydney (1990, p. 17) could write in the 1680s: “Liberty solely consists in an independency upon the will of another, and by the name of slave we understand a man, who can neither dispose of his person nor goods, but enjoys all at the will of his master.” And in the following century, the authors of *Cato’s Letters* could give a characteristically forceful statement to the theme. “Liberty is, to live upon one’s own Terms; Slavery is, to live at the mere Mercy of another; and a Life of Slavery is, to those who can bear it, a continual State of Uncertainty and Wretchedness, often an Apprehension of Violence, often the lingering Dread of a violent Death” (Trenchard and Gordon, vol. 2, pp. 249–250).

The easier-to-lose-freedom effect of opposing liberty to domination connects with the slavery theme because one of the striking things about slaves is that they remain slaves even if their master is entirely benign and never interferes with them. As Algernon Sydney (1990, p. 441) put it, "He is a slave who serves the best and gentlest man in the world, as well as he who serves the worst." Or as it was put by Richard Price (1991, pp. 77–78) in the eighteenth century: "Individuals in private life, while held under the power of masters, cannot be denominated free, however equitably and kindly they may be treated. This is strictly true of communities as well as of individuals." There is domination, and there is unfreedom, even if no actual interference occurs.

I mentioned that the first effect of opposing freedom to domination helped the defenders of the American cause to argue that while those in Britain were not made unfree by the law, given that the law could not be arbitrarily imposed there, those in America did not enjoy a similar status under the law. I should add that the second effect enabled them to sheet this argument home. They were in a position to argue that even though the British parliament did not interfere much in American affairs – even though it only levied a small tax – still the fact that it could levy whatever tax it wished, without any serious restraint on its will, meant that it related to the American colonists as master to slave.

Joseph Priestly (1993, p. 140) offers a nice example of this line of argument.

> Q. What is the great grievance that those people complain of? A. It is their being taxed by the parliament of Great Britain, the members of which are so far from taxing themselves, that they ease themselves at the same time. If this measure takes place, the colonists will be reduced to a state of as complete servitude, as any people of which there is an account in history. For by the same power, by which the people of England can compel them to pay *one penny*, they may compel them to pay the *last penny* they have. There will be nothing but arbitrary imposition on the one side, and humble petition on the other.

16.1.5 Three Further Remarks

My comments on the two main differences associated with treating freedom as nondomination rather than noninterference should serve to make the notion intelligible. I want to add three further remarks, however, to underline some points that are important for understanding it fully.

First, while freedom as nondomination is constituted by one agent's having the capacity to interfere on an arbitrary basis in the affairs of another, some plausible empirical assumptions entail that it is going to be systematically associated with a shared awareness on the part of the individuals or groups involved that this capacity exists. The question of whether you are undominated is bound to be of interest to anyone, and the facts that make you undominated, if indeed you are such – the facts about your comparative resources, for example, and about the degree to which you are protected by legal and other means – are bound to be salient to all involved. Under standard assumptions as to people's inductive and inferential abilities, it follows that the fact of nondomination is going to be a matter of common recognition among the individuals in question (Lewis, 1969, p. 56). And that is something of the greatest significance. For it means that under standard ways of achieving it, freedom as nondomination is going to be intimately linked with the ability to look others in the eye, without having to defer to them or fear them. Montesquieu (1989, p. 157) emphasizes this theme when he writes: "Political liberty in a citizen is that tranquillity of spirit which comes from the opinion each one has of his security, and in order for him to have this liberty the government must be such that one citizen cannot fear another citizen."

Second, if someone is to enjoy freedom as nondomination, then it is not enough that the other people are very unlikely to exercise arbitrary interference; those other people must lack the capacity to interfere arbitrarily in that person's life, not just be unlikely to interfere. Suppose that you are subject to interference on an arbitrary basis from someone who, as it happens, really likes you and is extremely unlikely to want to interfere. If it still remains the case that, by the ordinary standards of free-will attribution, they have a capacity to interfere or not to interfere, and this on a more or less arbitrary basis, then you are dominated in some measure by them and are thus unfree. This is not a very hard line to take because you clearly suffer an evil to the extent that the person has the capacity to interfere arbitrarily with you: to the extent that such interference is accessible to them as an agent, however improbable it is that they will exercise it. Their capacity for arbitrary interference means, for example, that you lack grounds for the subjective state of mind that goes with freedom as nondomination; you have reason to defer to the person in question and to look for their continued favor.

My last point is the most important. When Bentham and his associates came to reject the notion of freedom as nondomination, freedom as nonslavery, one theme in their reflections was that this sort of freedom did not come in degrees and so, unlike the rival conception, lent itself to "panegyric and careless declamation" (Paley, 1825, pp. 359–360; Long, 1977, ch. 4). John Lind (1776, p. 25) expressed the criticism strongly in his attack on Richard Price's talk of the American colonists as slaves. "Things must be always at the maximum or minimum; there are no intermediate gradations: what is not white must be black." The third point I want to make is that this perception is mistaken. Freedom as nondomination is not an all-or-nothing matter.

The point should be obvious on a little reflection. Agents may have a more or less ready capacity to interfere. And the interference for which they have a capacity may be more or less serious and may be available more or less without cost, say, without risk of retaliation. Thus, the freedom as nondomination of those they are in a position to affect may be more or less intense; the weaker the agents, the greater the freedom of those they may affect.

Intensity, I should add, is only one dimension in which freedom as nondomination may vary. As it is more or less intense, so freedom as nondomination may also be of one or another extent. It may be available for a smaller or larger number of choices, for choices that are more or less costly, and for choices of intuitively lesser or greater significance. Even if we have attained the highest possible intensity of nondomination for people in a society, there may be room for improving the range of undominated choice that is available to them: we may make the range of choice larger or less costly or intuitively more significant. Even if we remove all compromising influences on freedom as nondomination, there may still be room for how far we can remove conditioning influences as well.

That freedom as nondomination may be increased in either of two broad dimensions makes for a problem in deciding how those dimensions are to be weighed against one another (Pettit, 1997a, ch. 3). Indeed, a similar problem arises with freedom as noninterference because this will be increased in intensity as far as interference is blocked and increased in extent as far as the range of unobstructed choice is expanded, say, by providing people with extra resources. But I can overlook such problems of weighting here, as I shall be concerned with the promotion of these values only in the dimension of intensity. The question I address in the next section bears only on what is required for maximizing equal nondomination in the dimension of intensity.

16.2 The Significance of the Republican Ideal

16.2.1 The Paley Connection

Perhaps the most important figure in the demise of the republican ideal – someone more important even than Constant – is William Paley. Paley may have been the only writer in his time to recognize clearly the shift that was taking place, the shift indeed for which he argued, from the received notion of freedom as nondomination, freedom as security against interference on an arbitrary basis, to freedom as noninterference. He sets out his view with admirable clarity in *The Principles of Moral and Political Philosophy*, which was first published in 1785 and was continually reprinted throughout the nineteenth century (Paley, 1825).

Paley recognizes in this work that the usual notion of civil liberty, the one that agrees with "the usage of common discourse, as well as the example of many respectable writers" (p. 357), is that of freedom as nondomination. "This idea places liberty in security; making it to consist not merely in an actual exemption from the constraint of useless and noxious laws and acts of dominion, but in being free from the *danger* of having such hereafter imposed or exercised" (p. 357; original emphasis). But Paley argues against this received notion and in favor of a Benthamite version of freedom as noninterference on an extraordinary basis. He argues that the ideal in question, however well established, is excessively demanding on the state:

> Those definitions of liberty ought to be rejected, which, by making that essential to civil freedom which is unattainable in experience, inflame expectations that can never be gratified, and disturb the public content with complaints, which no wisdom or benevolence of government can remove. (Paley, 1825, p. 359)

How could Paley have thought that freedom as nondomination was the received ideal of freedom and yet that it was too demanding on the state? My hunch is that for Paley, as for most progressive thinkers of the late eighteenth century, it was no longer possible to think that the political citizenship and consideration could be restricted, as traditional republicans had taken it to be restricted, to propertied males: women and servants could not be systematically and permanently excluded from concern. "Everybody to count for one, nobody for more than one," is the slogan ascribed to Bentham by John Stuart Mill (1969, p. 257). Thus, whereas traditional republicans could think that everyone relevant to the state's concern – every propertied male – might aspire to be free in the sense of not being subject to anyone's domination, egalitarians like Paley could not say this without seeming to embrace a wholly revolutionary doctrine: a doctrine that would require the upturning of relations between men and women, masters and servants. Their response was to deflate the ideal of freedom – to reduce it from nondomination to noninterference – at the same time that they argued that the constituency of political concern should be expanded. What they gave with the one hand, they took away with the other.

Why is the republican conception of liberty politically significant for the modern state? In a word, because it would recall the state to performing in relation to citizens generally the service that a republic – even a republic hidden under the form of a monarchy – was expected to perform for traditional elites. It may indeed have been impossible for someone like Paley or Bentham or Constant to envisage a state that would liberate servants as well as masters, women as well as men. But this is no longer an obviously infeasible ideal, even if it is obviously unattained. For the limits on what we can envisage the state doing, and the limits on what we can imagine civil society allowing the state to do, have shifted dramatically over the last couple of centuries or so. Republicanism went underground at the time when the state began to become inclusivist, thereby permitting the state to become simultaneously more or less minimalist. It is high time that the doctrine was restored to prominence, allowing us to consider the direction that an inclusive republic – a republic dedicated to the general promotion of freedom as nondomination – would have to take.

I have tried to display the significance of the republican perspective elsewhere, examining the impact of the republican ideal on our notions of equality and community; on the policy commitments that we prescribe for the modern state; on the way we conceive of constitutional and democratic values and institutions; on the approach that we take to issues of regulation and control; and on the image we have of how the state should relate to civil society (Pettit, 1997a). Together with John Braithwaite, I have also looked at the significance of republicanism for thinking about criminal justice (Braithwaite and Pettit, 1990; Pettit, 1997b). To illustrate the way in which the republican perspective can affect our thinking, I will concentrate here on its significance for how we think of issues of redistribution. This theme is particularly relevant because it is at the center of contemporary political discussions, and it also connects up with the hostile reaction of Paley and others like him to the republican ideal.

16.2.2 Redistribution and Freedom as Noninterference

How far is the maximal equal distribution of freedom as noninterference consistent with inequalities in other dimensions? How far is it consistent, for example, with different levels of provision in basic goods like food and shelter, modes of transport, and media of reliable information; in basic services like medical care, legal counsel, and accident insurance; in human capital of the kind associated with training and education; in social capital of the sort that consists in being able to call with confidence on others; in political capital such as office and authority confer; and in the material capital that is necessary for production? How far is it likely to require putting inequalities in these matters right or at least alleviating their effects: in particular, coercively putting them right or coercively alleviating their effects under state initiatives? How far is it likely to require what I shall describe, in a word, as redistribution?

The common wisdom on this question is that the maximal equal distribution of freedom as nonintefeference would leave a lot to be desired in regard to redistribution: it would fall short, under most conceptions, of achieving distributive justice (Rawls, 1971). That wisdom is well placed, and I wish to argue that in this respect freedom as nondomination represents a sharply contrasted ideal: the maximal equal distribution of such freedom requires a much more substantial commitment to redistribution.

Before coming to that argument, however, it will be useful to see why the connection between freedom as noninterference and distributive justice is so loose. Two questions arise from the viewpoint of freedom as noninterference when any such issue of redistribution is considered. First, how far will redistribution entail interference in people's lives by the state? Second, how far will redistribution lower the probability of interference by other agents?

The answer to the first question is that redistribution will always entail a degree of interference by the state. For even the most basic form of redistribution involves taxing some to give to others and that constitutes interference; it deprives those who are taxed of a choice in how to use their money. That taxing issue aside, most forms of redistribution also require setting up inspectors and other officials to oversee the operation in question. Thus, the redistributive measures involve the creation of new possibilities of interference in people's lives.

The answer to the first question means that the onus of proof always lies, from the perspective of freedom as noninterference, with those who would counsel redistribution. Whether redistribution in any area is to be supported, then, depends on whether the answer to the second question shows clearly that the margin whereby redistribution will reduce interference in a society is greater than the margin whereby it introduces interference itself. The margin of projected improvement will have to be large enough to ensure that even when we discount for the less-than-certain nature of the projection, the argument squarely favors redistribution.

Finding grounds to defend the required answer to the second question is never going to be easy. The reason is that it is always going to be possible for the opponent to argue that as long as we do not think of the relatively advantaged as downright malicious, we must expect them not to be generally disposed to harm the disadvantaged and not to be generally in need of curtailment by the redistributive state. Perhaps employers are in a position under the status quo to interfere in various ways with their employees. But why expect them to interfere rather than striving for good and productive relationships? Perhaps husbands are able, given their greater strength and greater cultural backing, to abuse their wives. But why expect them to practice such abuse rather than remaining faithful to their affections and commitments? Perhaps those who lack medical care and legal counsel are prey to the unscrupulous. But why expect doctors and lawyers to be unwilling to provide essential services pro bono, especially when they can make good publicity of providing such services?

I sympathize with the drift of these rhetorical questions, believing that it is a mistake to demonize the relatively advantaged and see them always as potential offenders (Pettit, 1995). But the effect of the questions in the context of endorsing an ideal of freedom as noninterference is what concerns me now, not the propriety of raising them. The effect is to lead those who take the ideal as the only relevant yardstick of social performance not to require much in the way of redistribution: not to require much in the way of what we intuitively describe as distributive justice. It is quite possible to believe that the regime under which freedom as noninterference is equally distributed at maximal levels is a regime that allows great inequalities in other regards.

It is because of the looseness of the connection between freedom as noninterference and redistribution that thinkers in the broad liberal tradition tend to divide, roughly, into left and right. What unites those thinkers is the belief that freedom consists in noninteference and that equal freedom for all is the primary political value; this belief lies behind Rawls's (1971) first principle of justice, for example, according to which society should look for an equal system of maximal liberties. But what divides such liberals is that, whereas those on the right – libertarians, in particular – think that this is all that the state should try to achieve, those on the left shrink from such a minimalist vision. They argue that the state should concern itself with something over and beyond liberty proper: with the relief of poverty, for example – this may be held to make liberty effective (Van Parijs, 1995) – or with the achievement of something close to equality. Rawls's second principle of justice represents the ideal of something that is close to equality in this sense; the principle tolerates inequality but only to the extent that the worst-off person in the society benefits indirectly from the existence of that inequality.

16.2.3 Redistribution and Freedom as Nondomination

We can begin to recognize the significance of the republican ideal of freedom when we notice the difference between how it connects with redistribution and how freedom as noninterference does so. We have seen that the project of equalizing freedom as noninterference at the maximal possible level is hostile to redistribution in two ways. First, it introduces a presumption against redistribution; it casts the onus on the side of anyone who wants to argue for redistribution. And second, it ensures that any argument for redistribution will have to be probabilistic in a manner that is bound to make it easy to resist. I wish to argue that the ideal of maximizing freedom as nondomination at the maximal level possible differs from the associated ideal of freedom as noninterference in both these respects.

Freedom as noninterference introduces a presumption against redistribution because redistribution is itself a species of the evil of interference. But no corresponding argument is available with freedom as nondomination. If the redistributive measures adopted can be pursued and are pursued under what I described intuitively as a fair rule of law, then they do not themselves introduce any form of domination.

I assume that many of the redistributive measures contemplated in discussions of distributive justice can be pursued under a fair rule of law. I assume, that is, that the measures can be introduced under procedures designed to filter out discriminatory proposals – proposals that fail to reflect the politically avowable interests or ideas of some sector of the population – and to guarantee nonarbitrariness. If the assumption is sound, then the ideal of freedom as nondomination does not introduce any presumption against redistribution of the kind associated with freedom as noninterference. If redistributive measures are used in the promotion of nondomination, the good at which they are directed does not have to be balanced against a violation of that very good in the process of production; the process of production need not itself represent a form of domination.

Is the assumption about the nonarbitrariness of redistribution likely to be sound? Any redistribution is going to deprive the rich of resources they would otherwise have but it need not represent an arbitrary form of interference. The republican state is built upon the premise that freedom as nondomination is the primary goal of the state. And if it is a matter of general assumption that the state should do whatever is needed to ensure such freedom in the community, then a transfer of resources that is essential to that goal can be justified to the rich on the basis of an interest they share. Certainly, the rich have an interest in keeping their own resources but pressing that interest against the clear demand envisaged will constitute special pleading; it will be akin to looking for special treatment by the criminal law.

The process of redistribution will not be entirely innocent, of course. As we mentioned, any rule of law, and certainly any redistributive rule of law, is going to remove certain choices or raise the costs of pursuing them. But this way of restricting choice, this way of conditioning people's freedom as nondomination, falls far short of compromising such freedom on their part. If it succeeds in reducing the extent to which the freedom of the poor or the sick or the needy have their freedom compromised, then this cost in the conditioning of the freedom of people generally is going to be well worth paying.

Here is another way of thinking about the point. Redistribution under a fair rule of law counts in the republican ledger as a form of conditioning of liberty on a par with the conditioning effected by natural factors like poverty, disability, or illness. Redistribution involves something akin to moving around the factors that serve as conditioning influences on freedom and this without dominating anyone or without compromising anyone's freedom as nondomination. If that reshuffling of freedom-relevant factors can itself increase the degree of equal freedom in the society, then there is little or no question to raise about it. There is no reason to have a presumption against it.

But we should not be complacent about the dangers of redistribution. Complacency will be justified only up to a certain level and only under a certain kind of redistribution by a state. Suppose that the redistribution allowed involves the exercise of unconstrained discretion by individual agents of the state; the discretion may arise in the way goods are taken from some, for example, or in the way goods are given to others. Or suppose that the redistribution is so extensive, or subject to such frequent adjustments, that people hardly know where they stand relative to the state. Under any such suppositions, the prospect of redistribution is going to look very unattractive from a republican point of view.

The republican tradition of thinking has always put the state under severe scrutiny, for fear that state authorities will ever become, or ever support, relatively arbitrary powers. In arguing that the ideal of freedom as nondomination is not hostile to redistribution, in particular not hostile in the manner of freedom as noninterference, I do not mean to reject that tradition. If we treasure freedom as nondomination, then we have to be vigilant about not allowing the state certain sorts of power; we have to be careful to see that it is subject to all sorts of constitutional and other constraints. My point has been only that provided a state can be sufficiently constrained – and that may be a very big proviso – there is nothing inherently objectionable about allowing it to use redistributive means for promoting freedom as nondomination.[5]

[5] Libertarians often say that they are against big government. Republicans are also against big goverment, but in a different sense. They object, not necessarily to government's having redistributive rights and responsibilities, but rather to the government's being able to act arbitrarily in the pursuit of redistributive ends; the pursuit must always be governed by a fair rule of law.

The second point that we noticed about the redistributive significance of equalizing noninterference was that the question of whether any redistributive measure increased people's freedom as noninterference remained inevitably a probabilistic matter. Perhaps we can interfere with employers to ensure that they do not interfere in certain ways with their employees. Perhaps we can interfere with husbands to ensure that they do not interfere in certain ways with their wives. But before we think of practizing interference, we have to convince ourselves that a very shaky arithmetic comes out right. We have to convince ourselves that there is a suitably high probablility of a suitably large reduction in the practice of interference by employers and husbands. That thought may well give pause to any projects of redistribution that the ideal of freedom as noninterference is otherwise likely to sponsor.

But as on the matter of the presumption against redistribution, the ideal of freedom as nondomination has quite a different impact here. Suppose that an employer has the capacity in some measure to interfere arbitrarily in the affairs of an employee. Employment is so scarce and the prospect of unemployment so repellent, that the employer can alter agreed conditions of work, make life much tougher for employees, or even practice some illegal interference in their affairs, with relative ease. And suppose now that we contemplate introducing a system of unemployment benefits, a set of health and safety regulations, or an arrangement for arbitrating workplace disputes that would improve the lot of employees. Do we have to do a range of probabilistic sums before we can be sure of the benefits of such a redistributive regime?

Assuming that the regime is consistent with a fair rule of law and does not introduce an independent source of domination – provided it does not have any dominational side effects – it should be clear that no such sums are necessary. For just the existence of reasonable unemployment benefits is bound to reduce the extent to which an employee is willing to tolerate arbitrary interference by an employer and is bound by the same token to reduce the capacity of the employer to interfere at will and with impunity in the lives of employees. There is no uncertainty plaguing the connection. Or at least there is no uncertainty of the kind that makes the connection with freedom as noninterference so problematic.

Similar points go through on a number of fronts. The fact that people are poor, illiterate, ignorant, unable to get legal counsel, uninsured against illness, or incapable of getting around – the fact that they lack basic capabilities in any of these regards (Sen, 1985) – makes them subject to a certain sort of exploitation and manipulation. Other things being equal, then, any improvement in their lot is bound to reduce the capacity of others to interfere more or less arbitarily in their lives. And that means that other things being equal – dominational side effects being absent – any such improvement is bound to increase their freedom as nondomination.

The crucial difference in this second respect between the ideals of freedom as noninterference and freedom as nondomination occurs because the first ideal is compromised only by actual interference and the second by the capacity for interference, in particular the capacity for arbitrary interference. It may be very unclear whether a given measure will actually reduce the overall level of interference practiced by the more advantaged, while it is absolutely certain that the measure will reduce their capacity for interference.

Suppose the employer in our earlier example is actually benign or committed to a smooth and productive workplace, and that this ensures a negligible probability of ever interfering in the affairs of employees. The introduction of employment benefits, health and safety regulations, or arbitration procedures will not significantly reduce the probability of interference in such a scenario; that probability is already negligible. Still the introduction of any such scheme will certainly reduce the employer's capacity for arbitrary interference. Whether the employer interferes will no longer be dependent on their good grace; it will be substantively determined by factors outside the employer's will.

Some will retort at this point that there is no reason we should want to reduce the capacity of an employer to interfere with employees, especially given the cost of doing so, when it is certain that no interference will actually occur. But that shifts the issue from what the ideal of freedom as nondomination would require – and, in particular, from the observation that it would require, other things being equal, that the employer is constrained – to whether it is an attractive ideal. My aim is not to argue that it is an attractive ideal (on this issue, see Pettit, 1997a), only that it is a redistributively demanding one.

Still, I cannot resist adding a few words to explain the republican point of view. The situation in which another person has the capacity to interfere on an arbitrary basis with me but is very unlikely to do so, perhaps because of a morbid desire to be liked, will leave me relatively happy as far as I can regard that individual as a probabilistic device. But it is of the essence of human relations that we do not regard one another in this way: that we take one another to be responsible and free agents for whom the most improbable, out-of-character options remain accessible to choice (Strawson, 1973; Pettit and Smith, 1996). Thus, in dealing with the imagined person, it will be a matter of common expectation between us that I will do nothing to alienate and as much as I can to placate them. If I choose to discount their greater power, acting as if I know they will not strike against me, then that will constitute an act of defiance: an act that may very well transform their psychology and change them into a dangerous presence. In view of such considerations, it should be unsurprising that republicans denounce any situation where a person lives in the power of a lord – *in potestate domini* – even in the power of a lord who is judged unlikely to cause them any harm.

We saw earlier that freedom as noninterference may be maximized under the constraint of more or less equal allocation without any significant redistribution of resources being required. In this respect, as in so many others, freedom as nondomination is quite different. The republican ideal may be capable of encoding the redistributive measures that many of us would think it reasonable to require of the modern state. While remaining an ideal of liberty, this ideal may give adequate expression to the more demanding aspirations that the nonlibertarians among us find compelling.[6]

References

Ayres, I., and Braithwaite, J. 1992. *Responsive Regulation.* Oxford University Press, New York.

Bentham, J. 1843. Anarchical fallacies. In *The Works of Jeremy Bentham*, Vol. 2, ed. J. Bowring. W. Tait, Edinburgh.

Berlin, I. 1958. *Two Concepts of Liberty*. Oxford University Press, Oxford.

Blackstone, W. 1978. *Commentaries on the Laws of England*, 9th ed. Garland, New York. (facsimile of 1783 edition.)

Blom, H. W. 1995. *Causality and Morality in Politics: The Rise of Naturalism in Dutch Seventeenth-Century Political Thought.* CIP-Gegevens Koninklijke Bibliotheek, The Hague.

Braithwaite, J., and Pettit, P. 1990. *Not Just Deserts: A Republican Theory of Criminal Justice*. Oxford University Press, Oxford.

Constant, B. 1988. *Constant: Political Writings*, ed. B. Fontana. Cambridge University Press, Cambridge.

Dworkin, R. 1978. *Taking Rights Seriously.* Duckworth, London.

[6] This paper is reprinted, with some small changes, from my article with the same title in Andrew Vincent, ed., *Political Theory: Tradition and Diversity*, Cambridge University Press, Cambridge, 1997. Copyright © Cambridge University Press 1997. Reprinted with permission. For recent, somewhat reformulated statements of my point of view see Pettit (2006, 2007). I am most grateful to Geoffrey Brennan and Michael Smith for helpful discussions of the material and to Andrew Vincent and John Weymark, as well as to an anonymous referee, for suggestions on how to improve it. I was enormously helped by comments received when an earlier version of the paper was presented to a meeting in New Orleans organized by the Murphy Institute of Political Economy, Tulane University, and the International Society for Economics and Philosophy.

Elkin, S. L. 1987. *City and Regime in the American Republic.* University of Chicago Press, Chicago.

Fontana, B., ed., 1994. *The Invention of the Modern Republic.* Cambridge University Press, Cambridge.

Harrington, J. 1992. *The Commonwealth of Oceana and A System of Politics*, ed. J. G. A. Pocock. Cambridge University Press, Cambridge.

Hobbes, T. 1968. *Leviathan*, ed. C. B. MacPherson. Penguin Books, Harmondsworth.

Hutton, W. 1995. *The State We're In.* Cape, London.

Lewis, D. 1969. *Convention.* Harvard University Press, Cambridge, MA.

Lind, J. 1776. *Three Letters to Dr Price.* T. Payne, London.

Locke, J. 1965. *Two Treatises of Government*, ed. P. Laslett. Mentor, New York.

Long, D. C. 1977. *Bentham on Liberty.* University of Toronto Press, Toronto.

Madison, J., Hamilton, A., and Jay, J. 1987. *The Federalist Papers*, ed. I. Kramnik. Penguin, Harmondsworth.

Michelman, F. 1986. The Supreme Court 1985 term. *Harvard Law Review* 100, 4–77.

Mill, J. S. 1969. *Essays on Ethics, Religion and Society.* Collected Works, Vol. 10. Routledge, London.

Miller, D. 1990. *Market, State, and Community.* Oxford University Press, Oxford.

Montesquieu, C. de Secondat. 1989. *The Spirit of the Laws.* tran. and ed. A. M. Cohler, B. C. Miller, and H. S. Stone. Cambridge University Press, Cambridge.

Oldfield, A. 1990. *Citizenship and Community: Civic Republicanism and the Modern World.* Routledge, London.

Pagden, A., ed., 1987. *The Languages of Political Theory in Early Modern Europe.* Cambridge University Press, Cambridge.

Paine, T. 1989. *Political Writings*, ed. B. Kuklick. Cambridge University Press, Cambridge.

Paley, W. 1825. *The Principles of Moral and Political Philosophy, Collected Works*, Vol. 4. C. and J. Rivington, London.

Patterson, O. 1991. *Freedom in the Making of Western Culture.* Basic Books, New York.

Pettit, P. 1995. Institutional design and rational choice. In *The Theory of Institutional Design*, ed. R. E. Goodin. Cambridge University Press, Cambridge.

Pettit, P. 1996. Freedom as antipower. *Ethics* 106, 576–604.

Pettit, P. 1997a. *Republicanism: A Theory of Freedom and Government.* Oxford University Press, Oxford.

Pettit, P. 1997b. Republican theory and criminal punishment. *Utilitas* 9, 59–79.

Pettit, P. 2006. Freedom in the market. *Politics, Philosophy and Economics* 5, 131–149.

Pettit, P. 2007. Republican liberty: Three axioms, four theorems. In *Republicanism and Political Theory.* ed. C. Laborde and J. Maynor. Oxford University Press, Oxford.

Pettit, P., and Smith, M. 1996. Freedom in belief and desire. *Journal of Philosophy* 93, 429–449.

Pocock, J. G. A. 1975. *The Machiavellian Moment: Florentine Political Theory and the Atlantic Republican Tradition.* Princeton University Press, Princeton, NJ.

Price, R. 1991. *Political Writings*, ed. D. O. Thomas. Cambridge University Press, Cambridge.

Priestley, J. 1993. *Political Writings*, ed. P. N. Miller. Cambridge University Press, Cambridge.

Rahe, P. A. 1992. *Republics, Ancient and Modern: Classical Republicanism and the American Revolution.* University of Chicago Press, Chicago.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA.

Robbins, C. 1959. *The Eighteenth Century Commonwealthman*. Harvard University Press, Cambridge, MA.

Sen, A. 1985. *Commodities and Capabilities*. North-Holland, Amsterdam.

Skinner, Q. 1978. *The Foundations of Modern Political Thought*, 2 vols. Cambridge University Press, Cambridge.

Skinner, Q. 1983. Machiavelli on the maintenance of liberty. *Politics* 18, 3–15.

Skinner, Q. 1984. The idea of negative liberty. In *Philosophy in History*, ed. R. Rorty, J. B. Schneewind, and Q. Skinner. Cambridge University Press, Cambridge.

Spitz, J.-F. 1995a. *La Liberté Politique*. Presses Universitaires de France, Paris.

Strawson, P. F. 1973. Freedom and Resentment. In *Freedom and Resentment and Other Essays*. Methuen, London, pp. 1–25.

Sunstein, C. R. 1990. *After the Rights Revolution: Reconceiving the Regulatory State*. Harvard University Press, Cambridge, MA.

Sunstein, C. R. 1993a. *The Partial Constitution*. Harvard University Press, Cambridge, MA.

Sunstein, C. R. 1993b. *Democracy and the Problem of Free Speech*. Free Press, New York.

Sydney, A. 1990. *Discourses Concerning Government*, ed. T. G. West. Liberty Classics, Indianapolis.

Taylor, C. 1989. Cross-purposes: The liberal-communitarian debate. In *Liberalism and the Moral Life,* ed. N. L. Rosenblum. Cambridge, MA: Harvard University Press, pp. 159–182.

Trenchard, J., and Thomas G. 1971. *Cato's Letters*, 6th (1755) ed. New York: Da Capo.

Van Parijs, P. 1995. *Real Freedom for All*. Oxford University Press, Oxford.

Viroli, M. 1995. *For Love of Country*. Oxford University Press, Oxford.

17

Rule Utilitarianism and Liberal Priorities

Jonathan Riley

17.1 Introduction

Liberals typically claim that equal rights and liberties should have suitable moral priority over competing values. Rawls (1971, 1993), for example, argues that social justice properly limits permissible conceptions of the good life in a democratic culture and that, within justice as he conceives it, a first principle of equal basic liberties has absolute priority over a second two-part principle. Political rights are given a special place within his first principle, in that they (unlike other basic rights) must have roughly equal worth (or "fair value") for all persons (1993, pp. 5–6, 324–331, 356–362). The rationale for this special treatment seems to be the allegedly essential role that exercise of the political liberties plays in "preserving the other liberties" (Rawls, 1993, p. 299). Thus, individuals must have a fair opportunity to exercise their basic political rights "because it is essential in order to establish just legislation and also to make sure that the fair political process specified by the constitution is open to everyone on a basis of rough equality" (Rawls, 1993, p. 330).

A major aim of Rawls in particular, and of liberal theorists more generally, has been to provide a more secure foundation than utilitarianism seems to provide for the suitable priority of equal basic rights.[1] As Rawls

[1] It should be noted that non-basic rights and liberties are distributed under Rawls's second principle of justice, more specifically, the first part of it known as the principle of fair equality of opportunity. Thus, the basic rights of his first principle take lexical priority over non-basic rights within his theory. In his view, the basic rights are essential to the latent democratic ideal of a citizen as a moral agent with the powers of rationality and reasonableness as he conceives them, whereas non-basic rights are not essential to that

I am grateful to Ken Binmore, Charles Blackorby, John Broome, David Donaldson, John Weymark, and an anonymous referee for helpful discussions. Responsibility for the views expressed remains mine.

puts it, the "first task" of his political theory is "to provide a more secure and acceptable basis for constitutional principles and basic rights and liberties than utilitarianism seems to allow" (1985, p. 226). It is open to doubt whether this "first" task has been accomplished by Rawls or any other antiutilitarian.[2] Leaving that aside, however, I shall confine attention to the suggestion that utilitarianism is incapable of providing a reasonable guarantee of the sort that liberal political culture requires, namely, a guarantee that equal rights and liberties will be duly privileged over competing values.[3]

Modern utilitarians have not been slow to respond to that challenge. Harsanyi (1977, 1992), for example, argues that rule utilitarianism offers a reasonable foundation for liberal claims. In his view, rights are "morally protected" in the sense that they "*cannot be overridden* merely because by overriding them one could here and now increase social utility – except possibly in some very special cases where fundamentally important interests of society are at stake" (1992, p. 689, emphasis original). Basic liberty is secured by moral rules with which all persons should comply to maximize social utility in the long term, except perhaps in rare instances where suspension of the rules is justified to avoid social catastrophe.

ideal. At the same time, because the principle of fair equality of opportunity takes lexical priority over the "difference principle" (the second part of his second principle), non-basic rights take lexical priority over social and political policies that are designed to maximize the position of the worst-off in society in terms of primary goods. In particular, justice apparently demands that even non-basic rights must be respected by any fair policy of income and wealth redistribution. In his critique of Rawls's theory, Arrow (1973) argues that an ordinalist version of utilitarianism can replicate the difference principle in spirit by maximizing the position of the worst off in terms of interpersonally comparable ordinal utilities rather than in terms of primary goods. Even so, the difference principle, whether spelled out in terms of utilities or primary goods, remains controversial. See, for example, Arrow (1977). In what follows, I shall not address the issue of what constitutes fairness in the distribution of income and wealth.

[2] Rawls asserts that justice and right have priority over the good and that equal basic rights have priority over other aspects of justice, for instance, but he does not really try to explain why these priorities are compelling other than to claim that they are in accord with our considered convictions as democratic citizens. True, his principles and priority rules can arguably be inserted as a module into any one of a reasonable plurality of comprehensive theories of the good. Yet, ambiguity remains as to why this module can reasonably be said to be sufficiently more valuable than competing considerations to do its work. The judgment that the political conception should have suitable priority seems to require an assessment of its value relative to the other ingredients of any comprehensive doctrine of good in cases of conflict. But Rawls apparently wishes to avoid such comprehensive assessments of comprehensive doctrines.

[3] For similar statements to that effect, see also Rawls (1971, pp. 22–33, 315–317; 1993, p. 37).

"Only rule utilitarianism," he insists, "can explain why a society will be better off if people's behavior is constrained by a network of moral rights and moral obligations which, barring extreme emergencies, must not be violated on grounds of mere social expediency considerations" (Harsanyi, 1982, p. 41). Act utilitarianism, which asks each person to choose without constraint the social-utility maximizing action in every situation, is, by contrast, "a super-Machiavellistic morality" (Harsanyi, 1982, p. 41). It ought to be abandoned by utilitarians themselves because it authorizes "infringement of all individual rights and all institutional obligations in the name of some narrowly defined social utility," inferior to that attainable under rule utilitarianism (Harsanyi, 1982, p. 41).

Strictly speaking, a rule utilitarian must hold that any conduct permitted or compelled by the moral rights and obligations distributed by an optimal code always produces at least as much general utility as is produced by any other conduct, given that the code has been duly tailored to admit exceptions to its general rules in special situations. By "duly tailored," I mean that the exceptions themselves are framed as special rules that override and limit the general rules in extraordinary cases. Given such an optimal code, acts or omissions that seem to produce more general utility than compliance produces do not really do so and instead rely on a fake overly narrow conception of general utility that ignores the social benefits of security, freedom, and so forth that are possible only under the code. If act utilitarianism is conceived such that it replicates an optimal rule utilitarianism, however, then there is no longer any distinction between act utilitarianism and rule utilitarianism. Although this is a possible move, I shall follow Harsanyi and use the term "act utilitarianism" as it is commonly used in the literature, to refer to a crude version of utilitarianism which assumes (mistakenly) that some acts or omissions really do produce more general utility than is produced by compliance with an optimal code (see, also, Riley, 2000).

Despite his liberal perspective, Harsanyi resists the idea that social justice requires absolute priority rules. He criticizes Rawls in particular for suggesting that conflicts "can be resolved by the simple-minded expedient of establishing rigid *absolute priorities* between different social values, for instance by declaring that *liberty* (or, more exactly, the greatest possible basic liberty for everybody so far as this is compatible with equal liberty for everybody else) shall have absolute priority over solving the problems of *social and economic inequality*" (1992, p. 696, emphasis original). True, a rule utilitarian moral code must itself have absolute priority over other considerations, he seems to admit: "To be sure, there is a clear case of *absolute priority* of one social value over all others. It is the absolute priority we must assign to our *moral duties* over personal interests and over all other nonmoral considerations" (2008, p. 74, emphasis original).[4] But common sense tells us that rigid rules of absolute priority must be avoided *within* an optimal code:

> Common sense tells us that social life is full of situations where we have to *weigh* different social values against each other and must find morally and politically acceptable *trade-offs* between them ... Surely, there will be cases where common sense will tell us to accept a *very small* reduction in our liberties if this is a price for a *substantial* reduction in social and economic inequalities. (1992, p. 696, emphasis in original)

By implication, we should accept that even basic rights will not be given absolute protection. Any particular right may have to be sacrificed at times in favor of competing rights or duties. Social utility may even demand that other social values, such as a fair distribution of wealth, should take priority over all of "our liberties" within the code.

Yet Harsanyi's distaste for absolute priorities is not shared by all utilitarians. J. S. Mill (1859, 1861), for one, suggests that, if social utility is to be maximized in any advanced culture, then some set of equal rights and liberties must be given absolute priority over competing social values. Moreover, certain basic moral rights and correlative duties should be indefeasible even by the other rights, he suggests, although he differs with Rawls over the nature of these basic rights. For instance, Mill argues that social utility maximization requires *absolute* rights to choose as one pleases among what he called "purely self-regarding" actions (1859, p. 224).[5]

I shall argue that an optimal code necessarily gives absolute priority to *some* set of equal rights and liberties over competing social values. In contrast to crude act utilitarianism, which denies individuals any freedom to choose which among multiple acts and omissions shall maximize the general welfare, rule utilitarianism recognizes that the general welfare can be permanently enhanced by assuring some measure of individual freedom to make choices. Moreover, it seems plausible to hold that certain moral claim-rights, including a claim to complete liberty of self-regarding conduct in Mill's sense, should be given absolute priority within an optimal code over other rights, including claims, liberties, powers, and immunities, whenever the latter conflict with them.[6] In effect, I am suggesting that basic

[4] I shall ignore other statements by Harsanyi (e.g., 1985a, p. 54) that seem to contradict this.

[5] For further discussion of Mill's doctrine of liberty, see Riley (1989–1990, 1998a, 2007b).

[6] Moral rights can be defined to include the various instruments that Hohfeld (1919) in the legal context calls claims, liberties or privileges, powers, and immunities. Claim-rights,

moral rights are strong co-possible claim-rights which other people have no powers to abrogate or alter, although the claim-holder (or his agent) might be empowered to waive his claims for extraordinary reasons.[7] The idea is that the acts and omissions permitted by these basic claim-rights or compelled by the correlative obligations are of a far more valuable kind than any competing acts or omissions. A way to ensure that this extremely valuable kind of conduct receives such absolute protection within the code requires modification of the orthodox expected utility theory adopted by Harsanyi.[8]

17.2 Harsanyi's Liberal Rule Utilitarianism

Any utilitarian doctrine supposes that the weight attached to equal rights and liberties relative to other considerations ought to be determined by social utility. If equal freedom is to be given relatively high weight, therefore, utilitarians must argue that general utility-maximization requires suitable priority for a code of rules distributing equal claims and liberties. In effect, Harsanyi's rule utilitarian doctrine does give priority to such a code of equal justice.

With the caveat that some tensions may exist between his various statements, Harsanyi may be fairly interpreted to view utilitarian moral judgments as one species (among others) of rational preferences. Preferences are understood as consistent rankings of possible options. Such rankings are not based on, and do not necessarily reflect, desire-satisfaction or feelings of pleasure. Rather, they are revealed by, and indeed are equivalent to, rational choice behavior, which *would be observed* if the agent were guided by the axioms that comprise a given conception of rationality.[9] Moreover,

which are claims on society to protect a person's vital interests from being unduly harmed by others, imply correlative duties for others. In contrast, liberty-rights do not correlate to duties for others. Rather, multiple individuals may each have liberties to compete with respect to the same outcome such as making a sale to a given buyer. Power-rights enable a person to create or alter other kinds of rights such as claims and correlative obligations, whereas immunity-rights disable other people from abrogating or altering a person's other rights.

[7] I shall henceforth take for granted that basic moral rights are complex combinations of claims, immunities, and possibly powers of waiver. But I shall not discuss immunities and powers in any further detail.

[8] It should be noted that the possibility of waiving his basic rights enables the individual to engage in heroic supererogatory conduct in which he willingly sacrifices his vital interests to rescue other people from grave harm. By implication, such supererogatory conduct might be classed as a far more valuable kind of conduct than even the conduct that is protected by basic rights and correlative obligations.

[9] Actual choice behavior might not reveal rational preferences, whatever the conception of rationality. Even if rationality is exhausted by the axioms of completeness, reflexivity, and

as Binmore (1994, p. 52) emphasizes, Harsanyi's utility theory is "entirely orthodox" among neoclassical economists. "Utility" is merely a term for the value of a function that is employed for convenience to represent the relevant rational behavior: "A utility function, in the modern sense, is nothing more than a mathematically tractable means of expressing the fact that an individual's choice behavior is *consistent*" (Binmore, 1994, p. 51, emphasis original).

To be called utilitarian, any person's preferences must satisfy Harsanyi's complex idea of a moral type of rational choice, in which a formidable array of rationality conditions is implicated, including expected utility criteria; a requirement that attitudes toward risk and uncertainty embedded in the form of a von Neumann–Morgenstern (vNM) utility function should be used to measure relative preference intensities; a full information condition; the equiprobability criterion, whereby a moral decision maker must forget his actual circumstances and adopt an impersonal viewpoint by imagining that he has an equal chance of occupying any person's social position with that person's preferences; Dworkin's (1977, p. 234) requirement that "other-oriented" preferences, defined over other people's positions, must be ignored altogether by impersonal observers; a fundamental similarity postulate, ensuring that identical interpersonal comparisons and moral choices will be made by all impersonal observers; and, finally, a requirement that impersonal observers must jointly constrain themselves by making a binding commitment to the same optimal moral code. The relevant rationality conditions are viewed as *hypothetical imperatives* (rather than Kantian categorical imperatives), in that their motivating force is contingent on the possession of certain attitudes, including attitudes toward risk and uncertainty as well as attitudes toward other people.[10] If his preferences do satisfy

transitivity, which define the idea of an ordering, for example, actual choices might violate transitivity, in which case the agent does not behave as if he has a preference ordering. *A fortiori*, if rationality is defined to include more conditions, actual choice behavior might violate those conditions, in which case the agent does not behave as if he has a rational preference of the relevant sort.

[10] A person's behavior might satisfy some but not all of the rationality conditions that comprise Harsanyi's moral type of rational choice. Thus, it is possible to be amoral yet still be rational in a weaker, nonmoral sense. Binmore (1994, 1998, 2005) does not go along with Harsanyi's moral species of rational choice, for example, although he shares Harsanyi's view that utility merely represents rational behavior and also agrees that a Bayesian conception of rationality is better than alternatives. Rather than introduce axioms (of impersonal choice and so on) to define a distinctive moral sort of rational choice, Binmore views moral behavior as a short-run phenomenon that tends to converge on Nash bargaining behavior in the medium run. As exogenous changes occur in the set of feasible options, moral conduct is associated with an innate moral sense that makes use of existing cultural standards of interpersonal

all of the relevant conditions, then the person behaves as if he were a rule utilitarian (whatever his desires and feelings may be). Moreover, social and moral utility is merely the name for the value of a (cardinal and interpersonally comparable) function, of the traditional utilitarian form, that represents rational preferences of this complex moral sort.[11]

Harsanyi argues that utilitarians have good reasons to commit themselves to a code whose rules distribute equal rights and liberties. Given that people are naturally biased in favor of themselves and close associates rather than completely impartial, a code that provides some freedom for the individual to display particular concern for family and friends comports better with human nature than crude act utilitarianism does because the latter recognizes no such freedom. Act utilitarianism, he insists, has "*intolerably burdensome* negative implementation effects" (1992, p. 688, emphasis original). It is simply too demanding for humans as opposed to gods or saints. Its "rigidly universalistic principles" would require "a complete suppression of our natural inclinations," which are "particularistic" in the sense that the agent gives "greater weight" to the interests of himself, his family and friends than to the interests of other people" (Harsanyi, 1992, pp. 675, 688). Such suppression "could be done, if it could be done at all, only by extreme efforts and at extremely high psychological costs" (Harsanyi, 1992, p. 688). Thus, a rule utilitarian code could recognize that the average person gains utility if he is duly permitted to make his own choices in pursuit of his particular interests, rather than forced always to do his strictly impartial duty as determined by social utility maximization in crude act utilitarian terms.[12]

comparison such that rational self-interested agents are led in the short run to coordinate on a proportional bargaining solution. But this moral solution reduces to the Nash bargaining solution if exogenous changes cease and cultural standards adjust to a medium-run social equilibrium. For further discussion of Binmore's theory of justice, see Riley (2006b).

[11] Broome suggests that Harsanyi's "use of 'utility' is ambiguous" because "he sometimes uses the term . . . as a name for the value of a function that represents . . . a preference ordering" but other times uses it "to mean good" (1991, p. 59, n. 19). But the two uses are not necessarily incompatible. For Harsanyi, utility does represent given *rational* preferences defined over a relevant set of objects. At the same time, the relevant *rationality criteria* also shape the preferences that constitute the meaning of good. Thus, utility can represent moral good as well as rational preference precisely because moral good is equated with a complex moral type of rational preference. Given that an individual who reveals a moral preference must be acting in accord with the rules of an optimal code, Harsanyi has considerable leeway to use utility to encompass freedom as permitted by the rights and liberties distributed by an optimal code.

[12] By assigning a "procedural utility" to "free moral choice," a rule utilitarian code could also recognize "the traditional, and intuitively very appealing, distinction between merely doing one's *duty* and performing a *supererogatory* action going beyond the call of duty" (Harsanyi, 1992, p. 689, emphasis original).

A second reason for choosing a rule utilitarian code that affords substantial equal freedom relates to the positive "expectation effects" (including "incentive effects" and "assurance effects") associated with its system of rights and liberties. Rules that distribute and enforce weighty private property rights, for example, "provide socially desirable *incentives* to hard work, saving, investment, and entrepreneurial activities" and "also give property owners *assurance* of some financial security and of some independence of other people's good will" (Harsanyi, 1992, p. 690, emphasis original). Indeed, "widespread property ownership" yields "a kind of assurance effect benefiting society as a whole," to wit, enhanced "social stability" and "personal and political freedom" (Harsanyi, 1992, p. 690). Act utilitarianism cannot recognize those positive expectation effects of an entire system of equal rights and correlative duties. It considers only the expectation effects of individual actions, which are "normally . . . negligibly small" (Harsanyi, 1992, p. 690).[13]

As Harsanyi admits, though, the relevant incentive and assurance effects are really only socially useful for people like us, who are naturally inclined to be partial toward our particularistic concerns. Unlike act utilitarian saints, for example, we may be little inclined to engage in socially useful production in the absence of a suitable system of private-property rights and correlative duties. Saints, however, would perform any acts of work and saving required to maximize social utility in act utilitarian terms, independently of any rights to own the fruits of their labor and saving. Similarly, we may refuse to lend money to relatively poor people unless we have the assurance that promissory obligations will be enforced. Saints will not care if the poor individual breaks his promise to repay a debt, however, and "will shower further wealth upon him until each has the same marginal utility for one dollar more or less" (Binmore, 1998, p. 157).[14]

It emerges that, for Harsanyi, a rule utilitarian code that suitably privileges equal rights and liberties maximizes social utility, not so much because humans have intellectual and moral deficiencies, but because the most

[13] For further discussion of this important point, see Harsanyi (1985a, pp. 42–46; 1985b, pp. 70–72).

[14] Binmore remarks that there would be no difference between rule utilitarianism and act utilitarianism in an ideal society, where everybody is an act utilitarian "taking the same far-sighted view of the nature of social utility" under perfect information, and everybody always jointly performs whatever actions are required to maximize social utility (1998, p. 156, n. 24). Like Harsanyi, however, he finds act utilitarianism too demanding for humans. As a second-best "bourgeois liberal" theory of human morality, he recommends his own Humean alternative to rule utilitarianism (Binmore, 1998, pp. 157–158). For related discussion, see Riley (1998b, 2000, 2006b).

rational and empathetic among us, recognizing our fixed particularistic natures, would place a peculiar value on the freedom of the individual to make his own choices and on the incentives and assurances made possible by a reasonable network of equal rights and liberties. Liberal utilitarian rules are not merely tools for mitigating our limited capacities, which act utilitarianism mistakenly supposes are unlimited. Rather, they are socially useful constraints that retain their utility for humans as opposed to saints, even if humans are assumed to be fully informed wizards perfectly capable of imagining themselves in the shoes of others while pretending to have those others' particular concerns.

More formally, consider Harsanyi's two-stage n-person game of rule utilitarianism with full compliance under complete information (1992, pp. 692–694). The first stage is a cooperative game, in which all players agree to apply the equiprobability model to choose an optimal code m^* from some given set M of feasible moral codes. Given the fundamental similarity postulate, we may speak of a single impersonal observer making the choice at this stage. Any feasible code $m \in M$ gives rise to a permissible strategy set $P(m)$, which is the same for all players who bind themselves to comply with the code.[15]

The second stage is a non-cooperative game in which each player k is free to choose a (pure or mixed) personal strategy s_k from his feasible strategy set S_k so as to maximize his personal utility, subject to the requirement that

$$\forall k\colon s_k \in P(m^*),$$

where $P(m^*)$ is the permissible strategy set associated with the optimal code m^* chosen at the first stage. Any person's rights and obligations are included among his permissible strategies. The idea of equal rights and liberties for all is reflected in the fact that $P(m)$ is the same for everybody. Every person also has a general obligation to refrain from making impermissible strategy choices.

To choose an optimal code m^* at the first stage, an impersonal observer must compare alternative feasible codes to ascertain which of them is associated with the most social utility. The implementation effects of any given code are "represented by the fact that the players' strategies will be restricted to the permissible set $P(m)$ defined by this moral code m" (Harsanyi, 1992, p. 693). The expectation effects are "represented by the fact that some players

[15] Harsanyi assumes that, for all $m \in M$, $P(m)$ is a nonempty compact subset of the feasible strategy set S, which is conveniently assumed to be the same for every player, that is, $S = S_1 = \cdots = S_n$. The latter assumption could easily be dropped to permit each player to have a distinct strategy set, however, in which case $S = \prod S_k$.

will choose different strategies than they would choose if their society had a different moral code – not because *m* directly requires them to do so but rather because these strategies are their *best replies* to the other players' expected strategies, on the assumption that these *other players* will use only strategies permitted by the moral code *m*" (Harsanyi, 1992, pp. 693–694, emphasis original).

But to make a reasonable choice, an observer requires knowledge of how players may reasonably be expected to behave at the second stage, under each of the alternative codes. Given that all will fully comply with the rules, predictions are needed of the permissible strategy choices that each player will actually make when interacting with others, assuming that the choices of each will be best replies to the given choices of the others. In short, the impersonal observer must be able to predict a Nash equilibrium point $\bar{s} = (\bar{s}_1, \ldots, \bar{s}_n)$ of the noncooperative game that will be played at the second stage. Different noncooperative games will emerge for different codes and their respective permissible strategy sets. Given a solution concept (see, e.g., Harsanyi and Selten, 1988), the observer may be assumed to define a "predictor function" π that selects, for every possible noncooperative game $g(m)$, an equilibrium $\bar{s} = \pi(g(m))$.

Harsanyi supposes that any player *k*'s personal utility function u_k takes the form $u_k = u_k(\bar{s}, m)$. He includes *m* as a variable "because the players may derive some direct utility by living in a society whose moral code permits a considerable amount of free individual choice" (1992, p. 693). Because any impersonal observer's social and moral utility function *w* is defined in terms of the personal utility functions $u_1, \ldots, u_n$, however, we must have

$$w = w(\bar{s}, m) = w(\pi(g(m)), m).$$

Social and moral utility maximization is achieved by joint commitment to an optimal code m^* which suitably privileges equal liberty, such that each person *k* is free to choose a permissible personal strategy $\bar{s}_k \in P(m^*)$ that is a best reply to the given permissible strategy choices of his fellows.

The two-stage structure of the game makes it clear that the impersonal observer is choosing an optimal element from a feasible set of *equilibrium* outcomes. That feasible set of equilibrium strategy combinations is also the possible set of incentive-compatible systems of equal rights and liberties. In effect, *prior* to selecting an optimal code, the observer must settle the question of which systems of equal rights are feasible, by predicting the relevant Nash equilibria. Thus, there is a sense in which absolute priority *is* given to the value of equal liberty over distributive and other considerations.[16]

Moreover, an optimal code involves an optimal system of equal rights and liberties, in which the relative weights attached to the different permissible strategies will reflect whatever balancing of distributive and other considerations the impersonal observer decides maximizes social utility. If the observer decides that a duty to redistribute wealth to the needy overrides the right to private property in some circumstances, for example, then the duty in question is correlative to a superior right of the needy to the redistribution. Thus, if an optimal moral code has absolute priority over nonmoral considerations, as Harsanyi apparently thinks, then an optimal system of equal rights and liberties has absolute priority over competing social values.

Evidently, the players in this rule utilitarian game must display remarkable intellectual and moral capacities to identify an optimal code, even if we agree that their commitment to such a code involves less stringent emotional demands than those of act utilitarianism. In addition to imagining themselves with equal probability in each person's position while making the particularistic choices that person would make, these agents must be able to predict how people's interactions will change across distinct codes and associated networks of rights and obligations to calculate which code maximizes social utility. Those interactions can be extremely complicated, necessitating consideration of highly intricate systems of rules involving many exceptional elements, higher-order rules to settle contradictions among lower-order rules, rules for amending the code as circumstances change and knowledge improves, and so on.

Given the requisite capacities, individuals can maximize social utility by jointly committing themselves to an optimal code m^* that suitably privileges equal liberty. Justice and right are thereby subsumed under moral or social utility maximization, in other words, any rule utilitarian agent's conception of the good. Their joint commitment does not merely prevent these individuals from lowering social utility by deviating from the optimal code so as to increase their own personal utilities at the expense of their fellows. As Harsanyi stresses, "the players' commitment to the jointly adopted moral code will [also] prevent them from violating the other players' rights or their own obligations in order to *increase social utility,*" that is, some false notion of social utility (Harsanyi, 1992, p. 694, emphasis original). Crude utilitarian deviations from an optimal code are no less impermissible than selfish deviations. They equally detract from the permanent

[16] For related discussion, see Binmore (1994, pp. 9, 127–131, 147; 1998, ch. 3).

benefits of freedom, productivity, and security made possible by the optimal code.[17]

17.3 Absolute Priorities within a Liberal Utilitarian Code

Liberal utilitarianism as interpreted is open to various objections, despite its considerable strengths. Even if utility is agreed merely to represent rational choice behavior, moral thinkers might not accept all of the rationality criteria entering into the complex moral sort of rational choice behavior that, for Harsanyi, a moral utility function of utilitarian form serves to represent. Utilitarians themselves might reject his Dworkinian condition of equal concern and respect for each person's interests, for example, which forbids impersonal observers to consider anybody's "other-oriented" preferences, including benevolent and malevolent choices (Harsanyi, 1992, pp. 704–706). Without further argument, the condition seems to generate a vicious circle: it precludes the moral calculus from even counting other-oriented utilities, for moral reasons that are not products of the calculus but are binding on it *prior* to its operation. Hare (1981, pp. 169–196) apparently views such a deontological constraint on utilitarian morality as arbitrary.

Nonutilitarians who accept orthodox utility theory may reject entirely his equiprobability model of impersonal choice together with the fundamental similarity postulate.[18] Binmore argues, for example, that the model in effect asks us "to envisage some kind of Kantian consensus [on interpersonal comparisons] that would exist *prior to society itself*. Not only this, we must also believe that the nature of this Kantian consensus can somehow be deduced by a sufficiently rational being from data that makes no reference to the store of common knowledge that [may be identified] with the culture of a society" (1994, p. 299, emphasis original). He offers an alternative model of moral choice, in which a consensus on interpersonal comparisons is a product not of impersonal rationality but of social evolution. Social evolution tends to favor the survival of whatever interpersonal comparisons are made by persons whose behavior patterns are perceived as more successful than others and thus more likely to be imitated and encouraged through education, in a particular cultural context. At an "empathetic equilibrium" (where nobody

[17] More general games of rule utilitarianism may be obtained by relaxing the assumptions of complete and perfect information and of full compliance. See, for example, Harsanyi (1985b, pp. 72–73; 1992, p. 694).

[18] As is well known, Rawls rejects utility theory altogether. He seeks to base his political principles of justice on "primary goods" allegedly valued by all rational and reasonable persons, despite their plural and incommensurable comprehensive doctrines of good.

has an incentive to depart from the interpersonal comparisons he chooses given the interpersonal comparisons made by other people), there is a cultural consensus on how to make interpersonal comparisons of utility, on the basis of which people with perhaps markedly unequal bargaining power coordinate their moral interactions for the moment (Binmore, 1998, ch. 4; 2005, ch. 8, 10, 11).

Nevertheless, I shall largely ignore such objections to contest Harsanyi's distaste for absolute priorities within an optimal liberal utilitarian code.[19] As already discussed, rule utilitarianism does capture something of the lexical priority given by Rawls to equal freedom, in the sense that what is *feasible* in terms of Nash equilibria must be settled by an impersonal observer prior to his choice of an *optimal* code that distributes equal rights and liberties. Absolute priority is given to an optimal code's permissible strategy set over all competing social considerations, ignoring the extraordinary case of social catastrophe that destroys the possibility of any code of morality. Yet, there is certainly no guarantee that a utilitarian system of equal rights and liberties must closely resemble the specific system held out by political liberalism. All that we know is that equal rights and liberties are to be distributed in any situation such that social utility is maximized.[20] *A fortiori*, there is no guarantee that certain basic moral rights must always outweigh other types of rights and duties in cases of conflict. Rather, as Harsanyi insists, absolute priority rules within an optimal rule utilitarian code appear unreasonably rigid because social utility maximization may demand that any kind of right should give way to other social and moral considerations in some situations.

Still, it seems possible to modify Harsanyi's theory such that a rule utilitarianism embodying rigid liberal priorities emerges. In this regard, his rejection of lexical priority rules within an optimal code is tied to his conception of the moral type of rational choice. That conception implies what I shall call *strong monism*, the view that all conflicts of values can ultimately be settled in terms of a single homogeneous social utility function $w = w(\bar{s}, m)$, reflecting a complete and transitive moral preference over

[19] For a critical discussion of Binmore's theory, see Riley (2006b). Among other things, I suggest that there is textual evidence that Hume was inclined to support some version of a utilitarian theory of justice.

[20] That abstract knowledge does distinguish rule utilitarianism from alternative moral theories such as the proportional bargaining theory proposed by Binmore (1994, 1998, 2005). Unlike rule utilitarianism, proportional bargaining might, in some cultural settings, call for a poor and needy person to be given the right to a medical benefit, such as an organ transplant, even though a wealthy person would experience a larger gain of comparable utility from the transplant. Moreover, as Binmore (1994, pp. 146–147) remarks, his theory does not guarantee that equal rights will be distributed to all.

possible Nash equilibrium outcomes. Strong monism involves the striking hypothesis that the various acts chosen by any person are all of the same kind for normative purposes, ultimately reducible to a single scale of homogeneous value. Yet, beyond the definition of rational and moral choice that manifests it, no justification is supplied for such a remarkable hypothesis. Indeed, the classical utilitarian route to it (i.e., crude quantitative hedonism) is dismissed as incredible by Harsanyi.[21]

Rule utilitarianism can embed absolute liberal priorities within an optimal code by adopting a theory of moral choice that does not imply strong monism. Such a theory will involve modifications to the array of rationality conditions entering into Harsanyi's conception of a moral type of rational choice. Consider, for example, the expected utility criteria together with his distinct claim that an expected utility maximizer ought to use his attitudes toward risk (as revealed in the form of a vNM utility function) to measure his relative intensities of preference. According to vNM theory, a rational person behaves under risk *as if* he has not only a preference ordering over his possible alternative acts and omissions but also consistent beliefs over alternative events, where his beliefs correspond to an objective probability distribution (identical for all rational persons). Actions (including omissions) are seen as functions $f \in F$ from states-of-the-world $e \in E$ to consequences $x \in X$. A particular consequence x is the joint effect of an action f and a state e, which occurs with probability $p_i \in p$. Preferences over consequences are a special case of preferences over actions because a consequence x can be identified with the act having constant value x . The vNM axioms imply that the person's preferences can be represented by an expected utility function $\Sigma u(f(e))p(e)$, where $u(\cdot)$ is a function over X, and $p(\cdot)$ is a probability distribution over E. That vNM utility function u is a cardinal function because it is unique up to a positive affine transformation.[22]

[21] Even Sidgwick (1907), who embraced quantitative hedonism and tried to combine it with rational intuitionism, concluded that practical reason is ultimately divided against itself such that it cannot always resolve conflicts between egoistic hedonism and universalistic hedonism (classical utilitarianism). This "dualism of practical reason" implies that strong monism is not feasible for human beings, although Sidgwick seems to entertain the possibility that some form of divine reason (about which we know nothing) might make monism feasible for a god.

[22] For further discussion of vNM theory, see Binmore (1994, pp. 266–278, 304–315). The extension of the theory by Savage and others to the case of uncertainty, where objective probabilities are not available, may be ignored for present purposes. This is not to deny that a rather heroic assumption is being made when it is supposed that all rational persons (including impersonal observers) behave under uncertainty as if they face an identical subjective probability distribution. As Binmore points out, however, orthodox subjective utility theory regards such behavior as virtually tautological.

It is doubtless true, as Weymark (1991) and others have insisted, that this rational person's preference ordering can equally be represented by many utility functions other than vNM utility functions. There is nothing in his choice behavior under risk to justify a restriction to the cardinal family of vNM utility functions. But Harsanyi (1992, p. 685) claims that, "in ethics," rationality also requires the person to employ his mental *attitudes* toward risk, as embodied in the form of a vNM utility function, to measure his relative intensities of preference over alternative actions. That rationality condition, as distinct from the requirement to *behave* in accord with expected utility criteria, restricts the representation of such behavior to vNM utility functions. Harsanyi apparently views the condition as a hypothetical imperative of reason, whose validity will be self-evident to anybody who becomes aware of the particular attitudes toward risk that lurk within him and help to define him as a person. But no attempt is made to explain the *formation* of such attitudes. Rather, their possession by any rational and moral human being is simply assumed. It might seem that Harsanyi merely picks them out of thin air, therefore, and relies on them without justification to smuggle in the cardinality that his rule utilitarianism needs to operate.[23]

Still, the claim that a rational and moral person possesses, and makes suitable use of, attitudes toward risk as well as other attitudes requisite to utilitarian ethics can be separated from the question of how any person comes to possess the relevant attitudes. I shall follow Harsanyi's lead in this respect, with the caveat that more needs to be said about how the relevant attitudes are formed. More will be said, however briefly, in the conclusion.

To go beyond strong monism in the way required to give absolute priority to certain basic moral rights within an optimal code, orthodox expected utility theory must be replaced by what might be called a "plural" expected utility theory (see, Sen, 1980). According to a Millian version of the latter, rational behavior under risk involves recognizing that the feasible set of actions must be partitioned into plural subsets or kinds, where the different kinds are arranged in a hierarchy such that any kind is intrinsically

[23] In contrast, Binmore argues that *social evolution* selects in favor of consistent behavior that can only be represented in terms of maximizing a vNM utility function. His view seems to be that successful individuals tend to behave as if they possess attitudes toward risk, which they use to measure relative preference intensities. The need to survive in a competitive process of social evolution explains the required restriction to cardinal utility functions: "Evolution does not reward consistent behavior because it is consistent. Only consistent behavior that is directed at its own survival is rewarded" (1994, pp. 324–5; see also pp. 273–274, n. 21). Evidently, Nash bargaining (which is central to Binmore's approach) relies on such cardinal utility information.

or infinitely more (less) valuable than others below (above) it in the hierarchy. Because there are plural kinds of actions, there are plural systems of expected utility criteria, one for each kind of action (implicitly, consequences). Define $F(\alpha)$ as the set of actions of type α, where any action $f(\alpha)$ is a function from states-of-the-world $e(\alpha) \in E(\alpha)$ to consequences $x(\alpha) \in X(\alpha)$ of type α. Similarly, define $F(\beta)$ as the set of actions of type β, and so on. Let $x_1(\alpha)$ and $x_q(\alpha)$ be the best and worst possible consequences, or outcomes, of type α for person k, $x_1(\beta)$ and $x_q(\beta)$ be the best and worst possible outcomes of type β for him, and so on. Define $p^{\alpha}(z_i^{\alpha})$ as a restricted lottery that offers $x_1(\alpha)$ with probability $p_1^{\alpha} = z_i^{\alpha}$, and $\mathrm{x}_q(\alpha)$ with probability $\mathrm{p}_q^{\alpha} = 1 - z_i^{\alpha}$, $i = 1, \ldots, q$. Similarly, define $p^{\beta}(z_i^{\beta})$ as a restricted lottery that offers $x_1(\beta)$ with probability $p_1^{\beta} = z_i^{\beta}$, and $x_q(\beta)$ with probability $p_q^{\beta} = 1 - z_i^{\beta}$, and so on. Then, following Binmore (1994, pp. 272–273), axioms of continuity, independence, probabilistic equivalence, and optimization can be defined for each kind of action.

The continuity axiom for actions of type α, for example, may be stated as:

α-Continuity: For every consequence $x_i(\alpha) \in X(\alpha)$, there exists some restricted lottery $p^{\alpha}(z_i^{\alpha}) = (x_1(\alpha), z_i^{\alpha}; x_q(\alpha), 1 - z_i^{\alpha})$, $0 \leq z_i^{\alpha} \leq 1$, such that person k is indifferent between $x_i(\alpha)$ and $p^{\alpha}(z_i^{\alpha})$.

This axiom says that it is always possible for k to find some probability z_i^{α} of getting his best possible outcome $x_1(\alpha)$ of type α that would lead him to reveal indifference of the highest kind α between any given sure outcome $x_i(\alpha)$ and a restricted lottery $p^{\alpha}(z_i^{\alpha}) = (x_1(\alpha), z_i^{\alpha}; x_q(\alpha), 1 - z_i^{\alpha})$. An analogous β-continuity axiom may be stated for actions of type β, and so on.

An α-independence axiom may be stated as:

α-Independence: Given that person k is indifferent between any $x_i(\alpha)$ and a restricted lottery $p^{\alpha}(z_i^{\alpha})$, $i = 1, \ldots, q$, then k is indifferent if the q sure outcomes $x_i(\alpha)$ in any lottery p^{α} are replaced by the q restricted lotteries $p^{\alpha}(\mathrm{z}_i^{\alpha})$.

In other words, person k must not object if a simple lottery is transformed into a compound lottery with what for him are equivalent outcomes or prizes of type α. Again, an analogous β-independence axiom may be defined, and so on.

Without going into further details, it is sufficient to emphasize that separate continuity, independence, probabilistic equivalence, and optimization axioms must be imposed in the context of each kind of action. Any person who behaves in accord with these plural expected utility criteria may be described as maximizing separate vNM utility functions u^{α}, u^{β}, ..., defined over the different kinds of (sure) outcomes, respectively. At the same time, a plural expected utility maximizer is required to give absolute priority to higher kinds of actions over lower kinds. If an action of a higher kind comes into conflict with one of a lower kind, he must always choose the higher kind. Leaving aside the issue of what distinguishes the natures of the different kinds of acts, a rational agent of this sort behaves as if he has a corresponding hierarchy of distinct preference orderings, one for each of the plural kinds of actions. Moreover, he is assumed to possess complex attitudes toward risk, which he employs to measure his relative preference intensities of each kind and to arrange the plural kinds in a lexical order (identical for all rational agents). The lexical arrangement reflects any rational agent's unwillingness to ever risk the loss of any amount of a higher kind of utility to gain any amount of a lower kind. His preference for the higher kind of actions is *infinitely* more intense than (and in that sense incommensurable with) his preference for the lower kind.

Rational behavior of this sort may be described as maximizing over a hierarchy of plural distinct kinds of vNM utility functions, one for each kind of action. When choosing among acts of the same kind, the individual behaves as though he maximizes the value of a vNM utility function restricted to the relevant set of actions. His relative preference intensities are measured by his attitudes toward risk in the context of actions of that kind. When choosing between acts of different kinds, however, he invariably selects the higher kind of action. He always behaves as if he floats to the top of the hierarchy when it is feasible to do so, ignoring his lower kinds of utility functions when maximization of a higher kind is possible.

If plural expected utility theory is suitably combined with Harsanyi's other conditions of moral choice (including the equiprobability criterion and so on), an impersonal observer behaves as if he maximizes over a set of heterogeneous social and moral utility functions w^{α}, w^{β}, ..., which are related to each other by lexical priority rules, as opposed to a single homogeneous function $w = w(\bar{s}, m)$. The highest kind of social and moral utility function, $w^{\alpha} = w^{\alpha}(\bar{s}^{\alpha}, m^{\alpha})$, represents the observer's judgments about moral rules to govern the kind of acts and omissions that he holds to be intrinsically more valuable than any other kind in cases of conflict. He is assumed to construct the portion of an optimal code that distributes claim-rights and correlative duties respecting the performance of the relevant acts and omissions, as the case may be. Because the relevant acts and omissions are intrinsically more valuable than any competing kinds, this portion of the code may be described as a top layer of rules whose directions cannot legitimately be overridden by any other part of the code. In effect, the observer maximizes w^{α} by constructing an optimal coherent system of equal claim-rights, which trumps any competing social and moral considerations.

A more complex Millian sort of liberal utilitarianism emerges, which recognizes the absolute priority of a system of equal basic rights over all other strategies permitted (and not permitted) by an optimal code.[24] Suppose there are two different kinds of actions, for example. Actions of the higher kind are so inherently valuable that an impersonal observer attaches basic moral rights to their performance, where the right-holder's claim correlates to moral obligations for others (either to omit to interfere with his action, or to act on his behalf). Actions of the lower kind are recognized by the observer as permissible in some situations, yet insufficiently valuable to be thus morally protected. Rather, anyone has a non-basic liberty or permission to perform such actions in competition with his fellows, no one of whom has any duty not to interfere with other people's actions of this kind. An optimal moral code m^* then consists in effect of two layers or subcodes $M^{\alpha *}$, $M^{\beta *}$, related by a lexical priority rule. Similarly, the permissible strategy set $P(m^*)$ resembles a layer cake, involving two distinct layers of permissible strategies arranged in a hierarchy. Absolute priority is given to the top layer of basic claim-rights s^{α} over the second layer of non-basic liberties s^{β} in cases of conflict. In this simple case, the absolute priority rule is manifested by the fact that moral obligations correlate only to claim-rights. Everyone has a duty not to interfere with any personal strategy choice classed by the code as a claim-right, whereas they have no duty not to compete with personal choices classed as mere liberties. Unless the code is suspended, claim-rights can never be overridden by mere liberties.

Without going any further, it is an open question whether such a liberal utilitarianism has more appeal than Harsanyi's more flexible brand of liberal utilitarianism. Evidently, clarification is needed of which actions are infinitely more valuable than others, and why. Also required are more adequate models of basic moral rights (which are complex combinations of claims, immunities, and so forth) and non-basic rights to make precise the sense in which the one kind of permissible strategy is more powerful than the other. In this regard, claim-rights might be seen as permissible strategies

[24] Recall that basic moral rights are defined such that they may include powers to waive the basic claims. Thus, supererogatory waiver of basic claims remains a permissible strategy and it may even be classified as far more valuable than a choice to enforce one's claims and correlative duties. Space constraints prevent me from discussing such possibilities any further.

that, by obliging other people to perform correlative actions or inactions, fix or sustain personal aspects of Nash equilibrium outcomes of the games played at the second stage of the two-stage rule utilitarian game. Such personal aspects are thereby put under the control of the right-holder, even if he is not always the agent of the relevant actions (duty-holders may have to act on his behalf). Naked liberties or permissions, however, are permissible strategies that do not impose correlative duties on others. This weaker kind of strategy fixes personal aspects of the Nash equilibria only with some positive probability less than unity, depending on the relative bargaining power of the liberty-holder. Two or more players may each have liberties to compete for the same scarce job, for example, which only one player can succeed in winning. But more work is needed in this area.[25]

17.4 Conclusion

My argument has been twofold. First, liberal rule utilitarianism does seem to provide a "secure and acceptable basis" for equal freedom, contrary to a received view that utilitarianism and liberal priorities are incompatible. Harsanyi's brand of rule utilitarianism gives absolute priority to some system of equal rights and liberties, by virtue that an optimal moral code (and its permissible set of strategies) has absolute priority over all competing considerations (Section 17.2).

Second, some versions of utilitarianism may even be able to consistently accommodate as a module a liberal theory of justice similar to Rawls's, in particular, a theory that gives lexical priority to a principle of equal basic rights over a second principle of non-basic rights and other social and political considerations. Although Rawls himself never accepted the possibility, it seems that a Millian liberal utilitarianism can accommodate absolute liberal priorities *within* its optimal code, by suitably transforming Harsanyi's

[25] For some beginnings along these lines, see Riley (2006a). There has been a tendency in the literature to see a sharp dichotomy between game form formulations and social choice formulations of rights and liberties. But Sen emphasizes that "the alleged dichotomy is more presentational than substantial" (1992, p. 153). Game form models concentrate on each person's permissible strategies, whereas social choice models concentrate directly on power to influence social outcomes. But the definition of permissibility is "worked out – directly or indirectly – in the light of the characteristics and consequences of combining different people's strategies" (p. 153). And an equilibrium "combination of strategies produces an outcome – a social state (even if the social state is described as no more than a particular combination of actions having occurred" (p. 153). For further discussion, see, for example, Gaertner, Pattanaik, and Suzumura (1992); Sen (1992); Pattanaik and Suzumura (1996); Deb, Pattanaik, and Razzolini (1997); and Sen (2002), esp. pp. 642–658.

orthodox utility theory into a pluralistic utility theory (Section 17.3). If we can distinguish plural kinds of choices some of which are intrinsically more valuable than others, a utilitarian code can distribute powerful basic claim-rights and correlative duties to govern the performance of the most valuable kind of choices such that no competing considerations are permitted to override the system of basic rights and duties. More needs to be said about how such a pluralistic utility theory might be made plausible. In this regard, Mill himself seems to have believed that the required distinctions among different kinds of choices could ultimately be rooted in a qualitative hedonistic theory of the different kinds of pleasures and pains that are reasonably expected to flow from the choices. Although Mill's approach has struck some influential philosophers as "contemptible nonsense," the case against such a qualitative hedonism is far from proven.[26]

Finally, it is a legitimate complaint against liberal utilitarianisms as conceived, whether of Harsanyi's variety or of the more complex Millian variety, that they presuppose something like Kantian impersonal observers, with remarkable intellectual and sympathetic capacities to make rational and moral judgments that seem to transcend any particular cultural context. In effect, an impersonal observer must have developed what Aristotle and other ancients called *phronēsis*, that is, such skill in the art of living that his choice behavior displays both the knowledge and motivation required to achieve an ideal harmony of social and moral values, which is then represented as the maximization of social and moral utility.[27] But few if any clues are provided about the process by which phronēsis may be acquired.

Liberal utilitarianism must at least implicitly identify an ideal sort of personal character that is feasible for most human beings and then rely on a psychological theory to explain how individuals can in principle develop the requisite sort of character. The process of education required for most in any social context to acquire the character of an impersonal observer may well be difficult, if even attempted. Yet, there seems no reason to think that humans are forever incapable of developing into something like impersonal

[26] For an excellent discussion of qualitative hedonism along Millian lines and its potential normative appeal, see Edwards (1979). Elsewhere, I have argued that Mill views the complex feeling of "security," which he associates with the moral sentiment of justice, as a higher kind or quality of pleasure, infinitely more valuable than any other kind of pleasure in cases of conflict. See, for example, Riley (2003, 2007c).

[27] For an account of phronēsis, see Annas (1993, pp. 73–84). As she explains, "phronēsis unifies the virtues" and "the fully virtuous person, with complete possession of phronēsis, is an ideal" (p. 83). Unfortunately, she seems to think that "maximizing reasoning" is incompatible with phronēsis, and with ancient moral theory more generally (pp. 86–87, 447–448).

observers. Still, a convincing psychological theory is needed to show how people can be motivated to develop in the direction of the ideal. Unfortunately, as far as I am aware, no consensus exists among scholars about what constitutes a genuine psychology.

Even so, unlike Harsanyi, classical and modern utilitarians alike typically build some theory of motivation directly into the concept of utility, such that utility denotes pleasure (including freedom from pain), desire-satisfaction, virtuous dispositions, or some other such motivation. As a result, ample confusion surrounds the terms *utility* and *preference* in the literature. Some equate utility and preference with desire-satisfaction, for example, whereas others view utility as a mere representation of whatever preference would be revealed by rational choice behavior, independently of desire-satisfaction. If desire-based preference rankings conflict with rational choice–based ones, however, utility cannot possibly refer to both. Only when the two rankings coincide can the two usages of utility and preference be reconciled. In that ideal case, the term *utility* might be used to refer, simultaneously, to the ultimate motivation underlying a preference ranking and to the value of a function that represents that same preference as revealed by rational and moral behavior.

In the absence of a genuine psychology that explains how individuals might be motivated to attain the good, the possibility that most might fashion themselves into something approaching an impersonal observer must be left open. In the meantime, lacking impersonal observers, a liberal utilitarian doctrine must be able to persuade us that people like ourselves, prejudiced in favor of our particularistic concerns, can devise and operate a political system that could more or less stand in for the missing observers. In effect, the doctrine must persuade us that the general welfare is maximized by establishing some form of liberal democracy in which a system of equal basic rights and correlative duties is given absolute priority over competing social considerations. Perhaps Millian liberal utilitarianism may be up to this task. In any case, Mill's view that his utilitarianism grounds a form of representative democracy in which rights to liberty of "purely self-regarding" conduct are viewed as absolute claims, defeasible by no competing considerations (including other claim-rights), deserves serious consideration.[28]

[28] A crucial caveat is that the class of purely self-regarding acts and omissions must be properly defined. Moreover, the right to complete self-regarding liberty does not imply that the individual is free to ignore social duties to other people. Failing to help others, for instance, or omitting to pay one's fair share of taxes, are harmful other-regarding actions. The individual has no right to choose as he pleases among self-regarding and other-regarding acts. For further discussion, see Riley (1988, 1998a, 2007a, 2007b).

References

Annas, J. 1993. *The Morality of Happiness*. Oxford University Press, Oxford.

Arrow, K. J. 1973. Some ordinalist-utilitarian notes on Rawls's theory of justice. *Journal of Philosophy* 70, 245–263.

Arrow, K. J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review, Papers and Proceedings* 67, 219–225.

Binmore, K. 1994. *Playing Fair*. MIT Press, Cambridge, MA.

Binmore, K. 1998. *Just Playing*. MIT Press, Cambridge, MA.

Binmore, K. 2005. *Natural Justice*. Oxford University Press, Oxford.

Broome, J. 1991. *Weighing Goods*. Blackwell, Oxford.

Deb, R., Pattanaik, P., and Razzolini, L. 1997. Game forms, rights, and the efficiency of social outcomes. *Journal of Economic Theory* 72, 74–95.

Dworkin, R. 1977. *Taking Rights Seriously*. Duckworth, London.

Edwards, R. B. 1979. *Pleasures and Pains: A Theory of Qualitative Hedonism*. Cornell University Press, Ithaca, NY.

Gaertner, W., Pattanaik, P., and Suzumura, K. 1992. Individual rights revisited. *Economica* 59, 161–177.

Hare, R. M. 1981. *Moral Thinking*. Clarendon Press, Oxford.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge.

Harsanyi, J. C. 1982. Morality and the theory of rational behavior. In *Utilitarianism and Beyond*, ed. A. K. Sen and B. Williams. Cambridge University Press, Cambridge, pp. 39–62. Originally published in *Social Research* 44 (1977), 623–656.

Harsanyi, J. C. 1985a. Does reason tell us what moral code to follow and, indeed, to follow any moral code at all? *Ethics* 96, 42–55.

Harsanyi, J. C. 1985b. On preferences, promises and the coordination problem. *Ethics* 96, 68–73.

Harsanyi, J. C. 1992. Game and decision theoretic models in ethics. In *Handbook of Game Theory*, Vol. 1, ed. R. J. Aumann and S. Hart. North-Holland, Amsterdam, pp. 669–707.

Harsanyi, J. C. 2008. John Rawls's theory of justice: Some critical comments. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls,* ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 71–79.

Hohfeld, W. N. 1919. *Fundamental Legal Conceptions as Applied in Judicial Reasoning*, ed. W. W. Cook. Yale University Press, New Haven, CT.

Mill, J. S. 1859. *On Liberty*. Parker, London.

Mill, J. S. 1861. Utilitarianism. *Fraser's Magazine* 64, 391–406, 525–534, 658–673.

Pattanaik, P., and Suzumura, K. 1996. Individual rights and social evaluation: A conceptual framework. *Oxford Economic Papers* 48, 194–212.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge, MA, rev. ed., 1999.

Rawls, J. 1985. Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs* 14, 223–251.

Rawls, J. 1993. *Political Liberalism*. Columbia University Press, New York, expanded ed. 2002.

Riley, J. 1988. *Liberal Utilitarianism*. Cambridge University Press, Cambridge.

Riley, J. 1989–90. Rights to liberty in purely private matters, Parts I & II. *Economics and Philosophy* 5–6, 121–166, 27–64.

Riley, J. 1998a. *Mill: On Liberty*. Routledge, London.

Riley, J. 1998b. Mill on justice. In *Social Justice: From Hume to Walzer*, ed. D. Boucher and P. J. Kelly. Routledge, London, pp. 45–66.

Riley, J. 2000. Defending rule utilitarianism. In *Morality, Rules, and Consequences*, ed. B. Hooker, E. Mason, and D. Miller. Edinburgh University Press, Edinburgh, pp. 40–70.

Riley, J. 2003. Interpreting Mill's qualitative hedonism. *Philosophical Quarterly* 53, 410–418.

Riley, J. 2006a. Liberal rights in a Pareto-optimal code. *Utilitas* 18, 61–79.

Riley, J. 2006b. Genes, memes, and justice. *Analyse & Kritik* 28, 32–56.

Riley, J. 2007a. Mill's neo-Athenian model of liberal democracy. In *J. S. Mill's Political Thought: A Bicentennial Reassessment*, ed. N. Urbinati and A. Zakaras. Cambridge University Press, Cambridge, pp. 221–249.

Riley, J. 2007b. *Mill's Radical Liberalism*. Routledge, London.

Riley, J. 2007c. Justice as higher pleasure. In *Mill: Bicentennial Essays*, ed. P. J. Kelly and G. Varouxakis. Cambridge University Press, Cambridge.

Sen, A. K. 1980. Plural utility. *Proceedings of the Aristotelian Society* 81, 193–215.

Sen, A. K. 1992. Minimal liberty. *Economica* 59, 139–159.

Sen, A. K. 2002. *Rationality and Freedom*. Harvard University Press, Cambridge, MA.

Sidgwick, H. 1907. *The Methods of Ethics*, 7th ed. Macmillan, London. Originally published 1874.

Weymark, J. A. 1991. A reconsideration of the Harsanyi-Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. Roemer. Cambridge University Press, Cambridge, pp. 255–320.

Index