Utility theory and ethics *

Philippe Mongin† and Claude d'Aspremont‡

Contents

1 Introduction **1**

2 Some philosophical and historical clarifications **4**

2.1 Preliminaries 4

2.2 Early utilitarian views: Utility related to pleasure and pain 8

2.3 Utility as a measure of actual preference satisfaction 12

2.4 Utility and well-being: critical arguments 19

2.5 Utility, well-being, and social ethics: some positive arguments 26

3 Some definitions and concepts from utility theory **34**

3.1 Utility functions in the case of certainty, and "economic" domains 35

3.2 Von Neumann-Morgenstern utility functions 36

3.3 Anscombe-Aumann utility functions 37

3.4 Interpersonal utility differences 42

4 The aggregative setting with interpersonal comparisons of utility **43**

4.1 The SWFL framework 44

*Reprinted from *Handbook of Utility Theory. Volume 1: Principles*, S. Barbera, P.J. Hammond and C. Seidl (eds.). Dordrecht, Kluwer Academic Publishers, 371–481, 1998.

†THEMA, Centre National de la Recherche Scientifique and Université de Cergy-Pontoise

‡CORE , Université catholique de Louvain, Louvain-la-Neuve, Belgium

4.2 Invariance axioms and interpersonal comparisons of utility 46
4.3 Further conditions on SWFL 48
4.4 Utilitarianism versus leximin 50
4.5 Further rules for social evaluation. The variable population case 54
4.6 Some conceptual problems of the multi-profile approach 58

5 The aggregative setting with choice-theoretic constraints 61
5.1 Harsanyi's approach to utilitarianism. The aggregation theorem 61
5.2 A SWFL reconstruction of Harsanyi's aggregation theorem 66
5.3 Further philosophical comments on Harsanyi's utilitarianism 68
5.4 The aggregative approach in the case of subjective uncertainty 74

6 The impartial observer, the original position, and fairness 84
6.1 Impartial-observer theories 84
6.2 State-of-nature theories and the "original position" 87
6.3 Harsanyi's impartial observer theorem, and the problem of "extended sympathy" 89
6.4 Rawls's "Original Position" and "Veil of Ignorance" 97
6.5 Alternative notions of fairness 101
6.6 Equality of Resources and Welfare 105

7 Concluding comments 108

1 Introduction

The technical sense of "utility", as in "utility theory", can be traced back to the work of 18th century British empiricists and, of course, of 19th century utilitarians. These writers began to use the word in a sense different from the common sense meaning of "usefulness" but did not agree upon a clear-cut explicit or even implicit definition. The emergence of Paretian microeconomics in the first half of the 20th century, and the later developments of utility theory as a quasi-separate field of study, brought with them further connotations, so that despite its short history of technical use, the world "utility" is now richly ambiguous. As a result, the relevance of utility theory to ethics cannot be taken for granted without prior clarification.

Section 2 of this chapter attempts to provide the needed clarification. We begin by distinguishing two received notions of "utility" in the technical context: (i) pleasure and pain, and (ii) the satisfaction of the individual's *actual* preferences. We then enquire whether utility can represent more generally: (iii) the individual's well-being. All of the three interpretations of utility have been endorsed, at least implicitly, by some ethical theories. Today's predominant philosophical view is that (i) and (ii) are not relevant in an ethical context of application. We shall restate this critique, and examine whether it leaves room for an ethical application of the utility concept. If utility can fully represent (iii), it becomes a relevant concept to use by all those ethical systems in which well-being is the underlying notion of good. We shall emphasize one particular strategy to confirm that utility can indeed represent well-being. It consists in interpreting utility as measuring: (iv) the satisfaction of *rational and well-informed* preferences, and then arguing that this interpretation essentially coincides with (iii).

Once the semantic ground has been cleared, it becomes possible to discuss various applications of utility theory. Section 3 introduces definitions and notation. Sections 4, 5 and 6 then investigate the ethical consequences of imposing various axiomatic restrictions on the utility concept, such as the von Neuman-Morgenstern (VNM) axioms of risky choice, and the Savage (or related) axioms of uncertain choice. The importance of these standard constructions for investigating ethical theories can be defended most easily on the background of the last-mentioned

interpretation – i.e., (iv) above. If one insists on an ethically relevant notion of utility that makes crucial reference to the rational formation of preferences, one is naturally led to consider the ethical consequences of adopting the definitions of rationality provided by utility theory.

The present survey will be essentially restricted to the better known parts of the theory. Having adopted the classic distinction between three contexts of rational choice – i.e., certainty, risk, and uncertainty – we shall apply the following standard constructions to them: the theory of choice under certainty, as in consumer microeconomics; von Neumann-Morgenstern expected utility theory; and subjective expected utility theory, respectively.

This chapter will also be narrowed down for a different and perhaps less defensible reason. Except in the rather general Section 2, we shall most of the time restrict the problems of ethics to those of the proper foundations of collective life. We are keenly aware that a majority of ethical systems would reject this identification – Bentham's utilitarianism being arguably and exception. To mention a few standard doctrines, hedonism, perfectionism, eudemonism, and Kantianism are conceptions of *both* individual and social ethics. Accordingly, the connection between utility theory and ethics should be explored also at a strictly individual level. For example, the famous and unresolved question of whether or not individuals should entertain time-preferences can be interpreted as being, or at least involving, an ethical question, and has been discussed as such among utilitarian circles [see Parfit (1984)]. The content of individual preferences – e.g., time-preferences and egoism – as well as the various formal constraints that have been imposed on them – e.g., the maximization principle or the sure – thing postulate – raise significant ethical questions. These issues will hardly be addressed here. What will occupy the forefront are the issues of distributive justice, and more generally those raised by *the evaluation of social states of affairs*. The expression "social ethics" provides a loose delineation of our object matter.[1] In sum, we are concerned with the logic and normative strength of a particular class of value judgements. For expository purposes, it is convenient to attribute these judgements to some ideal "social evaluator" (or "observer") who may or may not be also a

[1]Some writers would use "morality" to refer to that part of ethics which is "other-regarding" (as against "self-regarding"). We do not adopt this terminology here, and use "moral" and "ethical" more or less interchangeably.

"social planner". This ambiguity is typical of welfare economics and social choice theory, on which we heavily rely in this chapter. These two fields deliver not just technical explorations of normative principles, but also broad guidelines for public decisions; they connect with both political philosophy and prescriptive economics. More often than not, the theories presented below do not resolve the ambiguity between the "evaluator" and the "planner".

Even granting the scope restrictions just discussed, the ethical connections of utility theory can be appraised in a variety of ways. This chapter will emphasize the distinction between the *aggregative setting*, on the one hand, and the *original position* and *impartial observer* devices, on the other. In either case, the aim is to determine the nature of the ethical observer's evaluation rule from (among other things) antecedent rationality constraints that have been imposed on the individuals's preferences or utility functions. An important point for discussion within either framework is whether the ethical observer himself should be subjected to identical rationality constraints. In the aggregative setting, which is the topic of Sections 4 and 5, one or more axioms determine the functional relations holding between the individuals' relevant characteristics and the observer's. Section 4 reviews some ethically relevant concepts and results in the theory of "social welfare functionals" (SWFL), as expounded by Sen and others in 1970–1980. Section 5 elaborates on Harsanyi's Aggregation Theorem, the *ex ante* versus *ex post* debate in welfare economics, and the many results connected with these topics. Section 6 moves on to original position and impartial observer theories. The general device investigated in this section consists in identifying the ethical observer's preference, or utility, with those of the individuals when the latter are deprived of part of their information. Harsanyi's Impartial Observer Theorem and Rawls's "veil of ignorance" construction are the key references here. These and related theories provide an indirect derivation of ethical rules which is in some sense more constructive and concrete than the aggregation procedures of Sections 4 and 5. Both the aggregative and original position or impartial observer types of analysis have received extensive treatment in the literature, so that we might perhaps be excused for not offering an exhaustive survey.[2] As

[2]The recent years have witnessed book-length presentations of collective choice theory and normative economics broadly speaking. To date, Fleurbaey's (1995) is the most comprehensive. Binmore's (1994), Kolm's (1995),

to the distinction between the aggregative theories of Section 4 and Section 5, respectively, it depends in part on the chosen conceptual and technical frameworks, in part on the underlying information context (i.e., pure certainty in the former, and risk or uncertainty in the latter).

Most of the work surveyed in this chapter involves the philosophically questionable thesis of *welfarism* – i.e., that individual utility values contain all the information required to derive collective evaluation rules. This thesis has come under increasing criticism in recent years, notably in the wake of Rawls and Sen. New constructions have emerged that either make no reference at all to the utility concept or (more usually) employ it just as a subordinate device. We shall echo this recent debate in virtually every section of this chapter but more particularly in Section 2, which states a qualified defence of welfarism.

2 Some philosophical and historical clarifications

2.1 Preliminaries

This section surveys several influential interpretations of the utility concept, in order to see whether the latter can be made relevant to social ethics. The discussion will be organized around three notions of utility. The first is utility as related to the individual's pleasure and pain, a notion introduced by 19th century utilitarians (see Section 2.2). The second notion is utility as a measure of the individual's welfare or well-being, a third and more general interpretation which is arguably more relevant (see Section 2.4). To prepare the ground for the discussion, some philosophical terminology must first be introduced.

We need the classic distinction between two types of ethical theories, *teleological* theories on the one hand, and *deontological* ones on the other. This distinction is usually understood in terms of the two concepts that are perhaps most basic to ethical reasoning at large, i.e., the *right* and the *good*. In teleological theories the good has conceptual priority over the right. Typically, they define what it means for a thing or a state of affairs to be good, and then

Moulin's (1988, 1995) and Roemer's (1996) also contain useful material. The philosophical background of this chapter is covered in Singer's (1991) and Canto-Sperber's (1996) encyclopaedias of ethics.

what it means for an act or a life to be morally right. As Rawls (1971, p. 25) insists, a crucial point is that "in a teleological theory, the good is defined *independently* from the right". A possibly no less crucial point is that in such a theory, no considerations other than the good contribute to the definition of the right, i.e., the latter can be *entirely* derived in terms of the former. Consequently, nonteleological theories are exactly those which either dispute that there can be an ethically relevant independent definition of the good, or else do not dispute that but introduce considerations other than the good in order to determine their concept of rightness. Nonteleological ethical doctrines are commonly referred to as deontological. This terminology conveys the fact that nonteleological ethics generally emphasizes the notion of a duty or a moral obligation.

Since at least Kant's (1785, 2nd section) classic discussion, the example of promise-keeping has been used to illustrate the contrast between deontological and teleological views of morality. Following a typically deontological analysis, when we say that it is wrong to break promises, we mean that it is *intrinsically* wrong to do so. This moral judgement derives – or so the argument goes – from the correct understanding of the rightness/wrongness distinction. It is supposed to be independent of whether or not promise-keeping is a good thing (and in particular, of whether or not it leads to good consequences). Following the equally banal, though opposite interpretation, which is held by many teleologists, we predicate that the action of promise-keeping is morally right just because it is good for society under normal circumstances. Note emphatically that the conflict between teleological and nonteleological views is compatible with the endorsement of one and the same maxim of action.[3] Here, as in many other relevant applications, ethical disagreements relate to the proper way of founding commonsense morality. It is however not difficult to conceive of particular cases in which deontologically oriented and teleologically oriented moralists would give conflicting recommendations. The most famous representative of the former school, Kant, insists that promises should be kept in all and every circumstances, whereas moralists of the latter school typically make exceptions to the rule (i.e.,

[3] "Maxim of action" will be used here in the Kantian sense of the subjective principle underlying the action; see the *Critique of Practical Reason* (1788) and the comments by Verneaux (1973, 2, pp. 177–178).

whenever the consequences of following the rule involve a net harm rather than net benefit).

Teological ethics is best represented by utilitarianism, a doctrine which will receive much attention in this chapter. Roughly speaking, classical utilitarianism is associated with the views that: (i) pleasure is the relevant concept of the individual's good; (ii) the right action is that which maximizes the total sum of individual amounts of good. By and large, contemporary utilitarianism has faithfully maintained principle (ii), while often rejecting (i) in favour of alternative definitions of the individual's good. Adherence to (ii) should be qualified in view of the distinction between *act-utilitarianism* and *rule-utilitarianism*. Strictly speaking, (ii) is the act-utilitarian's maxim, whereas the rule-utilitarian's would read as follows: (ii') the right action is that which follows from the rule maximizing the total sum of individual amounts of good.[4] Utilitarianism is a pervasive doctrine, but teleological ethics has a much wider coverage as well as deeper historical roots. Plato's *Republic*, and more strikingly, Aristotle's *Nichmachean Ethics* count as the historical sources of teleological ethics at large.

Kant's practical philosophy is the classic example of deontological ethics. This tradition too has deep historical roots: Kantianism has often been described as a systematic and laicized version of Christian ethics.[5] Following the *Groundwork of the Metaphysics of Morals* (1785), the defining conditions for the right action are, for one, that it accords to duty, and for another, it is taken for no other reason than its conformity with duty (it is performed "out of duty"). For instance, if I keep my promise because I feel that false promising would have disadvantageous consequences, my action satisfies the former condition, but not the latter, and thus does not count as a right action. On Kant's view, then, the nature of motives is crucial to the moral judgement. Since outward behaviour is uninformative, and motives are largely inscrutable, it

[4]This classic distinction permeates the utilitarian tradition. Harrod (1936) and Harsanyi (1977c) endorse rule utilitarianism. Brandt (1958, 1992), Lyons (1965), and Hare (1981) have discussed it. Smart (1973) dismisses rule utilitarianism by claiming that when properly understood, it reduces to act utilitarianism.

[5]Kant's Categorical Imperative, to be discussed below, is distinct from, but definitely related to the Golden Rule of Jewish and Christian Ethics: "What you dislike don't do to others; that is the whole Torah" [B.T. Schabbat 31a, cited in Singer ed. (1991, p. 87)], "Always treat others as you would like them to treat you" (Matt. 7:12).

must be extremely difficult to ascertain whether or not a particular act is right. In the *Critique of Practical Reason* (1788) Kant himself had to recognize that perhaps not any single right action had ever been performed since the beginning of mankind, and many critics have eventually rejected his ethical system as being aloof from human concerns and virtually impracticable. At any rate, Kant's approach very clearly illustrates the priority of the concepts of right and duty in nonteleological ethics.

Some of Kant's conceptions have proved remarkably influential, even among non-Kantian philosophers.[6] The *Groundwork* is justly famous for introducing the following test for something to count as a moral principle: the maxim of an action is a moral principle only if it is *universalizable*. This condition is contained in the so-called Categorical Imperative: "Act only on the maxim through which you can at the same time will that it be a universal law". To elaborate again on the previous example, actions resulting from the maxim of promising falsely fail the universality test, because it is impossible to "will this maxim as a universal law". To break promises systematically has the effect of destroying trust, without which the very notion of promise becomes meaningless. Crucially, the Categorical Imperative does not only demand that we formulate some universal principle under which the particular act-description can be subsumed, but also that the "covering" principle be, for one, free from any contradiction, and for another, related to the agent's actual motive. We will mention examples of non-Kantian theories in which motives play no role but at least the non-contradiction requirement is kept, and even strongly emphasized (see 6.1 on Hare's "universal prescriptivism" and Harsanyi's reinterpretation of Kant). The *Groundwork* also states the Categorical Imperative in the following alternative form: "treat humanity in your own person, or in the person of any other, never simply as a means but always at the same time as an end". This Formula of the End in Itself makes respect (for oneself as well as others) a condition for the right action. Like the above mentioned Formula of Universal Law, it has exerted considerable influence. It indirectly suggests an appeal

[6]This paragraph discusses only the influence of Kant's ethics. Picavet (1996) has recently uncovered another connection between Kant's philosophy and the subject matter of this chapter. He argues that several important themes in Bayesian decision theory are foreshadowed in the first *Critique*.

to consent on which some contemporary theories are explicitly based (see Sections 6.1 and 6.2).

There are modern examples of theories in social ethics which are deontological, or at least clearly nonteleological. Rawls repeatedly claims that his own doctrine, "justice as fairness", prioritizes the right over the good (e.g., 1971, p. 451). Accordingly, it should be counted as a nonteleological doctrine. This holds regardless of the fact that the notion of the individual's good, as defined by his own ends and pursuits, is pervasive in Rawls's *Theory of Justice*. As a further illustration, Nozick's (1974, pp. 28–29) conception of "moral constraints" versus "moral goals" is best ranked among the deontological theories. After many others, Nozick is particularly concerned with protecting the exercise of *rights*, in the sense of the individual's rights to property and other "natural rights". A case can be, and was indeed sometimes, made for individual rights along the teleological line that they serve the collective good. But this is not the deontologist's argument. Rather, he would claim that rights are primitive concepts, and their recognition is to be viewed as a "side-constraint" (Nozick's expression) on any social arrangements.[7]

Utility theory plays a more important role in the context of teleological than of nonteleological ethics. This privileged connection emerges from the early developments of the field, since the rudiments of modern utility theory date back to classical utilitarianism.

2.2 Early utilitarian views: Utility related to pleasure and pain

In daily parlance "utility" and "usefulness" are near synonyms.[8] Thus, "utility" refers to the property of actually serving, or of being able to serve, an end or purpose. Like "usefulness", "utility" is normally predicated of concrete things; but it can also be employed for persons, actions or states of affairs whenever they are envisaged from the viewpoint of an end they serve. When we say that a thing is useful, we presumably mean that it does not only serve an idle

[7] Sen (1982) discusses Nozick's notion of right as side-constraints while defensing his own conception of rights.

[8] See for instance *Webster's New Dictionary of Synonyms.* Importantly, English is exceptional in having two words. Both "utility" and "usefulness" are translated into a single word in French, German or Italian ("utilité", "Nützlichkeit", "utilità"). This linguistic fact has attracted attention since the early days of utilitarianism. The duality of "utility" and "usefulness" in ordinary English has no doubt facilitated the separation of technical and nontechnical uses of the utility concept, but it has also created a semantic problem of its own.

purpose: the commonsense notion of utility seems to imply the view that some ends are relevant and some are not. Also, the property referred to as "utility", in the sense of usefulness, often conflicts with pleasurableness. This much is implied by the French cliché: "joindre l'utile à l'agréable".

By and large, the commonsense meaning of utility prevailed in ethical and political philosophy until utilitarians began to employ this word in a special sense. Bentham's early writings, such as the *Introduction to the Principle of Morals and Legislation* (1789), signal a major change in use. We shall argue that early utilitarianism manifests a crucial shift in emphasis – since with Bentham, utility becomes the foundation of a whole system – but only a partial shift in meaning.

Here is the beginning of the *Introduction*:

> "Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne ... The *principle of utility* recognises this subjection, and assumes it for the foundation of that system, the object of which is to rear the fabric of felicity by the hands of reason and of law" (1789, pp. 1–2).

Since pleasure and utility are normally viewed as noncoincident, and even divergent concepts, there is a definite semantic shift in these famous lines. But in another crucial respect, Bentham does *not* depart from the commonsense meaning of "utility": he employs it to refer to a particular instrumental property of things – i.e., that they serve the purpose of producing pleasure or avoiding pain. Consider Bentham's definition in the same passage (ib., p. 2): "By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness". This and many other supporting passages show that Bentham employs "utility" in the sense of "usefulness", but with a twist.[9] Contrary to the popular reconstruction among historians of economics, he does not identify "utility" and "pleasure" directly with each

[9]The present interpretation follows Broome's (1991b). See also Little (1950, p. 7): "Utility was a power in objects which would normally create satisfaction".

other. These words refer to mutually related but distinct concepts: a subjective feeling, on the one hand; a property of things, acts, or states of affairs, on the other.

Whenever utility is conceived of as a property of things, it becomes an *objective* concept in the various accepted senses of this word.[10] Thus understood, it has to do with the relation of man in general – instead of some particular individual – to external objects. It would then hardly make sense to predicate utility of people, as in today's economists' phrase: "Given a consumer x endowed with a utility function ...". It follows that the problem of interpersonal comparisons of utility, which will be an acute one in the later subjective value theories, hardly arises in Bentham's framework. To illustrate this, consider the passage from the *Pannomial Fragments*,[11] in which he anticipates the "law of diminishing marginal utility". This passage shows that Bentham understands functional reasoning, as well as the mathematical property of a decreasing first derivative. However, Bentham's function here is not a utility function in the sense of subjective value theories. Rather, it is a money-to-pleasure mapping, and crucially, this mapping is *not* indexed by individuals.

No doubt, the misunderstanding that Bentham's utility *is* pleasure minus pain was fostered by the fact that he endows the pleasure and pain concepts with a rich numerical structure, as in contemporary utility theories. Here again, however, one should be careful to avoid retrospective misunderstandings. Schumpeter's (1954, p. 409) claim that Bentham's "felicitic calculus" lays down the essentials of 20th-century value theory is inaccurate. For one thing, Bentham explicates pleasure and pain in terms of several (viz., seven) dimensions or "circumstances" (1789, Chapter 4). He borrows from Beccaria's *Dei delitti e delle pene* (1764) the basic distinction between intensity, duration, certainty, and proximity, and he complicates it with a further distinction of his own between fecundity, purity, and extent (i.e., the number of persons to whom the pleasure or pain extends). Bentham normally does not assume that there is a complete system of exchange rates between them. That is to say, his analysis of pleasure and pain remains truly

[10] This paragraph and the next are based on Mongin (1995b).

[11] In Bowring (1838, vol. 33, pp. 228–229). For a discussion of this important passage, see Halévy (1901–1904, I, pp. 83–84) and Stigler (1965, Chapter 5).

pluridimensional and partly qualitative. For another thing, Bentham repeatedly emphasizes that material gains have different psychological consequences from material losses.[12] To formalize this claim, one could possibly resort to the following nonstandard assumption in the style of Kahneman and Tversky (1979): the net pleasure aggregate is a function of net variations in, rather than absolute levels of, current wealth, and this function exhibits different concavity properties on gains and losses.

To investigate the later utilitarians' work in any detail would lead beyond the scope of this chapter. Suffice it to say that their analysis manifests a general trend towards simplification. The semantic shift in the use of the word "utility" is already complete in the following excerpt from John Stuart Mill:

> "Those who know anything about the matter are aware that every writer, from Epicurus to Bentham, who maintained the theory of utility, meant by it, not something to be contradistinguished from pleasure, but *pleasure itself, together with exemption from pain*" (1861, Chapter 2, p. 5, our emphasis).

Even more obviously, in Jevons (1871) and Edgeworth (1881), "utility" stands for "pleasure" rather than for the "tendency of objects to produce pleasure". Contemporary readers have become so accustomed to this definition, even to reject it, that they might overlook the gap between them and the early utilitarians' still colloquial understanding of utility.

A further accompanying simplification in the late utilitarians' work is related to the metric of pleasure and pain. In fact, Bentham never consistently used his distinction between seven "circumstances". Most of his reasonings in penal theory can be reconstructed by restricting attention to intensity, duration, certainty and proximity (the two of which can be identified with each other), and extent. Even this simplified list has its conceptual problems; along with the search for tractable physical analogies, this might explain why Jevons and Edgeworth just

[12]This insight played a crucial role in the early utilitarians' conservative assessment of property rights and income distribution. On the political economy of early utilitarianism, see the classic discussions by Stephen (1901), Halévy (1901-1904), and Viner (1949).

retained the two dimensions of "intensity" and "time". Moreover, late utilitarians lost interest in Bentham's insight that pain and pleasure had distinctive measurement properties.[13]

2.3 Utility as a measure of actual preference satisfaction

Following the most popular interpretation among 20th century writers, utility is a measure of actual preference satisfaction. "Actual" is meant to contrast the individual's preference underlying his behaviour with his rationally formed preferences. This interpretation underlies standard texts in economic theory, and pervades other social sciences as well as philosophy. Most of the time, it seems to be regarded as a stipulation rather than a substantial claim, and it is not stated very clearly. It can be made precise in a number of different ways but the crucial point that all these formulations share is this: the utility of a thing or an action reflects the extent to which that thing or action is preferred to others, and has no meaning beyond that. Thus, the modern technical sense of "utility" not only excludes the commonsense notion of utility as usefulness, but also supersedes the old technical sense of utility as being related to pleasure and pain. More generally, the modern sense conflicts with any conception which would take the utility concept as primitive. This rejection of established interpretations has led some to suggest that "utility", as in "utility theory" or "utility function", is just a misnomer, a confusing remainder from 19th century economics. What is important is that numerical values can be attached to objects in a way expressive of the individual's preferences between these objects. What matters in "utility function" is the second word, not the first.

Since it is claimed to be the true primitive, the notion of preference is in need of explication. In ordinary parlance, the word refers to a wide range of subjective comparisons between objects: I prefer Burgundy to Bordeaux wines, a studio in Paris to a house in Paris (Texas), an adventurous to a quiet life, peace to war, etc. My tastes, as well as my goals, interests and values, might

[13] J.S. Mill's multi-dimensional conception of pleasure (and derivatively, utility) strongly departs from Bentham's. Mill is famous for claiming that pleasures have different *kinds*, as well as different *intrinsic values*, a conception which has often been rebutted as being inconsistent with utilitarianism. On this issue, see, e.g., Riley (1988). Sen (1980-81) compares different senses in which utility can be said to be "plural".

contribute to explaining why I prefer x to y. Nothing in the ordinary linguistic use appears to exclude that I prefer x to y from one point of view (for instance my tastes today), and y to x from another point of view (for instance my tastes tomorrow, or my permanent values). Although this has been suggested by some philosophers, it does not seem to be a matter of definition that my preferences comply with a global structure. Notice also that the objects that are compared with each other can be of many kinds. I can entertain preferences over states of affairs and things, as well as over actions. Hence, there is no privileged connection between preferring and choosing, at least if one understands choice as action rather than just as a determination to act.

Here as elsewhere, economists have shifted the ordinary meaning of words, and to some extent have become unaware of the shift. For one, today's economists put in it more structure than is normally implied. They assume that there is a preference *scale* underlying (and perhaps causing) the individual's preferential comparisons. What this scale exactly consists of is a matter for technical investigations, but all economists agree that it is structured and enjoys some form of permanence. Significantly, "the agent's preferences" is increasingly used to abbreviate "the agent's preference scale", i.e. the same word is used to refer both to the comparisons and their explanatory factor. For another, preferences in the economists' sense have to do with *choices between objects or actions*; they would not normally consider preferences between states of affairs in general. This double change is apparent in the following excerpt from Hicks (1956, pp. 17–18):

> "We have to make some assumption about the principles governing (the consumer's) behaviour. The assumption of behaviour according to a scale of preferences comes in here as the simplest hypothesis (...) What I mean by action according to a scale of preferences is the following. The ideal consumer (...) chooses that alternative, out of the various alternatives open to him, which he most prefers, or ranks most highly. In one set of market conditions he makes one choice, in others other choices; but the choices he makes always express the same ordering, and must therefore be consistent with one another."[14]

[14] At least, Hicks is careful enough not to conflate "preference" with "scale of preference". Some of the philosophers' technical elaborations of the preference concept are more faithful to ordinary language connotations than

The notion of preference exemplified by this passage goes along with the methodological conviction that the theoretician should not take a position on the content of preferences, but only on their formal or structural properties. Another deep-seated conviction among 20th century economists is that preferences are specific to the particular individual, so that interpersonal comparisons appear to be inherently problematic. In sum, the standard doctrine of preference is at the same time *formalistic* and *realistic*. It is against this conceptual background that one should appreciate the polemics against utilitarianism that are typical of early expositions of modern economics. Since preferential behaviour, in the formal conception, does not have to be pleasure-oriented, the utility theory of Bentham and his followers has been rejected as being all too specific. Also, the ease with which 19th century utilitarians take for granted that preference intensities are identical from one individual to another has aroused puzzlement and irritation, given the prevailing relativism. Among others, Robbins (1932) and Schumpeter (1954) are emphatic on the rejection of utilitarianism, both at the individual and the collective levels.[15]

The preference satisfaction view of utility can be traced back to Pareto. In the *Cours d'économie politique* (1896-7) and the *Manuel* (1909) he made it clear that utility is a formal concept, although he expressed this in a psychologistic language which sounds old-fashioned to today's readers. The most important feature of his conception is that he does not regard utility as being *itself* a feeling (or any mental state): utility is one step remote from its psychological substractum, *la sensation*. Pareto was equally clear about the relativistic nature of the utility concept. The *Manuel* introduces the celebrated notion of an optimum in a purely technical way, leaving the philosophical interpretations open.[16] But it distinguishes clearly between intra-

are the economists'. For instance, Jeffrey's (1965) system defines preferences and utility over *logical propositions*, and is therefore capable of encompassing preferences, typically over states of affairs, which are unrelated to choices.

[15]The former insisted that utility theory should be rid of "the accidental deposit of the historical association of Economics with Utilitarianism" (1932, p. 141). The latter somewhat aggressively condemned "the unholy alliance between economics and Benthamite philosophy" (1954, p. 831). Similar complaints underlie the contributions of the "new welfare economics" in the 1930's. Most of the writings of the time betray a confusion between the *utilitarian type* of interpersonal comparisons of utility and the *general* possibility of making such comparisons.

[16]See Appendix 89 on *Maximum d'ophélimité* (1909, pp. 617–619). This primarily mathematical passage is compatible with virtually any conception of utility, which might have been a source of conceptual misunderstand-

and interpersonal comparisons of satisfactions, and emphasizes that the latter belong to an underdeveloped area of social sciences [see (1909, p. 149)]. The Paretian conception is perhaps most clearly stated in the passage in which he contrasts *utilité* in the ordinary sense with *utilité économique* (or *ophélimité*). The former, it is claimed, is concerned with the relation of mankind to external objects and has therefore an objective status; the latter is concerned with the particular individual under consideration and is therefore irreparably subjective (1909, p. 157).

Pareto makes limited use of the preference concept.[17] It was left for later writers of his school first to clarify, second to formalize it, as the structural feature that accounts for the individual's choice activity. Hicks's prewar studies of consumer theory and his *Value and Capital* (1939) might signal its first systematic occurrence as an *explicans* of utility. At a later stage, Arrow's *Social Choice* and *Individual Values* (1951) popularized the formalism of binary relations, and it is probably only then that preference was fully recognized as a distinctive technical entity. The fact that preference can be endowed with a mathematical structure no less precise than that of utility has had far-reaching consequences on developments in the field. Today's utility theory is essentially concerned with establishing *representation theorems*, i.e., equivalences that connect various properties of the preference relation with numerical properties of the "representing" utility functions.[18] The technical turn of the theory approximately coincided with the conquest of its autonomy. The rather large body of representation theorems and related results that have

ings. In his *Traité de sociologie générale* (1917–1919) Pareto is more explicit about the meaning of a collective optimum.

[17]Texts in history of economics perhaps underrate this fact. In the next paragraph our (all too brief) account departs from standard historiography on another score: we emphasize that the modern conception of preference is more general than both "ordinalism" and revealed preference theory. Compare, with, e.g., Blaug (1980), or Screpanti and Zamagni (1993).

[18]The abstract structure of representation theorems is investigated in Krantz, Luce, Suppes and Tversky (1971). Under the name "measurement theory", these authors have developed mathematical tools to analyze the relations holding between qualitative (i.e., set-theoretic) structures and the numbers that represent them. The resulting algebra can be applied to problems in physical measurement as well as to the representation theorems that are specific to utility theory. For an introduction, see Suppes (1981).

become available by now goes far beyond the theoretical needs of microeconomics. So it is more appropriately referred to as *decision* or *choice theory*, a relabelling which conveys the broad scope of application and avoids the misleading allusion to utility.[19]

Before closing this retrospective we need to emphasize that the actual preference satisfaction interpretation of utility is more general than two views with which it has often been associated: (i) the view that utility is a purely ordinal concept, and (ii) revealed preference theory. Concerning (i), the Paretians are famous for being critical not only of the strong utilitarian assumption that utility differences are *inter*personally comparable, but also of the weaker assumption that they are *intra*personally meaningful. Definitionally, this amounts to denying that utility functions can be "cardinal". All the information they convey is about the ordering of alternatives: they are just "ordinal". This conclusion was endorsed by Arrow in his 1951 book; it is still very influential among economists. It should be clear that its conflation with the view that utility measures actual preference satisfaction is a matter of historical coincidence, not of logic. It is perfectly possible to adhere to the latter view while claiming a cardinal interpretation for the utility function. This position is perhaps best understood in terms of Suppes and Winet's (1955) axiomatic construction. These authors state axiomatic conditions on preference relations that give a meaning to the notion of preference differences (e.g., I prefer x to y more strongly than I prefer w to z), and then show that these conditions are equivalent to the existence of a cardinal utility representation. Prominent writers like Allais (1953) [see also Allais and Hagen (1994)] or Harsanyi (1977b) appear to adhere to both a cardinal interpretation *and* the standard representation-of-preference view of utility. Their conception seems to be implicitly grounded in the Suppes-Winet axiomatization. Section 5 will further elaborate on Harsanyi's cardinalism.[20]

A somewhat related warning applies to (ii). As expounded by Samuelson (1938), revealed preference theory is an attempt to extract the empirically testable content of Paretian consumer

[19] Arrow [e.g. (1984, 3, p. 56)] discusses this terminological change.

[20] The conception of cardinal *preference* suggested here should be contrasted with the following other two positions: (i) cardinal *utility* makes sense but cardinal preference does not, so that cardinalism can be defined only if utility is the primitive concept; (ii) neither cardinal preference nor cardinal utility makes sense. As an example of (i), see Loomes and Sugden (1982). Position (ii) is the more common of the two.

theory and axiomatize it in terms of relevant observable concepts. The (supposedly observable) "revealed preference relation" is the building block in Samuelson's reconstruction. More broadly, and somewhat loosely, revealed preference theory is the methodological claim that the preference concept receives its meaning from, and its completely expressed in, the agent's choices between objects. The standard conception, as exemplified above by Hicks, says that preference is the factor underlying the individual's actual choices. This is an altogether different claim from that of revealed preference theory. Actually, the two claims are not only different but openly conflict with each other. In the standard conception, preference has an existence of its own and assumes conceptual (and perhaps causal) priority relative to choice; in revealed preference theory, only choice exists in a substantial sense, and preference is just an abbreviative concept for the latter. Again, it is a matter of coincidence if some writers on economics or choice theory make the simultaneous claims that utility represents the individual's actual preferences *and* that these preferences are just another name for the individual's choices. Evan when it is assumed that actions are the only objects of preference, which we said earlier is a significant restriction, the former claim does *not* logically imply the latter.[21]

It is important to realize that actual preference satisfaction is the notion of utility underlying contemporary welfare economics. The object of welfare economics is to rank economic states of affairs, in particular those involving distributional consequences, public goods, and state provisions, in terms of "better" or "worse". To do so, the discipline needs a concept of good: officially, it is "welfare" or "well-being", presumably in some appropriately restricted interpretation. Since Pigou's pioneering *Economics of Welfare* (1920), specialists in the field have generally emphasized that they were concerned only with "economic" well-being. Abstracting from the (non-trivial) problem created by this restriction, it would appear that welfare economists should interpret utility functions as measuring individual welfare or well-being (in this chapter we shall not attempt to distinguish between these two words). Some important papers in the field sug-

[21]The technical side and analytical history of revealed preference theory are covered in Chipman et al. (1971). Sen (1973a) provides a thorough critical discussion of revealed preference theory in the broad methodological sense.

gest that this is indeed the case: for instance, Lange's (1942) classic piece on the computation of Pareto optima.[22] But even if individual welfare or well-being is theoretically at the centre of welfare economists' concerns, they are nearly silent about what this notion consists of. At least after Paretian economics established its grip on the field, the definition of utility effectively used in welfare economics, as against the official definition, went in terms of actual preference satisfaction – i.e. the very same definition as that which underlies *positive* theorizing.

A respected textbook in its time, Graaff's, resolved the tension between the two notions of utility by identifying them *by way of stipulation:*

> "a person's welfare map is defined to be identical with his preference map – which indicates how he would choose between different situations, if he were given the opportunity for choice. To say that his welfare would be higher in A than in B is thus no more than to say that he would choose A rather than B, if he were allowed to make the choice" (1957, p. 5).

Modern treatments are more cautious, but the conceptual difficulty remains. What Graaf made part of a definition, Boadway and Bruce explicate more lucidly as an informal, important and questionable postulate, as a "value judgement" (1984, p. 8). They also suggest that the relevant starting point of welfare economics is not so much objective welfare as "the household's view of welfare" (ibid.). This seems to be how it should be stated: welfare economics assumes that the relevant agents' conceptions of well-being are conveyed by a description of their actual preferences. This is a strongly loaded claim, and it is undefended within welfare economics itself.[23]

[22] He states that "welfare economics is concerned with the conditions which determine the total economic welfare of a community". Throughout his paper, Lange identifies "utility" with "welfare". In particular, he reinterprets utilitarianism as being that doctrine which equates "total welfare" with the sum of individual "welfares", and rejects it on the basis of this interpretation.

[23] For an early statement of this criticism, see Broome (1978). To make the household, rather than the individual, the relevant unit of analysis is to add another questionable assumption. In this chapter, we do not discuss the nature of agents but are most of the time concerned with individuals in the ordinary sense.

2.4 Utility and well-being: critical arguments

Is utility theory relevant to social ethics applications? In this subsection, we shall review some recent arguments, in particular by Rawls and Sen, to the effect that it is, at best, very little relevant. Some of these arguments assume a teleological framework in which the good is equated with individual well-being. Others are either teleological in another sense or quite clearly deontological. Most of the arguments discussed here are overtly critical of the "utility-based" approach in general. However, some are really directed towards the *standard interpretations*: their actual target is utility in either of the two received senses, i.e., pleasure minus pain and actual preference satisfaction, and the essential point made is that utility thus understood cannot really represent the individual's well-being. Other arguments are more disturbing because they suggest rejecting the utility concept *even if it could somehow be made to represent the individual's well-being*. This section attempts at recapitulating the negative case for the ethical use of utility theory, whereas the next subsection will attempt to make a positive case.[24]

Rawls and Sen are prominent among those writers on social ethics who have criticized the use of the utility concept in social ethics. Rawls's (1971, 1982) arguments are difficult to appreciate independently of his overall construction of the "well-ordered society". Sen's arguments are easier to detach. One of them is condensed in the following passage:

> "The choice-approach to well-being is really a nonstarter. But the other two – more classical and more reasonably defended – views of utility, viz., happiness and desire-fulfilment, are indeed serious candidates for serving as the basis of a theory of well-being ... [But] a person who is illfed, undernourished, unsheltered and ill can still be high up in the scale of happiness or desire-fulfilment if he or she has learned to have 'realistic' desires and to take pleasure in small mercies. The physical conditions of a person do not enter the view of well-being seen entirely in terms of

[24] If we were also concerned with *individual* ethics, our discussion would follow a different path. For instance, utility in the pleasure-pain interpretation is relevant to the assessment of a famous system of individual ethics, *hedonism*. We could restate in the language of utility the classic formulations and refutations of hedonism to be found in, e.g., Sidgwick (1884), but we shall not undertake that here.

> happiness or desire-fulfilment, except insofar as they are *indirectly* covered by the mental attitudes of happiness or desire. And this neglect is fortified by the lack of interest, of these two perspectives, in the person's own valuation as to what kind of a life would be worthwhile" (1985, pp. 20–21).

In this passage Sen assumes a teleological framework of social ethics, in which the relevant notion of good is the individual's well-being.[25] The "choice-approach to well-being" is what we called earlier revealed preference theory in the broad methodological sense. It seems permissible to identify the "happiness" and "desire-fulfilment" views with those discussed in Sections 2.2 and 2.3, respectively.[26] Hence, Sen's passage is intended to dispose at once of the more popular interpretations of the utility concept. That revealed preference theory has nothing to say about well-being can be argued as follows: the choices made by an individual might or might not be directed towards his good; by themselves, they indicate nothing about the causes and attainment of his well-being. Naturally, choices specifically directed towards his good are relevant; but revealed preference theory is concerned with choices in general, not with that particular class, and cannot explain how to draw the line. The argument just stated would be too sweeping if it were made against the "happiness" (i.e., pleasure and pain) and "desire-fulfilment" (i.e., preference satisfaction) views. At least, these two notions refer to particular goods.

The point against the "more reasonably defended" interpretations of utility is that well-being has crucial objective components, such as enjoying good health, which can be taken into account only *accidentally* by these views. If I find no enjoyment in smoking, if I prefer taking strolls in a clean countryside to lying on sun-burned and overcrowded beaches, so much the better for me, but plainly in this example, the achievement of better health coincides with higher

[25] Elsewhere he has emphasized freedom as one crucial aspect of the individual's good; see below. Sen (1982b) has also discussed "rights and agency" in a way which does not fit so easily with the teleological versus deontological distinction.

[26] The conflation of happiness with pleasure was not part of the Greek and classical philosophers' tradition but has become common since the beginning of utilitarianism. Desire-fulfilment and preference satisfaction are not really identical objectives, but have been identified with each other in 20th century economics. Pigou (1920) was aware of the conceptual distinction but claimed that it could be ignored for the purpose of his welfare analysis.

enjoyment, or higher preference satisfaction, just by chance. I could have altogether different idiosyncrasies. It is important to understand that by making this point, Sen and his followers revert to commonsense considerations that earlier writers in the field – both in the utilitarian and Paretian traditions – had in some way examined, and concluded to be irrelevant. Both in the pleasure and preference satisfaction interpretations, utility was implicitly claimed to capture not particular goods but an overall conception of the individual's well-being. To concentrate on the preference satisfaction version: if I prefer smoking to non-smoking, despite the fact that by smoking I will impoverish my health, then it must be said that smoking is after all better for me than non-smoking. On this view, it is implicit that preference comparisons are made *all things being considered*. Preference satisfaction can take care – *indirectly* but not *accidentally* – of any other factor pertaining to the individual's well-being. This conception is deeply subjectivist. If all things considered, I continue to smoke heavily, those who claim that I thus promote my own good put more weight on the individual's assent to his own fate than on his objective status. It is never said in this conception that there are objective factors, which are somehow taken into account by the subjective good.

We might interpret Sen's critical point as stating that those earlier writers on utility had not properly thought through their case. First, like clauses *ceteris paribus* in typical reasonings of positive economics, the clause *all things being considered* is the name of a problem, not its solution. In the absence of a proper analysis, Sen suggests, it is better to rely on our commonsense intuitions of what is objectively good for the individual. Second, not just any set of preference judgements satisfies the clause, even if it happens to meet standard formal requirements of utility theory, such as transitivity and completeness. This simple observation has important negative consequences. It was just explained that standard welfare economics relies on the notion of utility as representing *actual* preference satisfaction. By pointing out that only considered preferences appropriately represent the individual's good or well-being, one casts doubts on the ethical relevance of the welfare economics exercise. The crucial assumption mentioned at the end of Section 2.3 could prove to be not a "value judgement", but a brute logical *non sequitur*. Not all of those, mostly Paretian, writers who insist that preferences should

be understood as considered preferences have thought through the damaging consequences of this claim for the existing normative theory.

Elsewhere, Sen argues again against utility as "happiness" or as "desire-satisfaction", but on partly different grounds: any utility-based approach involving either of these interpretations

> "is a restrictive approach to taking note of individual advantage in two distinctive ways: it ignores freedom and concentrates only on achievements, and it ignores achievements other than those reflected in one of these mental metrics" (1992, p. 6).

The critical point made here is again that the utility-based approach contains an inadequate account of the individual's good; but it also introduces the consideration of freedom, with a view of connecting it with the earlier criticism. Sen is not so much concerned with free will as such, as he is with the *objectively defined* conditions of its exercise – then, with "real freedom" as against the formal notions provided by metaphysics or legal theory.[27] In his recent positive work he has come to emphasize the twin notions of *functionings*, to wit, "what the person succeeds in doing with the commodities and characteristics of functionings actually available to the person. For instance, "bicycling" is a functioning; it should of course be kept distinct from the commodity "bicycle" to which it relates. The corresponding capability is the set of bicycling possibilities, which evidently varies from one individual to another. Sen argues that the individual's well-being can be measured by a suitable index of functionings (1985, p. 25). He also emphasizes that normative economics cannot be based solely on a concept of *achievement*, be it the functioning concept, but crucially needs a concept of *opportunity*; capability is then said to be the relevant one. This notion constitutes the seed of Sen's current project in social ethics. We see this project as belonging to teleological ethics, albeit in the sense of some enlarged notion of the good: real freedom, as well as (and perhaps more crucially than) well-being, contributes to defining the good.

The various arguments made thus far share the following common component: they point out that the utility concept is *too limited or too narrow* in order to properly capture the notion of

[27] Or Kantianism for that matter. Kant's *Freiheit* excludes any empirical determination: it is the unrestricted exercise of the will in accord with pure reason.

well-being. They can be read either as completely dismissive arguments, or much less strongly, as critical arguments. For instance, if pleasure or happiness is constructed as a functioning, there might remain a relevant technical role for utility theory; it would provide a metric for at least one among the functionings, and perhaps also for the corresponding capability.

The further group of arguments to be reviewed now are critical or dismissive in a different, and actually nearly opposite way: they suggest that the utility-based approach is *too broad and too flexible*. The approach registers the effect on utility values of too many factors, including those which seem to be irrelevant from the ethical point of view. By and large, contrary to the arguments in the previous group, the arguments below do not dispute that utility values represent individual well-being; rather, they aim at showing that while being perhaps an appropriate representation of well-being, utility functions provide either worthless or insufficient information.

Zealous aspirations have long been an embarrassment to utilitarian writers. If the fanatic's pleasure or satisfaction outweighs the pain or dissatisfaction of the victims of his policies, then the ordinary intuition notwithstanding, the fanatic should have this way.[28] This unpleasant conclusion plagues not only utilitarianism but any theory in which utility, in either of the two senses, would be retained as a significant quantity. For the sake of the argument, let us conceive of a theory in which utility values are balanced against some non-utility-based index, say a relevant index of the individual's capabilities. It is unlikely that such a theory will ever recommend the fanatic's policies, but it might have to pay a taxing price in order to escape from this implication. The fanatic would have to be "compensated", and this might influence resource allocation to a disproportionate extent. This argument suggests that there are cases in which, perhaps, utility values should not be included *at all* in the social ethics evaluation. In the same spirit, and somewhat surprisingly, Harsanyi (1977a,b,c) has recommended to "censor" utility functions before computing the utilitarian sum.[29] Neither the argument nor the conclusion needs modifying when the standard interpretations of utility are replaced with well-being.

[28]Hare (1976, 1981) has repeatedly addressed this problem. Notice that *rule* utilitarianism might eschew it, contrary to *act* utilitarianism.

[29]See also Goodin's (1986) review of various cases of "laundering preferences".

Handicaps have often been mentioned to illustrate another alleged failure of the utility-based approach. A handicapped person needs more resources than an able person to reach the same level of utility; or equivalently, he or she reaches lower utility values when given the same amount of resources. Utilitarian rules will automatically imply that the handicapped person will receive fewer resources than the able one. This intuitively unattractive consequence is very likely to follow from utility-based social rules more generally than just utilitarianism. The problem here with utility values is not that they are altogether irrelevant, as in the fanatic's case, but rather that they are not sufficiently informative to guide the social evaluator. What should be made out of X's low utility value for going to museums and art exhibitions? X might be a philistine, but it might also be the case that X has refined artistic tastes and is disabled, while museums are not well-equipped with ramps, so that he is simply not fit to go. To ration X's access to art just on the grounds of his low utility value for art is to deal with these two possibilities as if they were identical from the ethical point of view. Many contemporary writers find this implication shocking. The counterexample is effective against the standard interpretations of utility but appears to work as well, if utility is construed as measuring well-being.

A formally related problem has been discussed in relation to expensive tastes. The general point is aptly summarized by Rawls: "Desires and wants, however intense, are not by themselves reasons in matter of justice" (1982, p. 171). More clearly than the previous two cases, the expensive taste example leads to assessing whether the individual is *responsible* for his particular utility values. A person who has expensive tastes is like a handicapped person, in that an average amount of resources would leave him or her with a below-the-average amount of utility. (In this argument also, utility can be taken as representing well-being). But there is a difference between the two cases. Only if those with expensive tastes are *not* responsible for these tastes can they be treated as they were handicapped, and thus claim a higher than average amount of resources. Notice the deontological assumption underlying this argument: it relies on our intuition of what a *right* distribution of resources is, irrespective of what its implications for the individuals' good may be.

Given this deontological undertone, it is perhaps not surprising that responsibility has come

to play a major role in Rawls's later work. He discusses it as follows:

> "It is not by itself an objection to (a theory of justice) that it does not accommodate those with expensive tastes. One must argue in addition that it is unreasonable, if not unjust, to hold such persons responsible for their preferences and to require them to make out the best they can. But to argue this seems to presuppose that citizens' preferences are beyond their control as propensities or cravings that just happen. Citizens seem to be regarded as passive carriers of desires" (ibid., pp. 168–69).[30]

Whether or not one is prepared to regard individuals as responsible for their expensive tastes depends in part on the purpose and scope of the theory. Under Rawls's "veil of ignorance", "citizens" decide about their plans of life and about the society which would best make them mutually compatible. From this *ex ante* perspective, it might be reasonable to regard individuals as responsible for their preferences. But from an *ex post* point of view, a case by case discussion seems unavoidable. In 1920, the underpayed professional taxi driver in Paris and the Russian immigrant aristocrat who had to drive a cab for his living were in similar objective situations. Presumably, because of his acquired tastes, the latter did much worse than the former in welfare terms. It is not easy to decide to what extent the Russian was responsible for his preferences, and whether this should influence the *ex post* social evaluation.[31]

Recent normative economics has been much concerned with the issues of expensive tastes, handicaps, and the counterpart of handicaps, talents. These issues have been on the agenda since *A Theory of Justice* (1971). From Rawls's own admission, his book abstracted almost entirely from the problems of handicaps and talents, and accordingly could provide only a first approximation of the desired theory of "justice as fairness". Besides the Rawlsian connection, all these issues have been discussed in relation to a time-honoured question recently revived by Sen (1980), "Equality of what?". We shall give here only a coarse sketch of some of the literature. Crucial to the work of Sen (1985, 1987, 1992), Arneson (1989), Cohen (1989), Roemer (1994,

[30] Responsibility for one's preferences was previously discussed in Scanlon (1975).

[31] Since responsibility is usually not a matter of all-or-nothing, a more appropriate wording is perhaps *to what extent* it should influence the *ex post* social evaluation.

1996) and others, is the notion of the individual's relevant *achievement*, which these authors envisage variously, and normally do *not* describe in terms of a utility function. Arneson (1989) might be the only writer who retains a utility-based notion of achievement while distancing himself from the traditional approach; his is a borderline case. Roughly speaking, these writers analyze individual achievement in terms of three explanatory variables, i.e., physical or economic resources r, talents or handicaps t, and "will" w – a catchword to refer to those factors which are under the individual's own control. While the chosen notion of resource is borrowed from standard microeconomics and regarded as unproblematic, the distinction between talents and handicaps, on the one hand, and "will" on the other, is variously construed and has led to active controversies. The point of the whole construction is to define an appropriate redistributive process. The above-mentioned writers agree with each other that the t factor should be compensated by transfers in r, on the grounds that individuals are not responsible for t, while they are for w, and that differences in achievements can be tolerated only if they can be traced to differences in factors for which individuals are responsible. How the redistributive process takes place is again a matter for discussion, but there are *prima facie* two broad schools of thought, one of which aims at equalizing the individuals' sets of possible achievements, the other is primarily concerned with factors and somehow tries to equalize the individuals' extended resource vectors (r, t).[32] These divergences can be accounted for in terms of different philosophical answers to Sen's question. Like responsibility, equality is of the keywords to the recent work in normative economics.

2.5 Utility, well-being, and social ethics: some positive arguments

As usually defined in recent normative economics and moral philosophy, *welfarism* is the thesis that the utility concept provides all the information required to construct a social evaluation rule. To the best of our knowledge, welfarism was never defended explicitly by any writer. But

[32]Typical representatives of each school are Arneson and Dworkin, respectively. For further elucidation the reader is referred to Fleurbaey's (1994) survey (on which the present paragraph is based). See also the accounts by Arneson (1990a), Hausman and McPherson (1996), and Roemer (1996).

as an underlying assumption, it is clearly shared by utilitarianism (both modern and classical), Paretian welfare economics, as well as a significant part of social choice theory. To reject welfarism is to reject all those approaches at once. Sen (1979) introduced this new concept to capture what he thought was a severe common limitation to all of them.[33] Welfarism is an "informational constraint" imposed on ethical judgements: if all the utility-relative information about two social states is known, one can judge them without knowing anything more about those states. Why this restriction might be exacting has been explained in the last subsection. Some of the arguments reviewed above point towards the conclusion that utility information is relevant, though insufficient. This seems to have been the essence of Sen's position at the early stage of his critique. Rather than being excluded from the analysis, utility values would have to be supplemented by direct descriptions of the states of affairs under examination. Other arguments point towards the stronger conclusion that utility information might be altogether irrelevant. Accordingly, we should distinguish between several positions among the "antiwelfarists". It is convenient (though no doubt oversimplifying) to classify them in terms of the achievement-factor model of Section 2.4:

- One position retains utility *as the* relevant index of achievement but relies on a factor analysis of utility values, in particular when it comes to assessing the role of handicaps, talents and expensive tastes; it is minimally antiwelfarist.

- Another position makes utility one among several indexes of achievements, for instance along with various "capability" indexes, and then proceeds to the factor analysis of each of these achievement concepts.

- Still another position is to exclude utility from the list of relevant achievement indexes; hence is maximally antiwelfarist.[34]

[33] See also Sen (1977, reprinted in 1982a, Chapter 11). Sen's "liberal paradox", as first stated in *Collected Choice and Social Welfare* (1970), is an early occurrence of his critique of welfarism. On this paradox, see Sen's (1976) and (1987) further elaboration and discussion of the extensive secondary literature. As to the word "welfarism", it was first used by Hicks, but with an altogether different meaning.

[34] Some writers, like Kagan (1992), have addressed "the limits of well-being" without really discussing the utility

Whatever its scope, the antiwelfarist critique crucially depends on the chosen interpretation of the utility concept. Enough has been said in Section 2.4 to suggest that the *two receives senses*, i.e., the pleasure and the actual preference-satisfaction interpretations, have little ethical import. They are relevant only as a roundabout way of referring to individual well-being, so it seems appropriate to consider the latter interpretation only. However, some important objections were raised in connection with this interpretation, too. It is by no means clear that today's utility theorists can *fully* meet the antiwelfarists' challenge. But something can be said in favour of a continuing use of choice-theoretic methods in social ethics. We shall sketch three lines of defence.

First, one could further elaborate on the contrast between actual and considered preferences, with a view to showing that the latter relate to individual well-being in an ethically relevant way. One key notion in Griffin's book *Well-Being* is that of "informed desire", to wit, a desire put though criticism and reflection, a desire "formed by appreciation of the nature of its object" (1986, p. 14). This deliberative conception clashes with that of ordinary economics, which views preference as a fixed attribute of the individual, and thus must regard desire, to the extent that it determines preference, as also being fixed. The "informed desire" view goes against a broader tradition, of which economics is the late offspring, that claims that deliberation is only of the means, not of the ends (*Nichomachean Ethics*, 1112b 12). Griffin follows an alternative tradition. The Kantians have been arguing that ends fall within the jurisdiction of reason. The ordinary meta-ethical intuition probably endorses the point that deliberation is of ends (if ethics excludes deliberations of that sort, what then is it about?). One strength of the "informed desire" view is that, in principle it can take into account those objective components of well-being which Sen wanted to emphasize against standard welfare economics. One apparent weakness of this conception, however, is that it is roundabout, and perhaps even redundant: can we not directly consider the objects of our reasoned desires, such as enjoying good health, improving our knowledge, and the like? Griffin's answer is summarized here:

> "The advantages of the informed desire account, therefore, seem to be these. It provides the material needed to encompass the complexity of prudential value. It has

concept. It would not be so easy to fit them in the present classification of antiwelfarist positions.

> the advantages of scope and flexibility over explanations of 'well-being' in terms of desirability features. It has scope because all prudential values, from objects of simply varying tastes to objects of universal informed agreement, register somewhere in informed preferences. It has flexibility, because not everyone's well-being is affected in the same way by a certain desirability feature, and we want a notion sensitive to these differences. We want to know not only that something is valuable, but how valuable it is, and how valuable to different persons" (1986, pp. 30–31).

One implication of this passage is that well-being has an irreducibly subjective aspect. This could well be the grain of truth contained in the utilitarian and Paretian writers' work – even if they overemphasized it to the point of absurdity. Another suggestion is methodological in character: a desire account of well-being avoids the charge of being unduly specific, a charge to which purely objective accounts are clearly open. The objection of undue specificity, if not of arbitrariness, has been raised against Rawls's list of "primary goods", as well as against Sen's examples of "functionings". We cannot here bring this methodological debate to its close, but are implying that the preference-based conception of well-being is probably in no worse predicament than the conception of well-being in terms of objective achievement lists.

To make progress with the previous account of well-being, the connection between the latter and preference satisfaction should be worked out more precisely. We suggest trying the following:

(*) x is better than y for individual i if and only if were i rational and well-informed, i would prefer x to y.

Notice that the conditional occurring after the "if and only if" is a counterfactual. It would be inappropriate to resort to *material implication*, as in the following statement:

x is better than y for individual i if and only if whenever i is rational and well-informed, i prefers x to y.

Here are the reasons why the material implication variant does not do the required job. If one takes the view that no agent is ever rational and well-informed, the variant is vacuously

true. If one takes the less extreme view that agents are not always rational and well-informed, one is led to the paradoxical conclusion that sometimes x is better than y, sometimes x is not, for reasons that we cannot intuitively connect to a change in relative "betterness". Importantly, (*) is meant to convey not a definition of what the individual's good consists of, but a criterion to recognize it. To distinguish here between an *essential definition* and *a criterion* is one way of trying to set welfare economics on its feet. It is clearly wrong to claim, as de Graaff does, that well-being *is* preference satisfaction. The link between the two concepts is at best an external one. (In the same way, very roughly, that choices can be linked to preferences: under certain conditions, choices make it possible to identify the individual's preferences, but choices never define what preferences are.) The welfare economists' further, and no less serious, mistake was to use actual, rather than rational and well-informed, preferences as their implicit criterion.

Some writers constrain the preference notion even further than we have just done in order to tie it to the notion of the individual's good. They insist that other-regarding desires should be kept aside. Accordingly, "self-interested" should be added to "rational" and "well-informed" in statement (*).[35] To add this proviso has much plausibility in the present context of discussion, where individual well-being is taken to be the relevant notion of good. (There are other teleological frameworks, such as perhaps Aristotle's ethics, in which the proviso would not make much sense.) However, one might argue that if self-interest is construed narrowly, for instance so as to exclude any impersonal motives,[36] the proviso will be unduly restrictive. A sufficiently encompassing definition of what it means for the individual to be rational and well-informed might very well account for the kind of long-term interests that one would intuitively like to connect with individual well-being. This leads to asking whether (*) should be construed as involving only *formal*, or also substantial rationality constraints. If "self-interested" is not added to the statement, rationality constraints should clearly be of the latter type. One should also inquire whether the agent's information should be complete and correct, or just partial and correct,

[35] See Broome (1991, p. 133), who however distances himself from the rational-preference-satisfaction theory of the good in whatever version. Griffin's own account of "informed desire" does not emphasize self-interest.

[36] That impersonal motives should be excluded has been defended by some utilitarians, but on different grounds: the proviso would be necessary to avoid *circularity* in defining moral rules; on this argument see Brandt (1979).

and in particular how far it should be extended beyond general knowledge of a law-like sort. In brief, there are a number of ways in which formula (*) can be made precise. To state and compare them would lead us beyond the purview of this chapter, while we just want to argue for the general plausibility of this formula.[37]

There is a second line of argument available. Most of the discussion thus far has assumed a teleological framework of ethics, in which the relevant notion of good is the individual's well-being. But there are alternative notions of good that are relevant to social ethics. We already mentioned that Sen has enlarged his framework of teleological analysis to make room for real freedom. Independently of this, he has sometimes also distinguished between two notions of good, i.e., well-being and "advantage" [e.g., Sen (1992)]. Actually, the whole list of well-received definitions of the good in moral philosophy (pleasure, happiness, human perfection, harmonious civic life, etc.) is *prima facie* relevant to social ethics. Rather than examining each in turn, one might be willing to abstract from particular definitions and attempt directly to comprehend the features *all* notions of good have in common. One might argue that educated people recognize at least broadly what the sensible use of the word "good" is, although they employ it in various senses and lack a theory to account for underlying regularities in their own use. Philosophers might then rely on the rich information provided by ordinary language and sketch the missing theory.[38] If this method works, it is unlikely to lead to a theory like, say, hedonism: it will be a *formal* theory, not a substantial one. To illustrate, when we say that x is better than y, and y better than z, we presumably take for granted that x is better than z. Hence, transitivity appears to be implied by our understanding of the word "good".[39] This seems like a meagre result, but it suggests that the same method may be applied in less obvious directions. What is relevant for our purposes is that transitivity is an axiom from utility theory: the concept of good might be similarly scrutinized from the vantage point of *any* axiom of that theory. In brief,

[37]For a discussion of self-regarding versus other-regarding preferences, see Collard (1978) and Kolm (1984). Indicentally, the restriction to self-regarding preferences is a classic way of escape from Sen's "Paretian liberal" paradox; see Gibbard (1974) and Hammond (1995).

[38]This type of analysis can probably be traced to Moore's *Principia Ethica* (1903).

[39]However, even this has been called into question; see Temkin (1987).

by adopting a formal stand in teleological ethics, one creates room for an ethical application of utility theory, at least as a mathematical framework of analysis.

The method sketched in the last paragraph has recently been illustrated by Broome. His *Weighing Goods* (1991a) makes two major claims: one can investigate the essential properties of good (or rather, the betterness relation) without adhering to any specific substantial theory, and the technical apparatus of preference and utility is relevant to that investigation. The two claims are mutually supporting: if he had not arrived at choice-theoretic conclusions about the structure of good, Broome would have had to face the charge that his ethical formalism is empty; in order to remain within teleology, he would have had to fall back on one of the substantial notions of good. His most significant conclusion is that the good satisfies the technical property of *separability*. This property will be defined below in Section 3.1. Among other things, it implies a warrant to utilitarianism, in the special interpretation of utility as referring to the individual's good. A remarkable feature of his analysis, Broome does not want initially to limit the scope of the good to the individual's good: persons are said to be "locations" of good in a way no different from times and states of nature. In order to argue that good is separable in its three locations, he investigates these properties of the betterness relation as if it were a preference relation. It is found to satisfy the "sure-thing principle", which is the technical expression for the separability property when that property concerns states of nature. Related arguments are intended to take care of separability relative to times and to persons. We do not mean to endorse these conclusions, but only to illustrate how the general method works.

To elaborate on the methodological point from a slightly different angle, remember that the basic structure of a teleological theory results from answering the following questions in succession: (i) How does the particular theory define the good? (ii) How does it relate goodness to rightness? Leaving aside question (i) and assuming a pretheoretic understanding of the good, there remains question (ii), which we argue utility-theoretic methods can help illuminate. The standard way of thinking of (ii) goes as follows: first, the chosen notion of good is attached to states of affairs; second, the rightness of actions is determined by the goodness of their consequences, which are particular states of affairs. Ethical theories based on this two-stage

process are said to be *consequentialist*. Teleological systems of ethics are consequentialist.[40] Now, part of utility theory is precisely concerned with the problem of how to evaluate a prospect in terms of the evaluation of its outcomes. *Prima facie*, there are many ways of constructing an overall evaluation for prospects. Rawls appears to skip over this problem when he simply states that in teleological systems, "the right is defined as that which maximizes the good" (1971, p. 24). It is a matter for technical investigation to formulate a maximization principle which takes into account the two-stage structure of consequentialist reasoning. Among other things, the present chapter is concerned with this issue. It will surface when we discuss Harsanyi's assumption that the social observer should obey the axioms of expected utility theory.

Finally, there is a third line of argument, which goes beyond the confines of teleological ethics. Typically, a non-teleological or deontological theory is one which does not determine the right action from the consideration of the good alone. In Kantian ethics, reason alone – without any role played by the good – determines the right action. In most other deontological systems, especially in the contemporary ones, such as Rawls's, the right action is derived by reason in the light of some consideration of the good. Understood one way or another, rationality plays a crucial role in deontological theories – a role in some sense more explicit than in teleological theories. So one wonders whether utility theory could not become relevant to the deontologist too, as a tool for clarifying his rationality concept. To make this claim precise would involve one in distinguishing between several notions of rationality. Choice theory – a better expression than "utility theory" in this context – deals only with *prudential* rationality, which is known to be of no major concern to Kantian ethics but is relevant to most other deontological systems. Choice theory is restrictive in another, more exacting sense: it is usually said to be "formal", in that it does not take any position on the individuals' aims, but just on the proper way of attaining them.[41] That choice theory leaves aside rationality considerations of interest to the deontologist is beyond doubt. But there is room for the use in deontological ethics of choice

[40] Much of the discussion of teleological versus nonteleological ethics has indeed been concerned with the merits and demerits of consequentialism. See in particular Williams (1973) and Scheffer (1982, 1988).

[41] This claim cannot be accepted without qualification; see Mongin (1984).

theory at least as a subordinate device – i.e., *qua* formal theory of prudential rationality. The main example in this chapter consists in Rawls's (cautious) reliance on choice theory when he derives the principles of justice from the "original position".

The first of the three lines of reasoning sketched above provides us with an interpretation on the *individuals*' utility functions for the welfarist constructions review in this chapter. Note emphatically that collective choice theories do not necessarily have to make the restrictions of a "rational", "well-informed", and (if needed at all) "self-interested" preference explicit in the formalism: some do, but others do not, while becoming nonetheless relevant once the technical notions of utility and preference receive the suggested interpretation. Utility-theoretic restrictions, such as the ordering axiom (always) or the expected utility axioms (sometimes) will also be imposed on *the evaluating observer himself*. To account for these technical restrictions we think that one of the last two interpretations should come into play. The axioms will either serve as tools to investigate the structure of consequentialism, or else refer to primitive rationality constraints that the observer should satisfy. By making these various interpretative points in favour of welfarism, we do not mean, of course, to imply that all the welfarist constructions reviewed below deliver good ethical theories. Each will have to be assessed on its own merits. What we claim at this stage is simply that they belong to the province of ethics, and should be discussed as such.

3 Some definitions and concepts from utility theory

Throughout, we shall denote the set of alternatives by X and the relevant domain of utility functions by $\mathcal{U} \subset \mathbb{R}^X$ (where $\mathbb{R}^X$ is the set of real functions on X). Most of the formal results of this chapter depend on assuming that there is a fixed population of n numbered individuals; denote $\{1, \cdots, n\}$ by N. We shall normally consider the elements of $\mathcal{U}$ as our primitives. We want the technical developments of this and the following sections to be compatible with both the now prevailing notion that utility functions represent preferences of some kind and those earlier theories which exclusively relied on the utility concept. However, a brief reminder of axiomatic

justifications of utility in terms of preferences is to the point in this preliminary section. There will be three major cases.

3.1 Utility functions in the case of certainty, and "economic" domains

As explained in microeconomics textbooks, consumer theory takes the set of alternatives X to be $\mathbb{R}^p$ or some suitably restricted subset of $\mathbb{R}^p$, where p is the number of commodities, and it takes $\mathcal{U}$ to be the set of continuous functions on X. An element in $\mathcal{U}$ is then seen as a representation of the consumer's preference over commodity vectors. Given an *ordering* (i.e., transitivity and completeness) and a *continuity* axiom imposed on the binary preference relation, the existence (and uniqueness up to strictly increasing transformations) of $u \in \mathcal{U}$ follows from a classic representation theorem by Debreu (1960). It is also known that further axiomatic restrictions on the preference relation can be translated into corresponding properties of u. *Monotonicity* properties of preferences directly translate into monotonicity properties of utility representations. *Convex* and *satiable* preferences lead to representations which are quasi-concave and bounded from above, respectively. Another particular case of interest is *separability*. When the preference relation over $X = \mathbb{R}^p$ induces well-behaved preference relations over each component or each subset of components, it is said to be "weakly" or "strongly" separable, respectively. These properties lead to special utility representations; typically, strong separability leads to additive representations with respect to the components.[42]

The more basic axioms of choice under certainty – i.e., the two ordering axioms, transitivity and completeness – have been accepted by most economists and a large number of philosophers and social scientists as being compelling for any rational agent. As the formal theory of choice functions demonstrates,[43] these two axioms essentially exhaust the meaning of "optimization" – i.e., of the notion that the agent's choice maximizes some preference relation – and "optimization" is widely seen as unproblematic.[44] Most writers regard continuity as a purely technical requirement; it serves to bridge the mathematical gap between the preference binary relation

[42] See Gorman's (1968) classic paper and the chapter by Blackorby, Primont and Russell in this volume.

[43] See for instance Richter's (1971) survey.

[44] For a dissenting view, see Mongin (1984, 1994b).

and its numerical representation. As to the other conditions, such as monotonicity and separability, they play only an occasional role in the theory of choice under certainty. They enter the definition of specific "economic" domains of utility functions that have been investigated in both social choice theory and normative economics (see 6.5 for an illustration).

3.2 Von Neumann-Morgenstern utility functions

In von Neumann-Morgenstern (VNM) theory the typical objects of choice are lottery tickets, as in games of chance, or rather some idealization of them. A standard formalization assumes that there is a measurable space of final outcomes Γ and that the alternative set is $X = \Delta_s(\Gamma)$, i.e., the set of *simple* probabilities on Γ. (A simple probability has a finite number of values.) Clearly, this variant as well as the more general one in which $X = \Delta(\Gamma)$ – i.e., the set of *all* probabilities on Γ – takes for granted that the material presentation of lottery tickets, or lotteries for short, does not matter. For instance, compound lotteries are unproblematically identified with simple lotteries.

The basic utility notion in VNM theory is expected utility, which can be formalized in either of the following ways. Granting the standard identification of outcomes γ with sure lotteries, any x on X gives rise to a function v on Γ; the latter is defined to be the restriction of u to Γ. Then, the VNM property is the familiar one that u is *expectational*: i.e., for all $x \in X$,

$$u(x) = \sum_{\gamma \in \Gamma} u(\gamma)x(\gamma).$$

Following an alternative definition, the VNM property states that u is mixture-preserving (MP): i.e., for all $x, y \in X$, all $\lambda \in [0, 1]$,

$$u(x\lambda y) = \lambda u(x) + (1 - \lambda)y(y),$$

where $x\lambda y$ stands for the convex combination $\lambda x + (1 - \lambda)y$. This shorthand will be used throughout. (Of course, $x\lambda y$ is itself a probability).

In either formalization the algebraic restriction on u translates the well-known VNM independence axiom, to the effect that the preference between x and y is equivalent to the preference

between mixtures $x\lambda z$ and $y\lambda z$ (for $\lambda \neq 0$ and any z). When added to the usual ordering and continuity requirements, this axiom implies that there is a utility representation satisfying the VNM property in either the expectional or the mixture-preserving sense, depending on the properties of the set X. It is also the case that the VNM representation, in either formalization, is unique up to a positive affine transformation. For this classic result the reader is referred in particular to Luce and Raiffa (1957, chapter 2) and Herstein and Milnor (1953). Fishburn (1970, 1982, 1988) provides up-to-date presentations.[45] The expectational definition of VNM utility functions always implies the MP one but the converse need to hold in the more general versions of the theory. At least, whenever $X = \Delta_s(\Gamma)$, it is readily seen (by finite induction) that the two definitions are equivalent.[46] We shall denote the set of VNM utility functions on X by $\mathcal{V}(X)$.

That the VNM axioms are compelling for rational agents has been argued in various ways, and similarly disputed. The average opinion leans towards acceptance of this claim. Hammond's (1988a) "consequentialist" reconstruction provides a normative argument for the ordering and independence axioms; see his chapter on "objective" expected utility, and see also McClennen's (1990) thorough critique. Allais (1953) is famous among those who deny that VNM independence is normatively compelling.

3.3 Anscombe-Aumann utility functions

In terms of a time-honoured distinction the VNM theory describes the individual's choice under *risk* rather than under *uncertainty*. That is to say, it takes probabilities in $\Gamma(\Delta)$, or $\Delta_s(\Gamma)$, as *given*, be they objective in the technical sense of expressing relative frequencies, or – more broadly – derived in some antecedent and unspecified way. The former sense is a particular case of the latter, which is the relevant one to consider here: VNP theory *per se* does not take any position on the origin of probabilities. In a multi-individual context, the use of VNM utility

[45] A number of technical variants have evolved from von Neumann and Morgenstern's (1944) initial theorem; see Fishburn (1989) for a historically oriented survey.

[46] See Hammond's chapter on "objective" expected utility for this and other results in VNM utility theory.

theory is tantamount to assuming that probabilities have somehow been agreed upon. Again, the theory does not explain why agreement should prevail among individuals. This puts a severe limitation on the use of VNM theory in social ethics. Roughly speaking, it is inappropriate in any context where the agents have reasons to differ in their assessment of factual matters as well as ultimate objectives.

In contrast, the more thorough Subjective Expected Utility (SEU) theory both provides an explicit derivation of probabilities and allows for disagreements between individuals. Essentially, it views the existence of probability assignments as reflecting coherent preferences over betting schemes. According to the celebrated *Dutch Book argument*, a probability assignment exists if and only if the individual cannot be involved in a system of bets which would leave him with a net loss, whatever the actual state of the world turns out to be. The argument can be phrased in such a way that it ensures not only existence, but also uniqueness of the subjective probability assignment. Crucially, it does not prescribe specific probability values. Starting with Ramsey (1931) and de Finetti (1937), writers in the SEU framework have always insisted that the individual's probability values depend on his particular information. For instance, he might or might not take relative frequencies into account. SEU theorists normally take for granted that whatever the individual's incoming information, he should process it according to Bayes' s rule any time that the latter applies meaningfully (i.e., whenever the prior probability of the conditioning event is nonzero). This further piece of doctrine is only loosely connected with the Dutch book argument. Despite this fact, we shall follow a widespread practice and also refer to SEU theory as "Bayesianism", especially in contexts where the existence of subjective probabilities is the important feature to emphasize (as in 5.4.).[47]

In the main, there are three axiomatizations of SEU in current use – Savage's (1954 and 1972), Anscombe and Aumann's (1963), and Jeffrey's (1965 and 1983). Each of the three starts from the preference concept as a primitive, and derives both a unique subjective probability and

[47]SEU theory should also be seen as part of the philosophical and statistical tradition of subjective probability theory which developed independently of choice theory. For this important connection see, e.g., Fine (1973), or the more accessible survey by Fishburn (1986).

an expected utility representation of preferences, where the expectation is taken with respect to the (endogenous) subjective probability. As in VNM theory, the SEU representation is unique up to positive affine transformations. In Jeffrey's system the objects of preference are *propositions* in the logician's sense. Accordingly, the set X of all propositions is endowed with a Boolean algebra structure. Jeffrey's axiomatization has become the classic version of SEU theory among English-speaking philosophers. Economists usually rely on Savage's or Anscombe and Aumann's axiomatizations, in which the assumed objects of choice are altogether different from Jeffrey's.[48] Given a set Ω of states of the world and a set $\mathcal{C}$ of consequences, the individual's objects of choice are functions $a : \Omega \to \mathcal{C}$. Clearly, these state-dependent functions are meant to represent subjective lotteries – i.e., lotteries which assign consequences (prizes) to each state *without* specifying the state probabilities. It is also appropriate to view state-dependent functions s those acts which are available to the agent. The intuition underlying the latter interpretation is roughly this: a consequence is entirely determined by the joint data of the individual's action and the state of nature; hence, to choose an action is tantamount to selecting a particular way in which states influence consequences. Savage insists on the "act" interpretation of state-dependent functions, whereas Anscombe and Aumann just regard them as subjective lottery tickets. (They call subjective lottery tickets "horse lotteries" and contrast them with the "roulette lotteries" of VNM theory, which carry given probabilities.) We shall always refer to the a functions as *acts*, even in the context of Anscombe-Aumann applications, thus avoiding any possible confusion between two kinds of lotteries. Notice that any consequence c can be identified with a relevant *constant act*, i.e., that act a_c which has constant value c across states.

A remarkable common feature of these three systems is that each involves a seemingly irrelevant structural assumption relative to the alternative set X. Jeffrey requires the Boolean algebra of propositions to be "atomless". A related condition appears as (P6) in Savage's axioms: it says in effect that the event space $(\Omega, \mathcal{A})$ is infinitely divisible; in particular, there must be infinitely many states in Ω. The corresponding mathematical restriction in Anscombe and Aumann is

[48] To the best of our knowledge, Broome's (1990) axiomatization of utilitarianism is the only application of Jeffrey's theory in the field of economics at large.

that the consequence set $\mathcal{C}$ is convex; typically (though not necessarily), it is a probability set. Contrary to Savage, Anscombe and Aumann can deal with acts having finite domains; but they have to constrain the range of acts somewhat artificially, by requiring it to be convex. There are deep – and by now, well-recognized – logical difficulties surrounding the axiomatization of SEU theory in a completely general context.[49] By and large, there is a gap between the preference axioms and the numerical data (i.e., the probability and utility functions) representing them. To bridge the gap, a restriction on the structure of the alternative set turns out to be unavoidable. This fact has serious implications for the philosophical assessment of SEU theory. For the axioms are meant to express the rationality justifications of the expected utility formula. Ideally, they should not exceed the logical content of the latter. No such difficulty occurred in the simpler case of VNM theory.

Conceptually, Savage's axiomatization is the most interesting among the three. It involves the widely discussed *sure-thing principle* which can be seen – very roughly speaking – as the subjective lottery variant of VNM's independence axiom. The sure-thing principle – Savage's (P2) – plays a crucial role in the derivation of both the additive property of subjective probabilities and the expectational form of the utility representation. As was mentioned earlier, most normative discussions of SEU theory have focused on the sure-thing principle. However, an important component in Savage's axiomatization – i.e., postulates (P3), (P4) and (P5) – has to do in effect with the Dutch book argument and the possibility of deriving the individual's subjective probability uniquely. Despite – or perhaps because of – its explicitness, Savage's approach is less elegant than Jeffrey's and less handy than Anscombe and Aumann's. In this chapter we shall base our technical developments on the latter, and only allude to the other two.

As in Anscombe and Aumann (AA), we introduce a finite state set Ω and a consequence set $\mathcal{C}$ endowed with the following special structure: $\mathcal{C} = \Delta_s(\Gamma)$ for some measurable set Γ, i.e., $\mathcal{C}$ is the set of lottery tickets – in the VNM sense – over some given set of final outcomes. This heavy but convenient restriction made it possible for AA to use antecedent results from VNM theory

[49] See Krantz et al. (1971) or the more accessible discussion in Suppes (1981). Wakker (1989, IV.6) also provides comments, extensions and references.

in order to derive their representation theorem. Notice that the chosen definition implies that $\mathcal{C}$ is convex. The alternative set is $X = \mathcal{C}^{\Omega}$, i.e., the set of all acts. Given the above-mentioned identification of c with a_c, any function u on X gives rise to a function v over $\mathcal{C}$. The AA axioms imposed on the preference relation imply that there is a unique probability p on Ω , and a utility u on X having the expectational form: for any $a \in X$,

$$u(a) = \sum_{\omega \in \Omega} p(\omega) v(a(\omega)).$$

As in VNM theory, the set of positive affine transformations of u is exactly the set of all utility representations having the expectational form. We shall denote the set of Anscombe-Aumann utility representations on $\mathcal{C}^{\Omega}$ by $\mathcal{A}(X)$.

Importantly, both Savage's and Anscombe and Aumann's axiom systems deliver *state-independent* SEU representations, i.e., expected utility representations in which the v function depends on states ω only through the chosen act $a(\cdot)$. The above AA representation should be contrasted with the following, more general variant:

$$u(a) = \sum_{\omega \in \Omega} p(\omega) v(\omega, a(\omega)).$$

It is not difficult to axiomatize this *state-dependent* SEU representation, but the resulting preference axioms will imply that the subjective probability p is essentially indeterminate. In other words, the Dutch book procedure for identifying subjective probabilities is no longer operative in the state-dependent case. This well-known difficulty explains why most writers in choice theory have preferred to avoid it altogether in spite of the serious loss of generality implied by state-independence.[50]

That the axioms of SEU theory are compelling for rational agents has been both strongly argued for and strongly disputed, with the dissenters typically going in one of these two directions: either they retain the expected utility property but take the expectation with respect to a non-standard ("non-additive") concept of probability distribution; or they retain probability but apply a non-expected utility formula to it. As for VNM theory, the average opinion

[50] On the topic of this paragraph, Fishburn (1970) and the survey of Schervish, Seidenfeld and Kadane (1990) are good sources.

is biased towards acceptance of the normativity claim. Hammond's (1988a) consequentialism encompasses SEU theory and therefore provides an argument for it; see Chapter 6 on subjective expected utility in this volume. The dissenters often base their case on a famous counterexample by Ellsberg (1954). The normative appraisal of the Dutch Book argument has led to lively discussions among philosophers.[51]

3.4 Interpersonal utility differences

At some point we shall have to formally express the assumption that the individuals' utilities strongly differ from each other. A attractive rendering is to say that for each individual, there is a pair of alternatives such that he is not indifferent over this pair while the other individuals are. Formally, if $f_1, \cdots, f_n$ denote the individuals' utility functions over X, we shall require that:

(*) $\quad (\forall\, j \in N)(\exists\, x_j, y_j \in X) f_j(x_j) > f_j(y_j)$ and $f_i(x_j) = f_i(y_j), i \neq j$.

This requirement should be compared with the more abstract property that:

(**) $\quad f_1, \cdots, f_n$ are affinely independent.

Recall that a family of real-valued functions $f_1, \cdots, f_n$ is said to be *affinely independent* if whenever $\sum \alpha_i f_i + \beta = 0$, then $\alpha_1 = \cdots = \alpha_n = \beta = 0$; otherwise, it is said to be affinely dependent. Equivalently, define the f_i to be affinely dependent if there is $j \in N$ such that f_j can be written as an affine combination of the remaining f_i; and define the f_i to be affinely independent if otherwise. Clearly, affine independence strengthens the more familiar concept of linear independence, and the two concepts collapse into each other in the particular case where the f_i ae probabilities. Now, it follows from the definitions that (*) $\Rightarrow$ (**). It turns out that whenever the f_i are expected utility representations, the implication (**) $\Rightarrow$ (*) also holds.[52] Hence, for most of the technical developments of the present chapter, the interpretable property (*) and the less transparent (but mathematically handy) property (**) turn out to be identical. This fact will be put to use in Section 5.[53]

[51] See in particular Howson and Urbach (1989).

[52] This result was proved by Fishburn (1984) in the VNM case and by Mongin (1995a) for SEU representations.

[53] Notice this perhaps not very intuitive implication of defining different utility functions in terms of either (*) or (**): opposite utility representations (i.e., such that $u_j = -u_i$) will not be treated as being "different".

4 The aggregative setting with interpersonal comparisons of utility

This section discusses the connection between social ethics and utility theory within a particular aggregative setting that has been provided by social choice theory. Whereas Arrow's celebrated *Social Choice and Individual Values* (1951) considered only preference relations and embodied the assumption that these preferences are noncomparable, Sen's *Collective Choice and Social Welfare* (1970) opened the way to a more general kind of social choice-theoretic investigation involving utility as well as preference, and allowing for both comparability and noncomparability assumptions.[54] The crucial step was to redefine Arrow's aggregative process appropriately. The Arrovian *social welfare function* (SWF) maps individual preference relations into the binary relation that models the collective preference relation. Sen and his followers' major tool of analysis is the *Social Welfare Functional* (SWFL), which maps individual utility functions to a collective preference. In this framework, Arrow's noncomparability assumption, as well as the various comparability assumptions that suggest themselves, can be captured under the guise of particular *invariance principles*. Arrow's impossibility theorem can then be restated and compared with possibility results which follow from selecting weaker *invariance* principles than noncomparability; prominent among the latter are characterizations of utilitarian rules. Most of the results of SWFL theory can be obtained by a two-step process, one leading to technical welfarism (as defined in Section 4.1), the other to the specific formula (e.g., utilitarian) according to which the individuals' utilities are to be combined (see Sections 4.2 and 4.3). Since this material has been extensively reviewed elsewhere,[55] we will restrict technicalities to essentials and lay the emphasis on interpretations (in particular in Sections 4.4 and 4.6).

[54] Arrow's approach is implicitly related to the welfare economics of the Paretian school, as it took shape in the 1930s. Sen (1970) and d'Aspremont (1995) discuss this connection. Arrow's later work (e.g., 1973a) shows that he became less adamant in his rejection of interpersonal comparisons.

[55] In particular by Sen (1982a, 1986), Blackorby, Donaldson and Weymark (1984), d'Aspremont (1985), Moulin (1988). The reader is referred to Bossert and Weymark's chapter in volume II on utility in social choice for a complete exposition of SWFL theory.

4.1 The SWFL framework

The notation will be in keeping with that of Section 3. It is meant to suggest relevant connections between SWFL theory and individual utility theory. As before, X will refer to the alternative set, which is taken here to be identical for the various individuals and the social observer. This is not an insignificant assumption. When the bearer of collective preference refers to the state, its alternative set is obviously non-coincident with those of the members of society (which might themselves differ from each other). To reduce a situation of initially heterogeneous alternative sets to the present framework is not logically impossible, but might require some care. For any given set $\mathcal{U}$ of utility functions, we define a social welfare functional, or SWFL, to be a function:

$$F : \mathcal{U}^n \to 2^{X \times X},$$

such that for every $U = (u_1, \cdots, u_n) \in \mathcal{U}^n$, $F(U)$ is a weak ordering on X (i.e., it is a transitive, reflexive and complete binary relation). $F(U)$ refers to the collective weak preference. Denote the induced strict preference and indifference relations by $P(U)$ and $I(U)$, respectively.

The following axioms belong to standard SWFL theory.

Axiom I. (Independence of Irrelevant Alternatives)

$$\begin{aligned}
&\forall\, U, U' \in \mathcal{U}^n, \forall\, x, y \in X, \\
&U(x) = U'(x) \\
&U(y) = U'(y) \;\Rightarrow xF(U)y \text{ iff } xF(U')y.
\end{aligned}$$

Axiom PI. (Pareto Independence)

$$\begin{aligned}
&\forall\, U \in \mathcal{U}^n, \forall\, x, y \in X, \\
&U(x) = U(y) \;\Rightarrow xI(U)y.
\end{aligned}$$

Axiom PWP. (Pareto Weak Preference)

$$\begin{aligned}
&\forall\, U = (u_1, \cdots, u_n) \in \mathcal{U}^n, \forall\, x, y \in X, \\
&u_i(x) \geq u_i(y), i = 1, \cdots, n \text{ [denoted as } U(x) \geq U(y)] \\
&\Rightarrow xF(Y)y.
\end{aligned}$$

Axiom WP. (Weak Pareto)

$$\forall U = (u_1, \cdots, u_n) \in \mathcal{U}^n, \forall x, y \in X,$$
$$u_i(x) > u_i(y), i = 1, \cdots, n \text{ [denoted as } U(x) \gg U(y)]$$
$$\Rightarrow xP(U)y.$$

Axiom Strict P. (Strict Pareto)

$$\forall U = (u_1, \cdots, u_n) \in \mathcal{U}^n, \forall x, y \in X,$$
$$u_i(x) > u_i(y), i = 1, \cdots, n \& \exists j : u_j(x) > u_j(y)$$
$$\text{[denoted as } U(x) > U(y)]$$
$$\Rightarrow xP(U)y.$$

Axiom SP. (Strong Pareto)

$$\equiv \text{ (Strict P) \& (PI)}$$

Clearly, **(Strict P)** $\Rightarrow$ **(WP)** and **(PWP)** $\Rightarrow$ **(PI)**. The vector inequality notation introduced above will be used throughout this chapter.

Standard SWFL theory deals with the case in which X is any set of at least three elements and $\mathcal{U} = \mathbb{R}^X$. This domain restriction leads to a classic lemma which says in effect that all the information relevant to the ethical preference is contained in the *set of values* taken by all utility function vectors U. Formally, the problem of ranking alternatives in X is reduced to the problem of ranking elements in the set $\cup\{Range U/U \in \mathcal{U}n\}$, which is readily seen to be equal to $\mathbb{R}^n$. Could we rely on a similar lemma in SWFL theory when VNM restrictions are assumed, i.e., when $X = \Delta_s(\Gamma)$ for some outcome set Γ, and $\mathcal{U} = \mathcal{V}(X)$? The answer is in the affirmative, provided that a (weak) dimensionality condition holds. We state formally the standard result as well as its VNM variant:

Lemma 4.1 (Welfarism Lemma) *Assume that F is a SWFL, and either of the following domain definitions holds:*

(i) Either X is any set of at least three elements, and $\mathcal{U} = \mathbb{R}^X$;

(ii) Or $X = \Delta_s(\Gamma)$ has (vector space) dimension at least 2, and $\mathcal{U} = \mathcal{V}(X)$.

Then F satisfies **(I)** *and* **(PI)** *if and only if the binary relation R^* on $\mathbb{R}^n$ defined by*

$$aR^*b - \exists U \in \mathcal{U}^n, \exists x, y \in X \text{ s.t. } U(x) = a, U(y) = b \text{ and } xF(U)y$$

is an ordering.[56]

The R^* relation defined in this lemma will be called a *social welfare ordering* (SWO), and used extensively below. We denote by P^* and I^* its asymmetric and symmetric parts, respectively.

The present notion of welfarism is a purely technical one. It is important to distinguish it from the *philosophical* notion of welfarism, which has been discussed in Section 2. The latter does not assume any specific formalization such as the SWFL framework used here. Being purely mathematical, the former would be compatible with any interpretation of the u_i symbol referring to other concepts than utility *if for that other interpretation, the underlying axioms* **(I)** *and* **(PI)** *could be defended.* For example, Kelsey (1987) applies the axioms of technical welfarism to "criteria" which might not be utilities in one of the received senses, and Roberts (1995) constructs a modified SWFL framework which allows for comparisons of *ways of comparing utilities as well as* for straightforward utility comparisons. Notice that within the SWFL framework, one can envisage mathematical definitions of welfarism of varying strength.[57] The definition here is the most common and the simplest among those available.

4.2 Invariance axioms and interpersonal comparisons of utility

The classical invariance axioms of SWFL theory must now be introduced:

Axiom CC. (Cardinality and Full Comparability)

$$\forall U \in \mathcal{U}^n, \forall \alpha > 0, \forall \beta \in \mathbb{R} \text{ s.t. } \alpha U + (\beta, \cdots, \beta) \in \mathcal{U}^n,$$
$$F(U) = (\alpha U + (\beta, \cdots, \beta)).$$

[56] For a proof of (i), see d'Aspremont (1985, Theorem 2.1). For (ii), see Mongin (1994a, Lemmas 1 and 2).

[57] For instance, Roberts (1980b, pp. 425ff) states a notion of welfarism for SWFL satisfying **(I)** and **(WP)**. Alternatively, axiom **(I)** rather than the Pareto conditions can be weakened to derive further technical notions of welfarism. See also the variant presented in Section 4.6.

Axiom CU. (Cardinality and Unit Comparability)

$$\forall U \in \mathcal{U}^n, \forall \alpha > 0, \forall \beta \in \mathbb{R}^n \text{ s.t. } \alpha U + \beta \in \mathcal{U}^n,$$
$$F(U) = F(\alpha U + \beta).$$

Axiom CN. (Cardinality and Noncomparability)

$$\forall U \in \mathcal{U}^n, \forall \alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n_{++}, \ \forall \beta = (\beta_1, \cdots, \beta_n) \in \mathbb{R}^n$$
$$\text{s.t. } (\alpha_1 u_1 + \beta_1, \cdots, \alpha_n u_n + \beta_n) \in \mathcal{U}^n,$$
$$F(U) = F(\alpha_1 u_1 + \beta_1, \cdots, \alpha_n u_n + \beta_n).$$

In the statement of the following two axioms $\varphi_1, \cdots, \varphi_n$, φ stand for arbitrary strictly increasing functions from suitably defined subsets of $\mathbb{R}$ to $\mathbb{R}$, and $\varphi \circ u$ stands for the operation of composing functions φ and u.

Axiom OC. (Ordinality and Comparability)

$$\forall U \in \mathcal{U}^n, \forall \varphi \text{ s.t. } (\varphi \circ u_1, \cdots, \varphi \circ u_n) \in \mathcal{U}^n,$$
$$F(U) = F(\varphi \circ u_1, \cdots, \varphi \circ u_n).$$

Axiom ON. (Ordinality and Noncomparability)

$$\forall U \in \mathcal{U}^n, \forall (\varphi_1, \cdots, \varphi_n) \text{ s.t. } (\varphi_1 \circ u_1, \cdots, \varphi_n \circ u_n) \in \mathcal{U}_n,$$
$$F(U) = F(\varphi_1 \circ u_1, \cdots, \varphi_n \circ u_n).$$

Clearly, **(ON)** $\Rightarrow$ **(OC)** $\Rightarrow$ **(CC)** and **(ON)** $\Rightarrow$ **(CN)** $\Rightarrow$ **(CU)** $\Rightarrow$ **(CC)**. The consequences of these axioms in terms of two classes of social choice rules will be spelled out below. This is not an exhaustive list. (For a more detailed analysis, see Bossert and Weymark's chapter. Their labelling of invariance conditions may occasionally differ from ours.)

Invariance axioms are intended to capture the *impossibility* of certain interpersonal comparisons of utility. For instance, **(CU)** implies that the ethical observer cannot compare the *levels* of cardinal utility functions. The comparisons which are possible (though not compulsory, of course) are exactly those which are not excluded by the given invariance axiom. For instance, **(CU)** allows for comparison of *measurement units* of cardinal utility functions. A

well-recognized consequence of the invariance formalism is that the stronger the chosen axiom is, the narrower the basis for interpersonal comparisons on the observer's part. At one end of the spectrum, to assume a pure SWFL framework without any invariance axiom is tantamount to assuming that *any* interpersonal comparison is possible. At the other end, the strong axiom **(ON)** says in effect that no comparison is possible. Many social choice theorists conclude that the invariance based approach to SWFL leads to the following paradox: the *ethical* content of social choice theories is inversely proportional to their *logical* content.[58] It should be emphasized that the paradox depends on the prior philosophical assumption that interpersonal utility comparisons, rather than the lack of them, are a problem. Such an assumption underlies Arrow's (1951) pioneering work, as well as (although to a lesser degree) his followers' contributions, but cannot be accepted uncritically. Our discussion of Bentham and Pareto in Section 2 provides some historical perspective on this problem. To pursue the issue of interpersonal comparisons of utility, the reader is referred to Hammond's chapter as well as his extensive (1991) survey and the many references listed in these two papers.

4.3 Further conditions on SWFL

Once technical welfarism holds, all unanimity axioms stronger than Pareto indifference, as well as all of the invariance axioms above, are automatically translated into corresponding conditions on the set of utility values, i.e., $\mathbb{R}^n$. For instance, using the definition of the R^* relation in the Welfarism Lemma, **(PWP)** becomes equivalent to:

Condition PWP*

$$\forall\, a, b \in \mathbb{R}^n, a \geq b \;\Rightarrow\; aR^*b,$$

and so forth for the remaining Pareto axioms. Similarly, **(CC)** becomes equivalent to:

Condition CC*

$$\forall\, a, b \in \mathbb{R}^n, \forall\, \alpha > 0, \forall\, \beta \in \mathbb{R},$$
$$aI^*b - \alpha a + (\beta, \cdots, \beta) I^* \alpha b + (\beta, \cdots, \beta),$$

[58] See the discussion of this point in Blackorby, Donaldson and Weymark (1984).

and so forth for the remaining invariance axioms. In the sequel the starred name of an axiom should always be understood as referring to the translation of that axiom in terms of the R^* relation. Starred conditions are more tractable than their initial counterparts in terms of social welfare functionals. This well-known technical advantage is put to use in the expositions of SWFL theory by Blackorby, Donaldson and Weymark (1984) and d'Aspremont (1985). If only for expository purposes, these writers assume the conclusion of the Welfarism Lemma to hold throughout.

The following axioms will be used in the sequel. One is the well-known axiom of Anonymity:

Axiom A

For any permutation $\sigma(\cdot)$ of $N = \{1, \cdots, n\}$ and any $U \in \mathcal{U}^n$,
$F(U) = F(u_{\sigma(1)}, \cdots, u_{\sigma(n)})$.

When technical welfarism holds, **(A)** is translated into:

Condition A*

For any permutation $\sigma(\cdot)$ of N and any $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$,
$aI^*(a_{\sigma(1)}, \cdots, a_{\sigma(n)})$.

At the general level, **(A)** stipulates that individuals should receive equal treatment in utility terms. The exact implications of this condition depend on whether and which (if any) interpersonal comparisons of utility are allowed. For instance, in the presence of technical welfarism and the utilitarian invariance principle, it will imply that each individual's utility *differences* are treated equally.

The most extreme case of unequal treatment is perhaps when one individual alone dictates the social preference. Following Arrow's (1951) definition, dictatorship prevails when one individual dictates the social *strict* preference. This leaves the social preference undetermined when the dictator is indifferent between two alternatives.[59] Restating Arrow's notion in the SWFL framework, we define *Nondictatorship* as follows:

[59]One possible strengthening of Arrow's notion is to introduce a *hierarchy* of dictators, where the $(n+1)$-th dictator rules if the n-th one is indifferent [Gevers (1979)].

Axiom ND

There is no $i \in N$ such that for all $x, y \in X$ and $U \in \mathcal{U}^n$,

$u_i(x) > u_i(y) \;\Rightarrow\; xP(U)y.$

Obviously, **(A)** $\Rightarrow$ **(ND)**, and under technical welfarism the latter condition will be translated into

Condition ND*

There is no $i \in N$ such that for all $a, b \in \mathbb{R}^n, a_i > b_i \Rightarrow aP^*b.$

The last condition is Continuity, to be defined here directly in terms of the R^* relation:[60]

Condition C*

For all $a \in \mathbb{R}^n$, the sets $\{a' \mid a' \in \mathbb{R}^n \text{ and } a'R^*a\}$
and $\{a' \mid a' \in \mathbb{R}^n \text{ and } aR^*a'\}$ are closed in $\mathbb{R}^n$.

Lemma 4.2 *For whichever domain restriction stated in Lemma 4.1, the following holds: if F satisfies* **(I)** *and* **(PI)**, *then each starred condition is equivalent to the corresponding axiom on SWFL, and:* **(A)** *&* **(CU)** $\Rightarrow$ *(C*).*

4.4 Utilitarianism versus leximin

We shall now restate and discuss some classic characterizations of social-choice-theoretic rules. The most famous among the utilitarian rules is the Benthamite *sum utilitarianism* rule:

$$\forall\, U = (u_1, \cdots, u_n) \in \mathcal{U}^n, \forall\, x, y \in X,$$
$$xF(U)y - \sum u_i(x) \geq \sum u_i(y).$$

A relevant variant is *mean utilitarianism*, where the equivalence just stated is replaced with:

$$xF(U)y - 1/2 \sum u_i(x) \geq 1/n \sum u_i(y).$$

Obviously, the two rule coincide when the size n of the population is fixed. Henceforth, the expression *standard utilitarianism* will refer to either of these classic rules.

[60] For brevity, we follow the standard exposition [see Maskin (1978)], but it would be more consistent to define continuity in terms of the primitive notion F.

At least for technical reasons, we need to introduce weaker variants than those of standard utilitarianism. After d'Aspremont (1985, p. 46), we define *generalized utilitarianism* as follows: there are nonnegative numbers $\alpha_1, \cdots, \alpha_n$, one of which is strictly positive, such that:

Rule GU

$$\forall U = (u_1, \cdots, u_n) \in \mathcal{U}^n, \forall x, y \in X,$$
$$\sum \alpha_i u_i(x) \geq \sum \alpha_i u_i(y) - xF(U)y.$$

A further variant is *weak utilitarianism*. It is defined by replacing **(GU)** with:

Rule WU

$$\sum \alpha_u u_i(x) > \sum \alpha_i u_i(y) \Rightarrow xP(U)y.$$

When $\alpha_i = 0$ for all except but one i, **(WU)** reduces to dictatorship in Arrow's sense, while **(GU)** delivers a different dictatorship concept.[61] Rule **(WU)** is hardly weaker than **(GU)**. Using completeness of the social preference $F(U)$, it can be checked that only one piece of information should be added to **(WU)** in order to recover **(GU)**:

Another widely explored social-choice-theoretic rule is the *leximin principle*. It says that alternatives x and y will be socially ranked according to the minimum individual utility values in each alternative; if the worst-off individual in x and the worst-off in y happen to have the same amount of utility, the rule prescribes to compare the minimum utility values in the remaining population, and so on. That is to say, it gives lexicographic priority to the worse-off (in utility terms).

Formally, given any utility vector $U(x) = (u_1(x), \cdots, u_n(x))$ define $\hat{U}(x) = (\hat{u}_1(x), \cdots, \hat{u}_n(x))$ to be any permutation of $U(x)$ which ranks the individual utility levels in a weakly increasing order, i.e., $\hat{u}_1(x) \leq u_2(x) \leq \cdots \leq \hat{u}_n(x)$. Now, the *leximin rule* can be defined as follows:

Rule L

$$\forall U \in \mathcal{U}^n, \forall x, y \in X,$$
$$xP(U)y - \exists\, m \in N \text{ s.t. } \forall\, h < m, \hat{u}_h(x) = \hat{u}_h(y), \text{ and } \hat{u}_m(x) > \hat{u}_m(y).$$

[61] If $\alpha_i = 0$ for all except but one i, and **(GU)** rather than **(WU)** holds, i imposes not only his strict preference but also his indifference relation. This is another strengthening of Arrow's definition of a dictator.

Like symmetric utilitarian rules, leximin implies that x and y are socially indifferent whenever $U(x)$ and $U(y)$ are identical up to a permutation.[62] Unlike symmetric utilitarian rules, leximin is not compatible with any further indifference case. It should be clear that **(L)** requires that utility levels be comparable from one individual to the other: unsurprisingly, **(OC)** will emerge as part of the characterization of this rule. Another – intuitively unattractive – concept must be introduced for formal purposes: define the *leximax principle* as that rule which gives lexicographic priority to the better-off.

Sen (1970, Chapter 9, and 1974) introduced the notion of leximin in the course of discussing Rawls's conception of justice.[63] He noted that Rawls's (1958, 1967) work pointed towards the simpler principle of *maximin*, which consists in giving priority to the worst-off (without paying attention to any higher ranks). For instance, if two alternatives x and y are described by utility vectors (1,4, 7) and (5, 1, 8) respectively, maximin would declare x and y to be socially indifferent. The example illustrates a gross violation of the Pareto principle (in the **(Strict P)** version). This is why Sen (1970, p. 138) argued for a refined, iteratively defined concept. Relevant SWFL characterizations of leximin were provided by Hammond (1976), Strasnick (1976), and d'Aspremont and Gevers (1977). The latter writers rely on some of the SWFL axioms explained above, as well as the following added principle of *separability*:

Axiom SE

$$\forall U, U' \in \mathcal{U}^n, F(U) = F(U') \text{ if } \exists\, M \subset N \text{ s.t.}$$
$$(i) \quad \forall\, i \in M, u_i = u'_i$$
$$(ii) \quad \forall\, j \in N \setminus M, u_j \text{ and } u'_j \text{ are constant functions.}$$

In this statement, individuals in $N \setminus M$ can be said to be *unconcerned* by which alternative is chosen. Hence, it says that unconcerned individuals do not influence the social prefer-

[62] All these rules also imply that x is socially strictly preferred to y if any permutation of the $U(x)$ vector Pareto-dominates $U(y)$ (in the **(Strict P)** sense). This property is Suppes's (1966) *grading principle*. It has sometimes been defended as a minimum equity principle for welfaristic contexts [e.g., Suzumura, (1983)].

[63] Kolm (1972, 1974) also contributed to introduce leximin under the label *justice pratique*. However, Kolm applies leximin to "fundamental preferences" which (by construction) are identical from one individual to the other (see Section 6.3). Consistently with his leximin approach, Kolm (1993) strongly argues against utilitarianism.

ence. D'Aspremont and Gevers's result states that leximin and leximax together are characterized by Independence of Irrelevant Alternatives, Strong Pareto, Ordinality and Comparatiblity, Anonymity, and Separability. A complementary result by Deschamps and Gevers (1978) shows that if Ordinality and Comparability **(OC)** is *weakened* into Cardinality and Full Comparability **(CC)**, and if leximax is excluded by assumption, the set of admissible rules is exactly leximin *and* weak utilitarianism. We restate here these two results together:

Proposition 3 *Take any SWFL which satisfies* **(I)**, **(SP)**, **(A)**, **(SE)** *and does not coincide with leximax. Then:*

(i) It coincides with either leximin or weak utilitarianism (with equal coefficients) if and only if it satisfies **(CC)**.

(ii) It coincides with leximin if and only if it satisfies **(OC)**.

(iii) It coincides with standard utilitarianism if and only if it satisfies **(CU)**.[64]

Part (i) comes close to saying that assuming the welfarism framework, an apparently modest requirement of separability and a sufficiently general axiom of interpersonal utility comparisons imply that there are no rules *other than egalitarianism (in the sense of leximin) and utilitarianism*. The difference between this loose wording and part (i) reflects the technical difference between weak utilitarianism and standard utilitarianism. In the case in which $\sum \alpha_i u_i(x) = \sum \alpha_i u_i(y)$, the conclusion that $xI(U)y$ does not necessarily follow; for instance, it would be permissible to break the tie by applying leximin. However, the loose wording captures the essential message of the Deschamps-Gevers theorem – arguably, one of the philosophically most instructive contributions of SWFL theory. This result makes it possible to construe utilitarianism and leximin as being *exhaustive* alternatives. In other words, it gives a formal explanation of why the debate between Rawls and Harsanyi is absolutely central to social ethics.

[64]For (i), see the proof of Theorem 2 in Deschamps and Gevers (1978). For (ii) and (iii), see the proof of Theorem 7 in d'Aspremont and Gevers (1977).

The above discussion of maximin and leximin is typical of the *welfaristic* reconstruction of Rawls's *Theory of Justice* (1971), a reconstruction which was authorized not only by Sen's early work, but also by Arrow's comments on Rawls (1973a, 1973b), and which quickly gained acquiescence among economists.[65] As is well-recognized by now, this interpretation leaves aside several crucial aspects of Rawls's conception. Both Rawls himself (1982) and his commentators have insisted that "justice as fairness" leads to rejecting the utility concept as a way of assessing the members of society's positions, and that accordingly, the welfarist reconstruction is inadequate. An "index of primary goods" should be used instead of utility functions. This distinction will be revisited in 6.4. At least, the SWFL framework used here has the didactic advantage of illuminating one important claim in Rawls's doctrine: he admits of no trade-off when it comes to the individuals' essential interests. His strongest disagreement with utilitarianism stems from the fact that the latter requires the sacrifice of the individual's interests whenever this is necessary for the greatest happiness of all. "Utilitarianism does not take seriously the distinction between persons" (1971, p. 27).[66]

4.5 Further rules for social evaluation. The variable population case

By introducing weaker invariance axioms than those just considered, more scope is given to interpersonal comparisons. Consider for example:

Axiom RS. (Ratio Scale Comparability)

$$\forall\, U \in \mathcal{U}^n, \forall\, \alpha > 0, F(U) = F(\alpha U).$$

This axiom has been used in the theory of income inequality, which in part connects with SWFL theory.[67] Under **(RS)**, **(I)**, **(SP)**, **(A)** and (C^*), restricting the domain $\mathcal{U}^n$ to positive-

[65] Arrow reviews *A Theory of Justice* in (1973a), while in (1973b) he critically investigates the dynamic properties of Rawls's theory, again by using a welfarist model.

[66] Some critics (such as Temkin, 1993, and Glannon, 1995) have argued that Rawls is concerned only about the distinction between persons among the worst off, since maximin (and to a lesser extent, leximin) leads to the neglect of losses, however large, when they are incurred by the better off.

[67] See in particular Blackorby and Donaldson (1978). The recent survey by Blackorby, Bossert and Donaldson

valued utility functions, one gets a large class of social welfare orderings, namely all orderings R^* that are representable by some increasing, symmetric, homothetic[68] and continuous "social-evaluation function" W from $\mathbb{R}^n_{++}$ to $\mathbb{R}$: aR^*b – $W(a) \geq W(b)$.

More generally, one could even assume *extreme comparability* – i.e., no invariance axiom at all. Then, under the same other axioms, one would obtain an even larger class of rules: all social welfare orderings (SWO) R^* that are representable by some increasing, symmetric and continuous "social-evaluation function" W from $\mathbb{R}^n$ to $\mathbb{R}$. There is one representation of particular interest: the *equally distributed equivalent* utility function w, which is uniquely defined by the equation:

$$W(a) = W(w(a), w(a), \cdots, w(a)),$$

for every a in the domain of definition of W. This notion is relevant to the theory of inequality measurement. A related concept will be mentioned in the context of fairness theory (see Section 6.5).

The equally distributed equivalent representation turns out to be useful also in some extensions of welfarism to the variable-population case. The ethical problems involved in population policies, life-saving social decisions, transfers to future generations, and so on, go far beyond the scope of this chapter.[69] We want, however, to indicate what new axioms are required to derive welfarist rules in this novel context. This exposition is limited to the particular approach introduced by Blackorby and Donaldson (1984). Like these authors, we assume that utility values are defined for the individual's life taken as a whole, not for each period of his life: this is the "lifelong utility" assumption, which by-passes the problems involved in constructing this aggregate from more elementary utility data.

The Pareto conditions, **(PI)** and **(SP)**, and the independence axiom **(I)** will remain un-

(1995a) also explores various connections between the two theories.

[68]Homotheticity means that W can be written as a composed function $\varphi \circ \lambda$, where λ is homogeneous of degree 1 and φ is monotonic.

[69]See in particular the collection by Silora and Barry (1978), Parfit's (1984) important book, and the continuing discussion on the latter. Among others, Hammond (1988b), Cowen (1989) and Brome (1992, 1996) discuss population issues.

changed. But they might have new implications when applied to alternatives involving populations of different sizes. Blackorby and Donadson argue that **(PI)** rules out social discounting of future utilities, a consequence which many authors would find undesirable. Axiom **(A)** should be formally modified to deal with comparisons involving two populations of the same size but with different individuals: under these circumstances, utility vectors that are identical up to a permutation will be declared to be socially indifferent. These axioms lead back to welfarism, but the social-evaluation function is now denoted by W^n, since the size n of the population determines the dimension of the utility vector. Then, assuming (C^*), we can again define the equally distributed equivalent utility function as an alternative representation of W^n; denote it by w^n.

The rules considered here will be variable population variants of utilitarianism. All involve the notion of a *minimal utility level*, to be interpreted as the level attached to a life which is just worth living, and normalized to zero. A first rule to consider is the extension of sum-utilitarianism: for any positive integer n and any $a \in \mathbb{R}^n$, $W^n(a) = \sum_{i=1}^n a_i$. The corresponding equally distributed equivalent is defined by average utility: $w^n(a) = 1/n \sum_{i=1}^n a_i$. For any two utility vectors, $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, aR^*b if and only if $W^n(a) = \sum_{i=1}^n a_i \geq W^m(b) = \sum_{i=1}^m b_i$, or equivalently in terms of averages, $nw^n(a) \geq mw^m(a)$. After the philosopher Parfit (1984), this utilitarian rule has been criticized for leading to the so-called *repugnant conclusion*: for any two positive average utility levels, assigning the lower one, however low, to a large enough population, will be preferred to assigning the larger one to a smaller population.

Mean utilitarianism provides one way to overcome this difficulty. Formally, (variable population) mean utilitarianism amounts to taking as a social evaluation function the equally distributed equivalent of (variable population) sum-utilitarianism. Namely, for any two utility vectors, $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$,

$$aR^*b \text{ if and only if } 1/N \sum_{i=1}^{n} a_i \geq 1/m \sum_{i=1}^{m} b_i.$$

Blackorby and Donaldson (1984) provide a simple characterization in terms of standard utilitarian axioms, i.e., **(SP)**, **(A)**, **(CU)**, and the following specific condition that they attribute

to the early 20th century economist Wicksell.

Wicksell Population Principle:

For any two utility vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^{n+1}$ such that $a_i = b_i$ for $i = 1, 2, \cdots, n$, bR^*a if and only if $b_{n+1} \geq w^n(a)$.

However, Blackorby and Donaldson choose to avoid the "repugnant conclusion" by defending a less "elitist" population principle than mean utilitarianism. They propose the following two principles, the first of which involves a critical utility level which they take to be positive, the other being a separability condition:

σ-Pareto Population Principle ($\alpha > 0$):

For any two utility vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^{n+1}$ such that $a_i = b_i$ for $i = 1, 2, \cdots, n$, bR^*a if and only if $b_{n+1} \geq \alpha$.

Population Substitution Principle:

For any two utility vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, $w^{n+m}(a, b) = w^{n+m}(w^n(a), \cdots, w^n(a), b)$.

The point of taking α positive is to avoid the "repugnant conclusion". (However, one should note than an "α-repugnant conclusion" can be opposed to critical-level utilitarianism.) In the presence of axioms **(SP)**, **(A)** and (C^*), these two principles characterize the class of SWO represented by *Critical-Level Generalized Utilitarian* rules (with the critical level fixed by α):

$$W^n(a) = h\left[\sum_{i=1}^{n}(g(a_i)) - g(\alpha))\right],$$

for some monotonic transformations h and g (where g is continuous).

Implicitly, the role of the weighting function $g(\cdot)$ is to take account of the social evaluator's attitude to inequalities in the utility distribution; typically, welfare inequalities will be smoothed out by one's choice of a concave $g(\cdot)$. The definition of generalized utilitarianism here is not the same as that used in Section 4.4; the weighting function can be (and typically is) non-linear.

This result, as well as the above characterization of mean utilitarianism, corresponds to a *static* framework of social choice. The only difference between this setting and that of Section 4.4 is that alternatives now involve populations of variable size. There is still no time

dimension involved in the social choice problem. Blackorby, Bossert and Donaldson (1995) attempt to remedy this by constructing a framework of dated individuals and principles. The main innovation in their framework is the principle of *Independence of the Utility of the Dead* (IUD), which states that the ranking of two alternatives remains unchanged when the dead are removed from these alternatives.[70] Principles like this make it possible to distinguish between individuals according to their birth and death dates, so that history matters to a certain extent. Technically, (IUD) plays in the intertemporal setting a role similar to the α-Pareto Population and the Population Substitution Principle taken together, so that this new axiom is the basis for another axiomatization of critical-level generalized utilitarianism.[71]

4.6 Some conceptual problems of the multi-profile approach

In 4.1 we emphasized the difference between Arrow's preference-based notion of a social welfare function (SWF) and Sen's and his followers' notion of a social welfare functional (SWFL). However, the two concepts share an important common feature: they belong to the *multi-profile* approach of social choice theory. In the *single-profile* approach only one vector of individual utilities or preferences is considered at a time. Then, it is hardly relevant to consider the collective preference as a "function" or "functional" of individual items (although, trivially, a function can be defined on a singleton domain). By contrast, the multi-profile approach makes it possible to consider *several* vectors of individual utilities or preferences at a time, and accordingly to relate these items to the collective preference in a truly functional way. To express the distinction between the two kinds of approaches in completely formal terms is a delicate

[70] Hammond (1988b) had also used a version of this principle.

[71] Loosely related to the theme of this subsection is the recent discussion of infinite utility streams, as in Nelson (1991), Vallentyne (1994), and Van Liedekerke and Lauwers (1997). These and other writers investigate the Pareto principle, the anonymity axiom, as well as several variants of utilitarianism, in the case of *infinite-dimensional* vectors of utility values – a mathematically natural extension of the welfarism framework which may be relevant to population issues. To avoid the paradoxes of infinity, standard axioms and social choice rules have to be reformulated using abstract methods; see in particular Lauwers (1995).

task.[72] But it is normally easy to recognize whether a particular axiom belongs to one or the other (or both). A moment's thought shows that the Pareto principle can be expressed in both frameworks. The conclusions it implies about collective preference depend on just considering one profile of individual utilities or preferences at a time. Conceptually, the Pareto principle is of the single-profile type. But it can also be expressed unproblematically in the multi-profile language of SWF or SWFL theory (as in Section 4.1 above). By contrast, Independence of Irrelevant Alternatives, as first introduced by Arrow (1951, p. 26), is a specifically multi-profile axiom. The conclusions it implies depend on comparing *two* profiles with each other; it states a low of variation of collective preference in terms of the relevant individual variables.

In the early days of social theory, objections were often raised against the multi-profile approach.[73] Since these objections hit the SWFL and the SWF frames of analysis with equal force, they should be reviewed here, albeit sketchily. The critics' common theme was that "laws of variation", such as Arrow's Independence axiom, assumed that analogies and disanalogies in individual preference profiles can be ascertained meaningfully, which they claimed was impossible. As Little put it forcefully:

> "we do not require that the difference between the new and the old ordering should bear any particular relation to the changes of taste which have occurred. We have, so to speak, a new world and a new order, and we do not demand correspondence between the change in the world and the change in the order" (1952, p. 423).

Even if one restricts the interpretation of preferences to tastes, as Little does in this famous comment, the critical point appears to be questionable. It is hard to see why tastes could not be compared with each other. A given individual's tastes in profiles 1 and 2 can be compared to the extent, for instance, that they lead to the same preference for wine over bread. Arrow's independence axiom does not require more than unproblematic comparisons of that sort. Presumably, Little's conviction that different individual preference profiles are like incommunicable worlds depends on the deeper point that there is no sense in considering individuals apart from

[72] See Rubinstein's (1984) elucidation in terms of mathematical logic.

[73] See Little (1952), Bergson (1954), and Samuelson (1967).

the preferences they actually have. He is in effect questioning the meaning of the expression "a given individual's tastes in profiles 1 and 2", in the last sentence. The social choice theorist should not be misled by his use of the same index i in profiles 1 and 2. Different *profiles*, the argument goes, refer to different *populations*. This line of criticism leads to the conclusion that Arrow's independence axiom involves a misunderstanding of the individualistic foundations of welfare economics. Once it is rephrased in this way, the Little-Bergson attack on the independence axiom becomes a conceptually relevant objection (though its exact force remains to be assessed). Notice that it appears not to depend on a taste interpretation of individual preference.

When technical welfarism is assumed right from the beginning, the distinction between the single-profile and the multi-profile approaches becomes largely irrelevant. Any social welfare ordering R^* might derive from either a multi-profile framework *or a suitably enriched single-profile framework*. Formally, take $\overline{U}$ to be any given profile of utility functions in $\mathcal{U}^n$, and $[\overline{U}]$ to be the class of relevant transforms of $\overline{U}$ (according to some chosen invariance condition). We introduce a SWFL restricted to $[\overline{U}]$ and require it to be constant on this domain. Now, we impose a "richness" condition on the domain, as well as a condition intended to play the role of Independence of Irrelevant Alternatives in SWFL theory. Respectively:

Axiom Unrestricted Utility Profile:

> For all a, b and c in $\mathbb{R}^n$ there exist x, y and z in X such that for some $U \in [\overline{U}]$,
> $U(x) = a, U(y) = b,$ and $U(z) = c,$

and

Axiom Relative Neutrality:

$$\forall\, U, U' \in [\overline{U}], \forall\, x, y, x', y' \in X,$$
$$U(x) = U'(x') \text{ and } U(y) = U'(y') \;\Rightarrow\; xF(U)y \text{ iff } x'F(U')y'.$$

These two axioms together imply that F can be extended to the complete domain $\mathcal{U}^n$ in such a way that both **(I)** and **(PI)** hold (see d'Aspremont, 1985, for a proof).[74] It then follows (by applying Lemma 4.1) that technical welfarism can be recovered in the *enlarged single-profile*

[74] An extension result can also be proved for a profile of VNM utility functions. For further elaboration of the

framework just introduced. This demonstrates that the welfarist approach of this section does not need the full force of the multiprofile approach, and thus has broader applicability than standard SWFL theory.

5 The aggregative setting with choice-theoretic constraints

5.1 Harsanyi's approach to utilitarianism. The aggregation theorem

Harsanyi's contribution to ethics is contained in two seminal papers (1953, 1955), his book *Rational Behavior and Bargaining Equilibrium* (1977a), and a number of philosophical or interpretative papers (in particular 1977b; see also his 1977c *Essays* and his 1992 restatement). The classic three-page 1953 article states Harsanyi's version of the "original state" or "veil of ignorance", to be compared with Rawls's (1971) altogether different version. This approach led Harsanyi to formulate the Impartial Observer Theorem: assuming that individuals value social positions as if they did not know who will hold them, and that this ignorance is captured by the VNM theory of risk, as applied to equiprobable lotteries, he concludes that the mean rule of utilitarianism prevails. The equally classic 1955 article states the following Aggregation Theorem: assuming a profile of individual utilities and a social or ethical utility, all of which are VNM functions on a lottery set, the Pareto-Indifference condition implies generalized utilitarianism, i.e., that the social utility is a weighted sum of the individuals' utilities. Harsanyi takes up the two theorems in his (1977a) book, where he also provides a direct philosophical defence of utilitarian interpersonal comparisons of utility. The common theme of the three piece is, of course, utilitarianism, although the mathematical form of the rule is not quite the same from one to the other – a technical problem which will be addressed below. Harsanyi's philosophy, if not always his formalism, leans towards *symmetric* utilitarian rules, and more particularly the mean rule which he thinks is superior to them ore popular sum rule.[75]

enlarged single-profile approach sketched in this paragraph, see Roberts (1980b), d'Aspremont (1985), and the earlier contributions referenced in these papers.

[75]However, Harsanyi does not normally discuss the variable population case. He also claims to be a rule – rather than an act-utilitarianism (e.g., 1977c), and provides philosophical arguments in favour of the former variant, but

The crucial philosophical problems raised by Harsanyi's two theorems are that, for one, they do not state explicitly the interpersonal comparison axiom that is necessary for utilitarian rules to hold; for another, they are not clearly connected with any ethically relevant concept of utility. These two criticisms have led Sen (1986) and Weymark (1991) to the strong negative conclusion that Harsanyi's theorems had no ethical relevance after all. We shall here take the more moderate view that the theorems are ethically significant, but that a nontrivial argument is needed to bridge the gap between the formal results and utilitarian ethics. This conclusion is common to Broome (1991a), Hammond (1987), and Mongin (1994a), even if they differ in their interpretations of Harsanyi's contribution.

A further important point for discussion in Harsanyi's method of deriving ethical conclusions from decision-theoretic premisses. He goes as far as to claim that "ethics is a branch of the general theory of rational behavior" (1977b, p. 42). Specifically, the crucial premiss in both the Impartial Observer and the Aggregation Theorems is that the individuals as well as the observer are VNM decision-makers. We shall reserve the discussion of the former theorem for Section 6. The present section is concerned with the latter and its numerous variants, applications and criticisms. It is interesting to record Harsanyi's own judgement on the relative strengths of his two major contributions: "[the Aggregation Theorem] yields a lesser amount of philosophically interesting information about the nature of morality than [the Impartial Observer Theorem], but it has the advantage of being based on much weaker – almost trivial – philosophical assumptions" (1977b, p. 48).

Proposition 5.1 (Harsanyi's aggregation Theorem) *Assume that the set of alternatives is* $X = \Delta_s(\Gamma)$ *or* $X = \Delta(\Gamma)$ *for some outcome set* Γ. *Assume also that the individual utilities* $U = (u_1, \cdots, u_n)$ *and the social utility* u_0 *are VNM functions on* X. *Then, the following condition holds:*

$$\textbf{(PI')} \quad \forall\, x, y \in X, U(x) = U(y) \;\Rightarrow\; u_0(x) = u_0(y)$$

this distinction appears to play no role in the interpretation of his two theorems. (To the best of our knowledge this distinction is not discussed in the choice-theoretic literature.)

if and only if there exist real numbers $\alpha_1, \cdots, \alpha_n$, β *such that:*

$$u_0 = \sum_{i=1}^{n} \alpha_i u_i + \beta.$$

Harsanyi's initial proof lacked definiteness, while his later argument (1977a, 4.8) involved irrelevant algebraic independence restrictions. Uselessly complicated constructions have been erected around Harsanyi's result, which can be proved most simply, as the following shows.

Proof: [Coulhon and Mongin (1989)] Denote $(u_1, \cdots, u_n)$ by U. Condition **(PI')** is equivalent to the property that $u_0 = f \circ U$ for some function f: Range $U \to \mathbb{R}$. The following shows that f is mixture-preserving (MP): for any $Y, Y' \in$ Range U, there are $y, y' \in X$ such that $U(y) = Y, U(y') = Y'$, and

$$\begin{aligned} & f(\lambda Y + (1-\lambda)Y') = f(\lambda\, U(y) + (1-\lambda)U(y')) \\ = \;& F(Y(y\lambda y')) = u_0(y\lambda y') \\ = \;& \lambda u_0(y) + (1-\lambda)u_0(y') = \lambda f(Y) + (1-\lambda)f(Y'). \end{aligned}$$

(Notice that the role of the assumption that both U and u_0 are MP). From the MP property of U, it is clear that Range U is convex. It is a fact that if a real function is MP on some convex subset of $\mathbb{R}^n$, it is affine on that subset. We conclude that u_0 is affine in terms of the u_i. ■

For simplicity we took X to be a lottery set but the proof carries through without change to the more general case in which X in convex subset of any vector space whatever.[76] This last observation delivers an interesting *non-stochastic* variant of Harsanyi's Aggregation Theorem. Take the alternative set X to be some convex subset of $\mathbb{R}^p$, as in consumer theory, and take the utility set $\mathcal{U}$ to be the set of all affine functions on X. These assumptions define a particular "economic" domain in the sense of Section 3.1; utility functions are strongly separable in each commodity. Now, if $u_0, u_1, \cdots, u_n \in \mathcal{U}$ satisfy **(PI')**, we can conclude that u_0 is affine in terms of the u_i, exactly as in the VNM case.[77]

[76] Actually the proof above applies to the even more general case in which X is a *mixture set* [in the sense of Herstein and Milnor (1953)].

[77] Hammond (1996) provides another nonstochastic variant of Harsanyi's Aggregation Theorem in which individual utilities $u_1, \cdots, u_n$ are defined on endowment vectors but depend only on the individual's own component.

The following features of Proposition 5.1 should be carefully recorded: (i) the derived coefficients $\alpha_1, \cdots, \alpha_n$ can be of any sign; (ii) they can be of any magnitude; (iii) the result is a single-profile one, i.e., the coefficients $\alpha_1, \cdots, \alpha_n$ depend on the given utility profile $u_0, u_1, \cdots, u_n$. Hence the conclusion reached above is still at a distance from Harsanyi's theoretical target, which was to derive the mean rule of utilitarianism. Feature (i) is the most troublesome of all. Utilitarianism can hardly remain an interesting ethical doctrine, if it is extended to the point of involving negative weights for some individuals. Feature (ii) would be compatible with the more plausible concept of *generalized utilitarianism*, which was introduced within the SWFL framework in Section 4.4. However, this is *not* Harsanyi's brand of utilitarianism, which follows a longstanding tradition in requiring weights to be equal. Finally, (iii) raises the equally important question of whether or not the individuals' weights should just depend on their identity (which is represented here by the index i) or *also on their utility functions*, as is the case in Harsanyi's own version of the Aggregation Theorem. Problem (iii) is already part of Sen's (1986) assessment of the ethical significance of Harsanyi's theorems. Except for this overlapping point, the discussion just sketched should be seen as a technical prerequisite to the Sen-Weymark critique.[78]

In both (1955) and (1977a) Harsanyi suggested that in order to remedy problem (i), it was enough to use a more demanding condition than the rather weak Pareto Indifference condition **(PI')**. The further results stated below clarify this intuition. We define the following strengthening of **(PI')**:

Axiom SP'

$$\textbf{(SP')} \equiv \textbf{(PI')} \ \& \ \textbf{(Strict P')}, \text{where } \textbf{(Strict P')} \text{ is}$$
$$\forall\, x, y \in X, U(x) > U(y) \Rightarrow u_0(x) > u_0(y).$$

It will also be interesting to investigate **(Strict P')** in isolation, as well as:

Axiom WP'

$$\forall\, x, y \in X, U(x) \gg U(y) \Rightarrow u_0(x) > u_0(y).$$

[78] This critique questions the philosophical relevance of the two theorems, but at least takes for granted that they are successful technically, i.e., that they derive rules which are *formally* identical to standard utilitarian rules. As just explained, the conclusion of Proposition 5.1 falls short of this requirement.

Neither **(Strict P')** nor **(WP')** implies **(PI')**. As their labelling suggests, the various Pareto conditions introduced here exactly parallel those of Section 4.

Now, the following proposition shows that **(SP')** exactly conorms with Harsanyi's intuition. But to reach similar results under **(WP')** and **(Strict P')**, a restriction of *minimum agreement among individuals*– **(MA)** below – must be added:

Proposition 5.2 Harsanyi's aggregation theorem with stronger Pareto conditions[79] *Assume that the individual utilities* $U = (u_1, \cdots, u_n)$ *and the social utility* u_0 *ae VNM functions on* $X = \Delta_s(\Gamma)$ *or* $X = \Delta(\Gamma)$.

1. *If* **(SP')** *holds, there exist positive numbers* $\alpha_1, \cdots, \alpha_n$, *and a real number* β *such that* $u_0 = \sum_{i=1}^{n} \alpha_i u_i + \beta$.

2. *Assume further that:*

   (MA) *There are* $x, y \in X$ *with the property that* $u_i(x) > u_i(y), i = 1, \cdots, n$.

Then, if **(WP')** *[***(Strict P')***] holds, there exist nonnegative numbers, not all of them zero [resp. positive numbers]* $\alpha_1, \cdots, \alpha_n$, *and a real number* β *such that* $u_0 = \sum_{i=1}^{n} \alpha_i u_i + \beta$.

Interestingly (and somewhat counterintuitively), the weak agreement condition **(MA)** can be derived from the assumption that individual utilities strongly differ from each other, or more technically, are affinely independent.[80]

We still have to deal with two technical issues among the three listed after Proposition 5.1. As it turns out, the technical problems (ii) and (iii) can be addressed at the same time by shifting from Harsanyi's initial framework to a SWFL one.[81]

[79] This proposition is borrowed from De Meyer and Mongin (1994). It is proved using convex analysis.

[80] See Weymark (1993, Proposition 5.2). Mongin (1995a) proves the corresponding statement for SEU theory (i.e. when the alternatives are acts).

[81] Among further relevant *single-profile* variants of Proposition s5.1 and 5.2, Zhou (1997) extends Proposition 5.1 to the case of an infinite population, and Blackorby, Donaldson and Weymark (1996) investigate a variant of Proposition 5.1 in which VNM lotteries are replaced with subjective probabilities which are identical from one individual to another. This last modelling is really a borderline case between Harsanyi's VNM aggregative framework and the SEU framework discussed at length in Section 5.4.

5.2 A SWFL reconstruction of Harsanyi's aggregation theorem

As in Mongin (1994a), the present reconstruction relies on the following social-choice-theoretic concepts: $X = \Delta_s(\Gamma)$ or $X = \Delta(\Gamma)$ for some outcome set Γ; X is required to have (vector space) dimension at least 2; and $\mathcal{U} = \mathcal{V}(X)$, i.e., the set of VNM functions on X. The SWFL

$$F : \mathcal{U}^n \to 2^{X \times X}$$

will be investigated under the special assumption that for all $U \in \mathcal{U}^n$, $F(U)$ satisfies VNM properties. Thus, we shall incorporate into the SWFL framework Harsanyi's assumption that both the individuals and the social aggregate are VNM maximizers. Let us define formally:[82]

Axiom VNM1

For all $U \in \mathcal{U}, F(U)$ satisfies the following continuity property:
$\forall\, x, y, z \in X, \{\lambda \in [0,1] : zF(U)(x\lambda y)\}$ and
$\{\lambda \in [0,1] : (x\lambda y)F(U)z\}$ are closed subsets of $[0,1]$.

Axiom VNM2

For all $U \in \mathcal{U}, F(U)$ satisfies VNM-independence, i.e.:
$\forall\, x, y, z \in X, \forall\ \lambda \in]0,1], xF(U)y - (x\lambda z)F(Y)(y\lambda z)$.

In order to complete the literal translation of Harsanyi's assumptions into the new framework, it is enough to add that F satisfies **(PI)** or some alternative Pareto axiom. However, if one stopped at that, the SWFL exercise would plainly be useless. The whole point of shifting to the SWFL framework is that it makes it possible to formulate axioms and (hopefully) results which, unlike **(VNM1), (VNM2)** and the Pareto conditions, relate to *several* utility profiles at a time. This is the essence of the *multi-profile* approach to social choice theory (see Section 4.6). Two relevant axioms are the already defined Independence of Irrelevant Alternatives **(I)** and

[82] These two axioms are adapted from one among the many axiomatizations of expected-utility representations. Of the two, **(VNM2)** is clearly the crucial one from the conceptual point of view. Axiom **(VNM1)** is introduced here as elsewhere for the well-known reason that numerical representations of preferences require some continuity condition to hold.

Anonymity **(A)**. By adding them to the literal translation of Harsanyi's assumptions, one can hope to strengthen his results in the right direction. Specifically, it is likely that by adding **(I)**, one will cancel the dependence of individual weights a_i on the given utility profile, and by further adding **(A)**, one will derive a symmetric summation rule, as required by standard utilitarianism. In brief, the SWFL approach is tailor-made to supersede the technical problems that remained unsolved in view of Propositions 5.1 and 5.2. The following proposition fulfils these expectations, while clarifying the connection between Harsanyi's VNM assumptions with earlier utilitarian conditions:

Proposition 5.3 *Take X and $\mathcal{U}$ as defined in the VNM case, and assume throughout that F satisfies* **(I)** *and* **(PI)**. *Then:*

(i) F satisfies **(VNM1)** *and* **(VNM2)** *if and only if there exists a vector $(\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$, unique up to a positive scale factor, such that for all $U = (u_1, \cdots, u_n) \in \mathcal{U}^n$,*

$$\forall\, x, y \in X, xF(U)y - \sum \alpha_i u_i(x) \geq \sum \alpha_i u_i(y).$$

(ii) Assuming **(Strict P)**, *the following holds:*

(ii.1) **(CU)** *and* **(VNM2)** *are equivalent to each other, and equivalent to Weak Utilitarianism (with strictly positive coefficients);*

(ii.2) The two pairs of conditions **(CU)** *&* **(C)** *and* **(VNM1)** *&* **(VNM2)** *are equivalent to each other, and each is equivalent to Generalized Utilitarianism (with strictly positive coefficients).*

(iii) Standard utilitarianism follows from adding **(A)** *to any of these restatements of Generalized Utilitarianism.*[83]

Thus, Proposition 5.3 solves the two remaining technical difficulties surrounding Harsanyi's formulation of the Aggregation Theorem. Its merit lies with its simplicity. Its weakness is

[83]For part (i), see Mongin (1994a), noticing the application of technical welfarism to the special VNM domain, as in Lemma 4.1 (ii) above. Part (ii) can be checked directly, which then leads to a variant proof of d'Aspremont and Gevers's (1977) **(CU)**-based characterization of utilitarian rules.

that the **(I)** axiom, for one, appears to be external to Harsanyi's initial problem-situation; for another, it gives rise to criticisms in its own right. In connection with the former point, we note that Hammond (1987) has also argued for a reconciliation of Arrow's social choice theory with Harsanyi's brand of utilitarianism. Such a *rapprochement* has the heuristic advantage of making comparisons easier, especially when it comes to discussing interpersonal comparisons of utility in Harsanyi's approach. In connection with the latter point, relevant criticism of Independence of Irrelevant Alternatives have already been raised in Section 4.6.

At least the following simple fact about the SWFL analysis should be clear. It appears to be impossible to derive classical, i.e., equal weights utilitarianism, without imposing either axiom **(A)**, or some variant *which must again be an interprofile axiom*. Hence, if not SWFL theory itself, an enriched single-profile approach, as introduced at the end of Section 4.6, is indispensable if one wants to move from Harsanyi's weighted sum formula to one which makes sense from the point of view of utilitarian philosophy. In the initial version of Harsanyi's theorem, which is purely single-profile, symmetric additive rules have – allegedly – been derived from suitable rescalings of individual utilities. From the axiomatic point of view this is a highly unsatisfactory procedure. The primitives of the reasoning are *fixed* numerical representations $u_0, u_1, \cdots, u_n$; they can be replaced with representations $u'_0, u'_1, \cdots, u'_n$ only if there is an invariance axiom to warrant their replacement. Whatever its exact wording, such an axiom cannot be insignificant, since it will indicate something about permissible and impermissible interpersonal comparisons of utility. We conclude that the usual "derivation" of symmetric rules is faced with a dilemma: either it relies on (some variant of) Proposition 5.3 above, or it should be objected to as being an *ad hoc* procedure.[84]

5.3 Further philosophical comments on Harsanyi's utilitarianism

A major objection raised against the ethical relevance of Harsanyi's Aggregation Theorem is that he implicitly relies on an interpersonal comparison assumption, but fails to make it clear which assumption it is. As we shall argue, this objection is misconceived. The whole point of

[84]This comment applies to Harsanyi's initial argument as well as to Broome's (1991a, chapter 10) exposition.

Harsanyi's approach is to show that if some choice-theoretic (i.e., VNM) and social-choice theoretic (i.e., Paretian) assumptions hold, then the collective preference conforms with a utilitarian interpersonal comparison principle. That is to say, the Aggregation Theorem is interesting just because it *derives* the relevant principle. Recall Harsanyi's own judgement as restated in Section 5.1: he regards the choice – and social-choice – theoretic premisses of the theorem as "trivial". No doubt, he does not regard the principle of utilitarian interpersonal comparisons as "trivial"; hence the interest in a formal proof that the former implies the latter. Harsanyi might be wrong in believing that his premisses are unproblematic; this would be a fair criticism. But to insist on reformulating the premisses so as to make the role of interpersonal comparisons apparent right from the beginning is, in our opinion, to miss the fine point of the Aggregation Theorem completely.

Admittedly, Harsanyi's initial single-profile formulation does not facilitate the analysis of underlying interpersonal utility comparisons. The SWFL reformulation is clearer than the initial one in this respect as well as in others. The essential message of the Aggregation Theorem is perhaps best summarized in Proposition 5.3 (ii.1): granting the decision-theoretic and social-choice-theoretic axioms, to assume that the social preference satisfies the VNM independence axiom is tantamount to assuming that it conforms with the utilitarian invariance principle **(CU)**. There are several ways in which this logical equivalence can be turned into an ethical argument. For instance, Proposition 5.3 "could be used against anybody who would be prepared to swallow **(I)** and "could be used against anybody who would be prepared to swallow **(I)** and **(SP)**, make VNM assumptions reflect individual rationality – and, say, turn Rawlsian, or hostile to any interpersonal comparison whatsoever, when it comes to assessing income distribution" (Mongin, 1994a, p. 349).[85]

[85]Broome's (1991a, pp. 219–220) interpretation is different, though compatible with the view that Harsanyi's derivation involves a "surprise effect". This interpretation emphasizes the role of the *completeness* axiom imposed on social preference. The latter assumption implies that the ideal observer can rank situations of conflicting individual preferences (i.e., situations x and y such that i strictly prefers x over y and j strictly prefers y over z). This suggests a close connection, between completeness and the admission of interpersonal comparisons of utility. Importantly, the completeness axiom does not indicate what *kind* of comparisons – i.e., utilitarian or otherwise –

The discussion above suggests that Harsanyi's critics should redirect their attack towards his *explicit* assumptions. Both Sen (1986) and Weymark (1991) indeed follow this strategy by questioning Harsanyi's exclusive reliance on VNM utility representations of social and individual preferences.[86] To analyze this further objection, one should carefully distinguish between two kinds of theoretical commitments, one to the VNM axioms imposed on the *preference* relation, the other to the use of VNM *utility functions* as representations of the underlying preference relation.

Philosophically, this distinction is in accord with Harsanyi's own view of utility as representing preferences of some kind. Harsanyi (e.g., 1977b, p. 54) has emphatically rejected the early utilitarian writers' interpretation in terms of net pleasure. He claims to be a *preference* utilitarian, not a *hedonistic* utilitarian, as was Bentham, or an *ideal* utilitarian, as Moore is sometimes described. This is not to say that he follows the welfare economists' footsteps in defining utility as representing *actual* preferences. To the contrary, he adheres to – and in our opinion should count as a major representative of – that philosophical school we discussed in Section 2.5 which takes utility functions to be relevant to the ethical exercise only if they represent *improved* rather than actual preferences. In Harsanyi's special terminology, "moral" preferences are constructed by aggregating "personal" preferences only after the latter have been "corrected" and "censored" (see, e.g., 1977a, pp. 61–62). Correction is needed because some preference judgements might depend on factual errors; the ideal observer does not make these mistakes, and should thus appropriately modify the individual's preference ranking. Harsanyi mentions the example of an individual who wrongly believes that a certain drug is efficacious when it is not. There are trickier cases than this one, and it is not clear where an enlighted observer should stop improving "personal" preferences in this way. Harsanyi suggests that at least in principle, it is possible to draw a line: the observer "will be justified in using a corrected utility function

are made.

[86] Weymark (1991) also suggests restating Harsanyi's theorems by including among the axioms some *very general* comparability assumption (such as Cardinality and Full Comparability). This restatement is again compatible with the view that the added value or "surprise effect" of the Aggregation Theorem consists in deriving the *utilitarian* form of comparison. Formally, it would require moving to a SWFL or related framework.

(for j) only if he thinks that j himself would approve of this" (1977a, pp. 61–62). Notice the recurring methodological theme: the preference concept relevant to ethical applications can only be delineated in terms of some counterfactual experiment (see Section 2.5).

As to censoring "personal" preferences, à la Harsanyi, this does not simply mean that irreflective preference comparisons, or those resulting from the individual's weakness of the will, should not be considered by the observer. Presumably, such a preliminary laundering of "personal" preferences can be taken for granted. It can be analyzed in terms of a counterfactual clause modelled after the previous one: under suitable circumstances, the individual *himself* would approve of the observer's overruling his initial preference. Censure refers to a deeper problem: some *considered* preference judgements are, Harsanyi claims, ethically inadmissible. The observer "will be perfectly justified in disregarding j's actual preferences in cases where the latter are based on clearly antisocial attitudes, e.g., on sheer hostility, malice, envy, and sadism" (1977a, p. 62).[87] As critics have noted, it seems unsatisfactory to recommend preference censoring, and Harsanyi himself sometimes comes close to recognizing that there is a serious difficulty here. But it is important to bear in mind that this is a general difficulty with *any welfarist approach* to social ethics. The sadist plagues Harsanyi's theory in the same way, roughly, as the fanatic does Hare's (see Section 2.4). In other words, if Harsanyi's defence of utilitarianism strikes one as being inconsequent, there is nothing here that is specific to him, nor even to the particular brand of utilitarianism he adopts. As we suggested earlier, there are ethical objections that welfarism appears to be inherently incapable of answering. The inadequacy of Harsanyi's recommendation of censoring "personal preferences reflects one of the 'limits of well-being'".

Moving on to the technical side of the utility-preference distinction, we remind the reader that the uniqueness part of the VNM representation theorem does *not* say that *only* expectational (or mixture-preserving) functions represent those relations which satisfy the ordering, continuity and VNM-independence axioms [see Fishburn (1970) and (1982), and Hammond's chapter on "objective" expected utility]. Actually, any nonlinear, strictly increasing transform of the VNM

[87] Notice the following variant: Harsanyi (1992, p. 704) has recently recommended censoring *all* other-regarding preferences, including benevolent ones, whereas he initially meant to exclude only malevolent ones.

representation provided by the theorem also represents the given binary relation. Sen and Weymark are then asking: Why should Harsanyi restrict attention to VNM representations only?

An answer to this question can be sketched along the following lines. What is crucial about VNM representations is that those, and only those representations, preserve the risk-attitude properties of the VNM decision maker. For instance, it is well known that the Arrow-Pratt coefficient of absolute risk-aversion remains invariant if and only if expectational representations are used to represent the VNM preference.[88] Now, following the heuristic underlying Harsanyi's approach in both the Aggregation and Impartial Observer Theorems, risk-attitude properties are ethically important data to record. In Harsanyi's opinion there is a close connection between the amount of risk that theVNM decision maker is willing to take in order possibly to receive x, and the strength of his desire for x. As he stated most clearly, "The VNM utility functions do express people's attitudes to risk-taking (in gambling, buying insurance, investing and other similar activities). But they do not merely express these attitudes; rather they try to explain them in terms of the relative importance (relative utility) people attach to possible gains and possible loses of money and other economic and noneconomic assets" [Harsanyi (1977b) in Sen and Williams (1982, pp. 52-53)]. After other writers in the field, Harsanyi essentially claims that VNM indexes provide *cardinal* information on the individuals' and ideal observers' preferences that can be used *universally*, i.e., in the contexts of both risky and riskless choices. We are not suggesting that this strong claim should be endorsed, but just stressing that it is Harsanyi's most likely response to the Sen-Weymark critique. At least his position is internally consistent. Granting his prior conviction that VNM theory delivers cardinally relevant information, there would be no point for him to enlarge the set of utility representations beyond the class of VNM functions.[89]

The further question is, why should one accept Harsanyi's conviction that VNM theory de-

[88] See Pratt (1964) or any textbook on the economics of risk.

[89] An irrelevant objection here would be that Harsanyi cannot be at the same time a *preference* utilitarian and a cardinalist. We explained in Section 2.3 that the preference concept can receive a cardinal interpretation, as in Suppes and Winet (1955).

livers cardinally relevant information? Starting with the early days of expected-utility theory,[90] there has been an active debate about the meaning of cardinality (or "measurability", in some writers' terminology) in VNM theory. Two simple lessons can be drawn from this ever lively discussion. For one, the fact that VNM indexes are unique up to positive affine transformations just defines a mathematical notion of cardinality. It does not in itself lean towards any psychological interpretation. For another, there are two *prima facie* psychologically relevant interpretations of cardinality in the VNM context: one is Harsanyi's; the other is the – by and large now prevailing – opposite view that the numerical information contained in VNM indexes is relevant to *risky choices only*, and thus irrelevant to such issues as income distribution among the members of society (in which no chance mechanism is involved).[91]

We have separated interpretative questions pertaining to VNM indexes from those pertaining to theVNM preference axioms. The question now arises, whether or not these axioms should have been assumed in the first place. Even granting the point that they embody a normatively compelling modeling of rationality, and the further philosophical point that the ideal observer should aggregate *rationally formed* rather than empirically given individual preferences, it could be asked, why should the ideal observer *himself* comply with the VNM axioms? This last requirement turns out to be crucial to the utilitarian-like results of Harsanyi's two theorems. As far as the Aggregation Theorem is concerned, the equivalences stated in Proposition 5.3 (ii) make the technical contribution of **(VNM2)** very clear: this axiom carries with it the exact force of the utilitarian principle of comparison. In the context of both the Aggregation and the Impartial Observer Theorems, several writers have questioned the assumption that the social preference – as against individual preferences – should be subjected to VNM independence. Their criticism usually runs as follows. When it comes to evaluating social arrangements, one is interested in the *exact distribution* of utility over the individuals. It is not enough to know the *mathematical expectation* of that utility distribution [e.g., Sen (1970, p. 143); see also Sen

[90] See in particular Luce and Raiffa's discussion of "Fallacy 2" (1957, p. 32).

[91] We refer the reader to Fishburn (1970, 1989) and Bouyssou and Vansnick (1990) for careful reviews of these conflicting doctrines.

(1973b, 1986)]. When stated in this way, the objection to VNM independence becomes a variant of the classic objection against the distributional consequences of utilitarianism, an objection which is also endorsed by Rawls (1971).

A related but more specific argument results from a classic example of Diamond (1967). This example relies on a two-person society, and two lotteries x and y involving some equal chance device, say tossing a fair coin. Under x, whatever the result of tossing, individual gets everything, and j gets nothing. Under y, i gets everything if heads, and j gets everything if tails. The ethical intuition, Diamond believes, recommends ranking lottery y higher than lottery x, and at the same time, to express indifference between the following two outcomes: i has everything while j has nothing, j has everything while i has nothing. But a moment's thought shows that this conclusion violates VNM independence, which thus could not apply to the ethical observer's choices. This ingenious example has attracted considerable interest.[92] It broadens the discussion of the normative standing of VNM independence (as well as related axioms, such as Savage's sure-thing principle) to the different issue of *fair lotteries*.[93]

The positive argument in favour of imposing VNM independence on the social preference is, of course, *coherence*. Why should the observer be exempted from the rationality axioms to which individuals are subjected? This question has also been raised with respect to the Savage (or related) axioms of uncertain choice. Since the latter are – in a sense – more general than the VNM axioms of risky choice, we might as well postpone the required discussion to the end of Section 5.4.

5.4 The aggregative approach in the case of subjective uncertainty

Harsanyi's reliance on the VNM axioms implies that he takes the relevant state probabilities as given. Although several interpretations can be devised for this assumption, as suggested in Section 3.2, his aggregative approach is inherently incapable of dealing with public or moral

[92] See in particular the comments by Sen (1970), Harsanyi (1975), Hammond (1983), Broome (1991a), Epstein and Segal (1992), and Karni (1996).

[93] On the issue of fair lotteries, see Broome (1990-1991), Wasserman (1996), and the references listed in the latter paper.

situations in which not only utility assessments but factual opinions significantly differ from one individual to another. To be accurate, there are cases in which the role of factual disagreements can plausibly be chanelled through utility functions. For instance, one might conceive of the individual's diverging conditional expectations of lifespan as being appropriately reflected in their time preferences. But such construals should be used with much care. They clearly do not apply to the largest number of cases of relevant factual disagreements. Consider public debates over tax reform, illness prevention schemes, or environmental programmes. Each of these policies is likely to lead to significant discrepancies in the citizens' expectations of their practical consequences, independently of their valuations of consequences. Notice also that Harsanyi's recommendation of "correcting" preference judgements when they depend on mistaken beliefs can attenuate the citizen's differences of opinion, but is unlikely to eliminate them altogether. Following the Bayesian tenet, such truly factual and substantial disagreements can, and even should, be formalized by assuming that individuals entertain different subjective probabilities. Accordingly, SEU theory becomes the relevant frame of analysis to pursue Harsanyi's aggregative approach to social ethics.

Several writers have followed this line of inquiry Hylland and Zeckhauser (1979), Hammond (1981, 1983), Broome (1990, 1991a), Seidenfeld et al. (1989), Mongin (1995a). The general lesson from this lively strand of literature is that Harsanyi's aggregative approach runs into severe difficulties, as soon as one replaces his own VNM assumption with any axiomatization of SEU theory. Essentially, the Pareto conditions clash with the requirement that both the individuals and the aggregate follow the SEU axioms. The authors just mentioned either properly demonstrate, or illustrate in the case of a two-agent society, impossibility theorems to the effect that only special cases, such as dictatorial rules or uniform probability and/or utility profiles, satisfy the two subsets of conditions at the same time. The last paper in the list has perhaps achieved maximum generality in stating these various impossibilities. The above contributions differ from each other not only in their chosen auxiliary assumptions, but also in more crucial axiomatic respects. Hylland and Zeckhauser's (1979) – which might count as the historical source of the present literature – uses a rich axiomatic framework of social choice theory but no axiomatiza-

tion of SEU theory *per se*. The same comment applies to Hammond's (1983). The remaining contributions share the common feature that they do not use the language of social choice theory, but – sometimes implicitly – assume some axiomatic framework of Bayesianism. Mongin's (1995a) is based on Savage's axiomatization of SEU theory, but its results carry through without significant loss of content to the more accessible Anscombe-Aumann version. We shall follow this convenient variant here.

The notation will be the same as in Section 3.3. Each individual will thus be endowed with an AA-representation: for $i = 0, 1, \cdots, n$ and any act a,

$$(*) \qquad u_i(a) = \sum_{\omega \in \Omega} p_i(\omega) v_i(a(\omega)).$$

For technical reasons, we introduce the following assumption of *Minimal Agreement on Consequences*:

Assumption MAC

$$\exists\, c, c' \in \mathcal{C} \text{ s.t. } \forall\, i \in N, v_i(c) > v_i(c').$$

This added assumption amounts to strengthening the nontriviality requirement which enters any SEU axiomatization for rather obvious reasons.[94] We should not appropriately reformulate the various *Pareto conditions* envisaged by Harsanyi. To do so, it is enough to revert to the definitions of **(PI')**, **(SP')** and **(WP')** in Section 5.1, and everywhere replace $x, y \in X$ with $a, b \in X = \mathcal{C}^{\Omega}$. Thus reformulated, the Pareto conditions have an *ex ante* interpretation: if all individuals agree on the ranking of two state-dependent functions (or, more pictorially, bets) a and b, so does the social observer. This stipulation holds before uncertainty is resolved, and should carefully be distinguished from *ex post* versions of the Pareto principle, which will be discussed below. Notice that **(MAC)** has the effect of making **(WP')**, hence also **(SP')**, nonvacuous.

[94] Without it preferences between acts would not reveal the agent's subjective probability. What **(MAC)** adds to this standard requirement is that the nonconstancy of the U_i can be checked on a *uniform* choice of c, c'. A minimal agreement assumption was already used in the VNM context of Proposition 5.2.

The impossibility theorem below involves various concepts of dictatorship. We shall say that there is a *utility dictator* if $v_0 = v_j$ (up to a positive affine transformation) for some $j \in N$, and an *inverse utility dictator* if $v_0 = -v_j$ (up to a positive affine transformation) for some $j \in N$. Similarly, there is a *probability dictator* if $p_0 = p_j$ for some $j \in N$. The negative conclusions of the theorems depend on assuming affine independence of either the probabilities or the utilities or both. We explained in Section 3.4 that affine independence is an algebraic rendering of interindividual diversity.

Proposition 5.4 *The following assumptions hold. There is a finite state set Ω, a consequence set $\mathcal{C} = \Delta(\Gamma)$ for some outcome set Γ, and the alternative set preferences that are represented by utility functions $\mathcal{U} = (u_1, \cdots, u_n)$ and u_0 respectively, as in (*) above. The utility functions satisfy* **(PI')**. *Then,*

- *(i) If the $p_i, i = 1, \cdots, n$, are affinely independent, there is a utility or inverse utility dictator j; if furthermore the v_i, $i = 1, \cdots, n$, are pairwise affinely independent, j is also a probability dictator.*
- *(ii) Symmetrically, if the v_i, $i = 1, \cdots,$, are affinely independent, there is a probability dictator j; if furthermore the p_i, are pairwise distinct, j is also a utility or inverse utility dictator.*

Proposition 5.5 *The assumptions are as in Proposition 5.4, plus* **(MAC)**. *Then, if* **(WP')** *holds, the conclusions are as in Proposition 5.4 (i) and (ii), except that inverse dictatorship becomes impossible. Assume now that* **(Strict P')** *holds. Then,*

- *(i) It the $p_i, i = 1, \cdots, n$, are affinely independent, the v_i , $i = 1, \cdots, n$ must be pairwise dependent.*
- *(ii) If the $v_i, i = 1, \cdots, n$, are affinely independent, the $p_i, i = 1, \cdots, n$ must be identical.*

Propositions 5.4 and 5.5 should be compared with their VNM counterparts, i.e. Propositions 5.1 and 5.2 above. It turns out that in the case of the weaker Pareto conditions – i.e., **(PI')**

and **(WP')** – Harsanyi's utilitarian-like aggregation rule degenerates into a form of dictatorship (which is stronger than Arrow's). The conclusion is even more negative in the case of the strongest of the Pareto conditions considered here, i.e., **(Strict P')**. Then, the assumptions are shown to impose a constraint on the *data* of the aggregative problem rather than its *solution*. Under **(Strict P')** even dictatorship might fail to provide an aggregative rule. This further impossibility is not so much in the style of Arrow's as of those social choice results which state that some "natural" set of assumptions involves an outright logical contradiction. To make this clear, we may formulate the essential conclusion of Proposition 5.5 in a slightly different way. The following weaker, but perhaps more transparent statement holds:

Proposition 5.6 (This is a variant of Proposition 5.5) *Assume the following. There are n individuals; they, as well as the social observer, have AA preferences represented by utility-probability pairs $(v_1, p_1), \cdots (v_n, p_n)$, and (v_0, p_0), respectively.* **(MAC)** *holds, was well as the following Interindividual Diversity condition: the p_i and the v_i are affinely independent. Then, if* **(WP')** *holds, there is an individual $j \in N$ who is both a probability and a utility dictator. To impose* **(Strict P')** *instead of* **(WP')** *would lead to a logical contradiction.*

Subjective probability measures p_i and utility functions v_i play completely symmetric roles in Propositions 5.4 and 5.5. Some variants of Proposition 5.5 (ii) above have been discussed earlier under the label "probability agreement theorem", mostly by Broome (1989), (1990), (1991a)]. This expression is slightly misleading because it might wrongly suggest a constructive interpretation for the theorem. Contrary to convergence results in Bayesian statistics and the economics of information, the theorem here does not state *reasons* why subjective probabilities should become equal. In the other results, asymptotic equalization of subjective probabilities is due to shared information between individuals and successive revising of posterior probabilities in view of this common information. In the present framework, the involved probabilities can be interpreted as being either prior probabilities or posterior probabilities in a context of differing private information. Conceptually, there is absolutely no reason why they should be equal; this

is why the result must be interpreted as an impossibility theorem.[95]

Propositions 5.4 and 5.5 are derived by adapting the proof of their counterparts in Savage's framework (Mongin, 1995a, Propositions 5 and 7).[96] The technical argument need not be pursued here. The technical argument need not be pursued here. Suffice it to say that the crucial properly to use in the AA framework is the lottery (or convexity) property of the consequence set $\mathcal{C}$. By and large, this property will play the same role as did Savage's (1954) divisibility postulate (P6) in the corresponding proof. These alternative mathematical restrictions are wroth emphasizing for the following two reasons. First, in the absence of them, it is possible to construct counterexamples to Propositions 5.4 and 5.5.[97] Second, they happen to be the very same mathematical restrictions that utility theorists need in order to derive SEU representations from axioms on preferences over acts.[98] Hence, there is a tight connection between the paradoxes of collective Bayesianism, as restated here, and the *axiomatic* versions of that doctrine. It follows that at least in the present version, they are paradoxes of *preference* rather than of *utility* theory. On hte face of it, they concern *improved* preference theories no less than *actual* preference theories.

We have been stressing the connection between Propositions 5.4 and 5.5 and Harsanyi's research programme in ethics, but these propositions can receive an alternative and equally relevant interpretation in terms of a classic problem of welfare economics. When uncertainty rather than risk prevails, even assuming SEU theory as the relevant theory of choice under uncertainty, it remains to choose between two families of collective evaluation rules. *Ex ante* rules result from aggregating the individuals' Bayesian preferences or SEU functionals over acts. Such rules take the individuals' Bayesian preferences or SEU functionals over acts. Such rules take the individuals' subjective probabilities into account since these probabilities are embodied

[95]For modelling purposes, economists often assume that prior probabilities are equal. It is important to realize that this "common prior assumption" is introduced for methodological reasons and does not have any choice-theoretic foundation. The interested reader is referred to Morris (1995).

[96]For more details on the AA variant, see Mongin (1996).

[97]Mongin (1995a, p. 337).

[98]The fact that special (and seemingly irrelevant) properties of Ω or $\mathcal{C}$ are needed to prove the SEU representation theorem is well-recognized and was explained above in Section 3.3.

in the aggregated characteristics. By contrast, *ex post* rules are obtained by first constructing a probability measure and a utility function on consequences for the collective entity, and second combining these two items in the way prescribed by SEU theory. Each type of rules give rise to specific variants of the Pareto principle. *Ex ante* Pareto conditions relate individual to social preferences or utility representations before uncertainty is resolved. *Ex post* Pareto conditions relate the former to the latter after uncertainty is resolved – i.e., only in terms of consequences. Given the usual identification of consequences with a particular subset of acts, *ex ante* conditions are logically stronger than their *ex post* counterparts. It is also true that each family of rules relies on a distinctive application of SEU theory. The *ex ante* point of view applies the theory to individual preferences, while the *ex post* point of view applies it to collective preference as well.

In sum, either type has good credentials in terms of either Paretianism or Bayesianism, but can at best be described as a partial application of these two doctrines. The question then arises of whether or not an *ex ante* rule can also be *ex post*. Since Hammond's work (1981, 1983), the technical answers to this question by welfare economists and social choice theorists have regularly been negative. Propositions 5.4 and 5.5 might be interpreted as stating a further version of the clash between *ex ante* and *ex post* rules. To see that, notice that the assumptions made in the above propositions are that: (i) the individuals are Bayesian; (ii) the Pareto principle holds in an *ex ante* sense; and (iii) the collective preference itself is Bayesian. We have just explained that (i) and (ii) together define the *ex ante* approach, whereas (i), (iii) and and appropriate weakening of (ii) together define the *ex post* approach. Hence the assumptions of Propositions 5.4 and 5.5 can be understood as formalizing the requirement that an *ex ante* Paretian rule be also *ex post*. The conclusions spell out the precise sense in which this requirement fails.

From the Bayesian point of view, only (i) is completely indispensable, and indeed most of the positive contributions to the *ex post versus ex ante* debate can be classified according as they sacrifice (ii) or (iii). The *ex post* school of welfare economics, as it has come to be called, expresses theoretical preference for (iii) over (ii).[99] Its basic objection against the *ex ante*

[99]A leading exponent is Hammond: he recommends "that a policy-maker maximize expected *ex post* welfare

version of the Pareto principle is that it applies the principle to the wrong set of preferences. The underlying argument can be reconstructed as follows. There are various defences of the Pareto principle, based on different interpretations of preference and utility, but all of them involve a basic distinction between factual and normative considerations. One way or another, the individuals are proclaimed to be sovereign about *normative* matters. This appears to mean two things: for one, "collective" normative judgements are derivative; for another, the individuals cannot be mistaken in normative judgements of their own. "Normative" here might be variously understood by reference to values, objectives, or even simply tastes, as in the famous "consumer sovereignty" doctrine. No defence of the Pareto principle has ever involved similar claims in terms of *factual* judgements. This hardly comes as a surprise: it would not make such sense to claim that collective factual judgements are derivative; it is of course grossly untrue that the individuals cannot be mistaken in their factual judgments. *Ex post* oriented writers are aware of the distinction between normative and factual judgements, and of the important implication that unanimity is compelling at most in terms of the former, never in terms of the latter. This is why these writers accept (and indeed recommend) applying the Pareto principle of preferences over *consequences*, but reject it when it comes to preferences over *acts*. Essentially, they see preferences over acts as relying on a critical mixture of factual and normative judgements which blocks the application of the Pareto principle. At the same time, they construe preferences over consequences as reflecting only normative judgements, which, in this case, leads to an unproblematic application of the principle.

The negative case against the *ex ante* version of the Pareto principle seems compelling, but let us reconsider the positive case for the *ex post* version. The underlying argument seems to be this: if the preference concept is suitably restricted, the Pareto principle can reap the benefits of the normative versus factual distinction. But is it possible to restrict the scope of individual preferences to the point where factual considerations do not matter anymore, or at least influence every individual preference identically? A moment's thought shows that

based on the best information available to him" (1983, p. 176); see also the conceptual comments in his 1982 article.

this must be very difficult. In real life, judgements about consequences of acts are infected with factual considerations, which typically diverge from one individual to another. It is only because they must stop the analysis somewhere that decision theorists, such as Savage, have introduced unstructured consequence sets. The formalism of the theory should not mislead its users. In any concrete application, the properties of the consequence set will become relevant, and the claim that the Pareto principle applies to this or that particular consequence set might well founder on the very same argument that earlier refuted the *ex ante* application. On reflection, the *ex post* version is not immune to the difficulties surrounding the *ex ante* version.[100]

Those who find the *ex post* versus *ex ante* debate inconclusive will be inclined to go back to the list of assumptions above and relax (iii) rather than (ii).[101] But a serious obstacle to this way of escape is Harsanyi's already mentioned coherence principle, to the effect that rationality principles are recognized once and for all, and apply to both the individuals' and the observer's preferences (see Section 5.3). Harsanyi expressed this principle as follows: "welfare economists are no more at liberty to reject the sure-thing principle or the other Bayesian axioms of rationality than are people following lesser professions" (1975, p. 67). If anything, "welfare economists" should aim at *higher* standards of rationality than ordinary people. This wording takes for granted that the bearer of social preference is an individual – an assumption which can be disputed. It might well be an appropriate assumption to make in the context of Harsanyi's own ethical theory. He endows each individual with *two* utility functions, or preference maps, one of which (the "personal" one) expresses the individual's tastes and interests, the other (the "moral"

[100]This paragraph echoes Broome (1990 and 1991a, Chapter 7). See also Hausman and McPherson's comment: "why should one believe that people in general are better at forecasting the consequences of lung cancer than the likelihood of getting it?" (1994, p. 398). For another discussion of the *ex post* Pareto principle, see Hild, Jeffrey and Risse (1997). These authors argue that the *ex post* principle is at odds with a basic requirement of adequacy for collective preference (i.e., that collective preference should remain consistent with earlier descriptions of preferences when these descriptions are progressively refined).

[101]In the context of risky choice, some writers [e.g., Myerson (1981); Kolm (1997)] have recommended the use of *concave* social welfare functions, and therefore implicitly rejected the analogue of assumption (iii), i.e., VNM independence. In the same context, Epstein and Segal (1992) have explored the implications of adopting a quasi-concave quadratic social welfare function.

or "social" one) his views about the interests of society as a whole.[102] There are no further utility or preference concepts to be considered in Harsanyi's system. Hence, the only interpretation left for "the social observer", "the bearer of collective preference", and like expressions, is: *any* individual in the society, whenever he adopts a "moral" rather than a "personal" point of view.

Granting this analysis, the coherence principle under discussion strikes one as relatively easy to defend. But there are alternative available notions of "the social observer". Following the tradition initiated by Arrow (1951), social choice theory describes the passive result of aggregating individual normative judgements. The properties of collective preference relations are entirely endogenous – they may or may not turn out to be well behave. As far as collective preferences are concerned, some properties might be *desirable*, but not *normatively compelling* in the way they would be if applied to individual preferences. Accordingly, one should not insist on endowing the collective entity with these properties if this contradicts the more basic requisits of the social choice problem. Transitivity is the most famous example of a requirement on collective preference that Arrovian theorists regard as desirable, though not indispensable.[103] A similar argument could be made for the sure-thing principle in those cases in which the bearer of social evaluation is not a concrete individual, but an abstract entity or a pure aggregate.

For the sake of completeness, we should mention that there are alternative ways of approaching the difficulties of collective Bayesianism. One of them consists in changing the meaning of "Bayesianism" in assumptions (i) and (ii) and more particularly to relax the property of *state-independence* which we explained in Section 3.3 is crucial to both the Savage and Anscombe-Aumann version of SEU theory. This solution is explored in Schervish et al. (1991) as well as in Mongin (1996). It can be shown that the negative results of Propositions 5.4 and 5.5 disappear from the pure state-dependent version of SEU theory but reemerge (though in a more complex form) when the theory allows for a mixture of state-dependence and state-independence. Thus, given relevant qualifications, state-dependent utility theory does *not* appear to be the

[102] See, e.g., Harsanyi (1975, pp. 65–66) and similar passages in his (1977a and b).

[103] On the status of transitivity in social choice theory, see in particular Arrow (1951, pp. 118–120), Buchanan (1954) and Sen (1970, 1982a).

way out of the predicament of collective Bayesianism. The impossibility of "consistent Bayesian aggregation" is a robust theorem.[104]

6 The impartial observer, the original position, and fairness

6.1 Impartial-observer theories

Following a long-standing tradition, *impartiality* is a distinctive feature of moral judgements on collective life – notably, when it comes to deciding whether or not a social situation or an institution is just. That is to say, these judgements should not depend on the individuals' identities and other particular circumstances. They should remain the same if the individuals and surrounding circumstances concerned are, *mutatis mutandis*, replaced by others. There are two distinctive currents of thought in which this general principle has been defended. One is typical of the Scottish writer of the 18th century, in particular Hume, Hutcheson and Smith, while the other is associated with Kant's practical philosophy.

The former current leads to the celebrated construction of the "sympathetic but impartial observer" in SMith's *Theory of Moral Sentiments* (1759). In Hutcheson's earlier formulation, "benevolence" – rather than "sympathy" or "impartiality" – was the key notion:

> "This universal Benevolence toward all men, we may compare to that Principle of Gravitation, which perhaps extends to all Bodys in the Universe: but, like the Love of Benevolence, increases as the Distance is diminished and is strongest when Bodys come to touch each other." (1725, paragraph 145)

This definition suggests that one and the same feeling is at work whether the individual is concerned with himself or the others – the difference between the two cases being only a matter of degree. "Universal benevolence" was Hutcheson's suggested foundation for morality.

[104]Levi (1990) provides another discussion which does not fit in with the previous pattern of choice between the *ex post* point of view and rejection of Harsanyi's coherence. Levi rejects the *ex ante* principle but does not restrict it as drastically as does the *ex post* school; he is concerned about justifiable unanimity in a truly *ex ante* context. The technical implications of Levi's justifiability criterion need to be sorted out.

It can be argued that it is a shaky foundation. Benevolence is just a *feeling*. Although it might be compatible with the exercise of judgement, it by no means implies that a judgement, let alone a moral judgement, is passed. Besides, even if it is "universal" in Hutcheson's sense of embracing all mankind and varying by degrees, benevolence is only a coarse approximation to the impartiality requirement of traditional morality. That I am moved by others' situations, and willing to act on their behalf, is no evidence that I am a moral person. Benevolent people, if they are just benevolent, lack the disinterestedness which is typical of morality. Hume's notion of "sympathy" has much in common with Hutcheson's "benevolence": it is an all-embracing feeling, the intensity of which varies monotonically with propinquity. However, more clearly than "benevolence", "sympathy" involves the exercise of judgement:

> "My sympathy with another may give me the sentiment of pain and *disapprobation*, when any object is presented, that has a tendency to give him uneasiness tho' I may not be willing to sacrifice anything of my own interest, nor cross any of may passions, for his satisfaction." (1736, p. 586, our italics)

Even if sympathy in Hume's sense is an admixture of feeling and judgement, one can argue that it is not yet an appropriate foundation for morality, because, like benevolence, it lacks disinterestedness. To act out of sympathy for my neighbour is not, in a deep sense, to act morally. One can interpret smith as clarifying this when he recommends that the moral observer should balance *sympathy* and *impartiality* with each other. The crucial point is that sympathy must be corrected by impartiality; in particular, we should avoid biasing our judgements towards those whose life we share, or those who resemble us. Conversely – but this seems to be a secondary theme – impartiality is stimulated, and in some sense preceded, by sympathy. Smith apparently believed that judgements of morality would simply be impossible – i.e., they would not really exist as acts of the mind – once separated from the psychological substratum of sympathy. By and large, this *sentimentalist* stand towards morality was shared by all of the 18th century Scottish writers.[105]

[105]The quotations of this paragraph are borrowed from Collard (1978). For other relevant extracts of the British moralists, see Monroe (1972).

Another, altogether different connection between morality and impartiality can be found in Kantian philosophy. On this view, the notion that the moral law is impartial must be explicated in terms of its conformity to reason. Kant is well known for emphasizing universality as a criterion of conformity to reason. The way in which he does is highly restrictive. He distinguishes between the conditional universality of *hypothetical* imperatives (which tell you what to do if you have certain ends) and the unconditional universality of the *categorical* imperative (see Section 2.1 above). Only the latter has the formal character that Kant claims to be the distinctive feature of reason. Hence, only the latter is said to be relevant to morality. Prudential reasoning, which relies on the former kind of imperatives, is declared to be void of ethical content. Many later philosophers have retained from Kant's *Groundwork* and *Critique of Practical Reason* the important claim that the status of a maxim of action as a moral law can be testes by examining whether and how it can be universalized. But a majority of these philosophers have also rejected the extreme formalistic stance of Kantian ethics. They have denied that the distinction between two kinds of imperatives can consistently be made, while still insisting that universalization provides a usable criterion for moral reasoning. This updated – some would say watered-down – style of Kantianism underlies Harsanyi's reconstruction of "Ethics in Terms of Hypothetical Imperatives" (1958). Rawls's (1980) reconstruction of Kantianism also tampers with Kant's initial distinction between two kinds of imperatives.

Contemporary impartial observer theories appear to borrow from both the Scottish and the Kantian sources. One way or another, these theories elaborate on Smith's basic idea that one should judge one's actions as if they were observed by some hypothetical other. But they appear also to rely on Kant's universalization device, if not on his sharp division between the intelligible and the empirical realms, or between the two kinds of imperatives. The connecting link is as follows: the action that is declared to be best from the impartial observer's point of view can also be said to have successfully passed the universalization test.

As a first application, consider Hare's (1976, 1981) "universal prescriptivism". It relies on the following idea of interpersonal permutations: the same moral prescriptions should apply to all situations obtained from a given one by permuting individuals. The observer should first evaluate

any social situation by adopting the identity of every individual in turn, and then weighting the resulting evaluations equally, in order to derive his final evaluation. Although Hare's theory does not provide any formal derivation, it is meant to be a defence of utilitarianism. Hence, it must be a teleological theory. However it also has a deontological side, since it is based on everybody's right to receive an impartial treatment. Universal prescriptivism "is nothing but a restatement of the requirement that moral principles be universalizable" [Hare (1981, p. 154)]. A weakness of Hare's theory is that it does not specify the defining properties of individual identities, and thus leaves relatively undefined the nature of the many evaluations that should be weighted against each other.

Other writers have brought the Impartial Observer approach to a higher degree of precision by calling upon the methods of utility theory. One prominent example, to be reviewed in Section 6.3, is Harsanyi in his 1953 Impartial Observer Theorem.

6.2 State-of-nature theories and the "original position"

Impartial observer theories have often been contrasted with state-of-nature theories. The latter claim that the basic principles of society derive from its members' *voluntary agreement*: the concept of a "state of nature" then refers to that particular situation which precedes and brings about the agreement. This important tradition of analysis is more closely linked to political than to moral philosophy. The writers who initiated it in the 17th and 18th centuries – Hobbes (1651), Locke (1690), and Rousseau (1755, 1761), to mention only the most famous ones – were primarily concerned with justifying the existence and defining the limits of political institutions. However, they also wanted to illuminate the contrast between the primitive and fully socialized stages of human life, so that their analysis can be read from other angles than just political theory. They can also be viewed as providing an ideal genesis of morality, a point well emphasized by Gauthier (1986, p. 10). The common feature to all of the 17th and 18th century notions of the state of nature is that they put special constraints on individual interactions before the founding agreement takes place. Depending on the particular author, these constraints can go in opposite directions: Rousseau's state of nature keeps individual interactions to a bare minimum,

but Hobbes's and Locke's constructions admit of wide-ranging interactions (which led Rousseau to complain that they had failed to capture the true meaning of a state of *nature*).[106] The classics were unanimous in considering the individuals' agreement on political institutions as a full-fledged *contract*; they modelled it after the corresponding notion in positive law. Hence the label "contractarianism" for these writers' and their followers' theories. Another important feature that is common to the classics is that they accounted for the transition from the state of nature to the political state in quasi-historical terms. Admittedly, their works are essays in *conjectural* rather than in real history. But the fact remains that their argument for justifying and assessing institutions cannot be stated without the device of a temporal set-up, however interpreted.

Those contemporary philosophers, such as Rawls (1971), or Gauthier (1986), who endorse the 17th and 18th contractarian philosophers' tradition, have distanced themselves from them in a number of respects. To the best of our knowledge, no contemporary writer has ever really endorsed the notions of a state of nature and of the ensuing contract – even in the conjectural history interpretation. Two definitely weaker concepts are used instead: respectively, that of a *reference situation* from which relevant institutions and aspects of collective life have counterfactually been eliminated, and that of a (possibly tacit and informal) *agreement* between the individuals.

Gauthier's (1986) construction of the reference situation and the ensuing agreement is based on the formal theory of bargaining – more precisely, on Kalai and Smorodinsky's (1975) solution to the bargaining problem. This and other constructions are representative of a whole class of state-of-nature theories which are formulated in terms of game-theoretic concepts (be they cooperative or non-cooperative), and thus go beyond the purview of this chapter.[107] By contrast,

[106] In Hobbes' state-of-nature individuals have unlimited claims to self-preservation and property, which leads to universal war. By contrast, in Locke's state-of-nature individual freedom and property rights are grounded in, and limited by Natural Law, and peace and mutual assistance prevail. For a thorough account of 17th and 18th centuries state-of-nature theories, in particular Rousseau's, the reader is referred to Dérathé (1951).

[107] On Gauthier's contractarianism, see Vallentyne's (1991) collection, and for reviews of bargaining theory, see Kalai (1985), Gaertner and Klemisch-Ahlert (1992), and Thomson (1995). (The latter extends the basic definitions

Rawls's reference situation – the "original position" – is not phrased in game-theoretic terms. Rawls requires unanimity to hold for the building up of collective institutions. The citizens' unanimous decision could well result from some preexisting negotiation, but Rawls construes it differently: individual choices are made separately, under special conditions of ignorance, which explain why these separate choices happen to coincide. This famous restatement of the "original position" in terms of a "veil of ignorance" links Rawls's construction to the theory of individual choice under risk and uncertainty; hence it fully belongs to our subject matter.[108]

6.3 Harsanyi's impartial observer theorem, and the problem of "extended sympathy"

Harsanyi's 1953 theorem, as further clarified in (1977a and b), is perhaps the simplest recent example in the class of impartial observer theories. Harsanyi adheres to Smith's notion that "the moral point of view is essentially the point of view of a *sympathetic* but *impartial* observer" (1977a, p. 49), while occasionally claiming for his theory the benefit of Kant's universalization maxim. He appears to recognize the difference between the hypothetical experiments involved in his quasi-Smithian construction, and the hypothetical histories that are typical of the state-of-nature approach. His novel contribution to the Smithian-Kantian tradition is twofold: he argues first that the observer's judgements can be reproduced as an individual's choices in a

to the variable population case.) Gauthier is not the only contemporary writer to base an ethical construction on bargaining theory. Binmore (1994) also does that, while expressing theoretical preference for Nash's (1950) original solution to the bargaining problem. Yaari (1981) also investigates Nash's solution viewed as a theory of justice. In an innovative empirical study, Yaari and Bar-Hillel (1984) compare respondents' attitudes towards several well-known bargaining solutions. Roemer (1994, Essay 9) questions the relevance of the bargaining approach as a whole to the theory of justice. His critique is largely directed against the informational limitation (i.e., the welfarism assumption) underlying this approach.

[108] Nozick's (1974) libertarian conception also involves a primitive reference situation and a founding agreement. The former is characterized by an "original acquisition of holdings". The latter results only in a "minimal" state. The current property distribution is justified only if it results from successive free transfers starting from the reference situation, but there is a – unanimously agreed – "Lockean clause" whereby this process should not worsen anyone's subsistence level.

relevant situation of ignorance (i.e., "complete ignorance of what his own position, and the position of those near to his heart, would be within the system chosen", 1953, p. 4), and second that this makes utility theory the relevant tool of analysis to resort to ("choice in that hypothetical case would be a clear instance of a 'choice involving risk"', ibid.). In Harsanyi's argument, the connecting link between complete ignorance and risky choice is *equiprobability*. If the individual does not know what his own "position" is, he should give equal probabilities to the various "positions" he can conceivably hold. This is a significant and disputable step in Harsanyi's argument. Bayesian statisticians often take for granted that complete ignorance should be rendered by a uniform prior, but SEU axiomatizations of Bayesianism, such as Savage's and Anscombe and Aumann's, do not logically imply this principle. Also, Harsanyi claims to provide an ethical system rather than just a theory of justice. His notion of ethics is restrictive in one sense, because he unexceptionally identifies the "moral" and "social" points of view, and thus appears to leave no room for private ethics. In another sense, his conception is an encompassing one, because it supposedly covers *all* ethical aspects of social relations.

Even assuming that the theory of risky choices and the equiprobability model are relevant here, one gets different formalizations, and possibly different ethical implications, depending on how one draws the line between what is known and what is unknown to the individuals in the hypothetical experiment. Lerner (1944) and Vickrey (1945) are sometimes given credit for anticipating Harsanyi, but they were imprecise in this respect as well as in others. Vickrey seemed t have in mind that the members of society should ignore their position on the *income distribution* ladder.[109] Harsanayi's 1953 article is still unclear. It can be understood in terms of Vickrey's interpretation, as well as of a richer (and philosophically more challenging) notion of what in the individual's position should be unknown to him. From Harsanyi's later comments, the wider interpretation emerges as the only relevant one:

"Individual i's choice among alternative social situations would certainly satisfy (the)

[109] According to Vickrey, society should chooses as would any individual, on the understanding that "once he selects a given economy *with a given distribution of income*, he has an equal chance of landing in the shoes of each member of it" (1945, p. 329; our emphasis). See also Vickrey (1960).

> requirement of impartiality and impersonality, if he simply did not know in advance what his own social position would be in each situation – so that he would not know *whether he himself would be a rich man or a poor man, a motorist or a pedestrian, a teacher or a student, a member of one social group or a member of another social group, and so forth"* (1977a, pp. 49–50, our emphasis).

This apparently indefinite list points towards the following interpretation: the information to be cancelled relates to everything that constitutes i's individuality – a point well clarified in Pattanaik's (1968) comparison of Harsanyi with Vickrey.

To formalize Harsanyi's point, one might want to borrow from the following, independently developed construction of social choice theory. In various social choice contexts, Arrow (1951 and 1963, 1977), Suppes (1966), Sen (1970, chapter 9), Kolm (1972, 1995), Suzumura (1983, 1994) and others have discussed versions of "extended sympathy" – Arrow's term. Essentially, these authors assume that each member of the society i is endowed not only with an actual preference relation defined on some alternative set X, but also with an "extended preference" relation defined on suitably modified alternatives, having the typical form (x, j). In these "extended alternatives" both the individuals' identities and the initial alternatives are treated as choice or evaluation variables. That i can make extended preference judgements is not a weak assumption. It means that given any two members of the society, j, k, individual i can decide which is better from these two: either to be faced with x while being j, or to be faced with x while being k.[110] Sen (1970, pp. 149–150) has provided a two-individual illustration in which j is a devout Muslim and k is a devout Hindu, and x means "to eat pork" and y "to eat beef". Any member of the society should be able, in effect, to decide whether breaking Hindu law is better or worse than breaking Muslim law. This famous example was not primarily meant to illustrate the demanding nature of extended preference judgements, but it does.

Like Harsanyi himself in his later work, we shall borrow the special preference concept just outlined, and formalize his Impartial Observer approach accordingly. However, our extended

[110] Of course, $i = j$ and $i = k$ are logical possibilities in this statement. "Alternative", here might refer to individual actions, as in Sen's example, as well as to social states of affairs.

preference relation will be defined on elements (x, t_j), where t_j denotes individual j's *type*. This notation is meant to convey the particular interpretation of extended preference that is suited to Harsanyi's approach: i will be assumed to "land in the shoes" of j by deducing what his choices or evaluations are, given j's relevant characteristics. To say that i records j's choices or evaluations directly would be to make an altogether different assumption.[111] Harsanyi (1977a, pp. 58–59) makes it clear that he is concerned with the former, not the latter. Accordingly, the second variable of our extended utility functions will not be an index of individuals but a symbol of their relevant characteristics. As in Harsanyi's (1967–68) classic exposition of incomplete information games, the notion of a type t_j will refer to the (supposedly meaningful) complete list of such characteristics.[112]

Formally, let us denote the set of all individuals' types by $T_N = \{t_1, \cdots, t_n\}$ and the initial alternative set by X. Then, $X \times T_N$ is the set of extended alternatives. We also need to introduce the set $\Delta_s(X \times T_N)$ of extended lotteries: they will be the objects of extended preference in our formalization. Of special significance are the following equiprobable extended lotteries: for any $x \in X$,

$$L_x = (1/n(x, t_1), \cdots, 1/n(x, t_n)).$$

In words, L_x promises alternative x with equal probabilities of being awarded any of the available types. Each individual $i \in N$ is endowed with three utility functions: a personal utility u_i on X, a moral utility w_i also defined on X, and an extended utility v_i, which represents his extended preferences on $\Delta_s(X \times T_N)$. We make the technical assumption that X itself is a lottery set. This will make it possible to apply the VNM axioms to u_i and v_i.

Our first axiom, to be called *Equal Chance*, is Harsanyi's account of the impartial but sympathetic observer:

[111] A related distinction is discussed in Suzumura (1983, pp. 133-136).

[112] d'Aspremont and Gérard-Varet (1991) elaborates on the connection between Harsanyi's Impartial Observer construction and his concept of a type in games of incomplete information. They restate the former in terms of Bayesian implementation theory.

Axiom EC

$$\forall\, x, y \in X, \forall\, i \in N, w_i(x) \geq w_i(y) \text{ iff } v_i(L_x) \geq v_i(L_y).$$

It says in effect that to compare x and y morally is to be ignorant of one's own identity, in the special sense of giving equal probability to each available type. This axiom should be compared with the symmetry requirements of social choice theory, such as **(A)**. While the Anonymity axiom used in Section ' can receive only an ethical interpretation, the present one has both an ethical and an epistemic connotation.

(EC) connects the moral preferences with the extended ones. The second axiom, or *Principle of Acceptance*, will connect extended with personal preferences:

Axiom PA

$$\forall\, x, y \in X, \forall\ i, j \in N, v_i(x, t_j) \geq v_i(y, t_j) \text{ iff } u_j(x) \geq u_k(y).$$

On the face of it, this seems another axiom about the impartial but sympathetic observer. In the extended sympathy literature (e.g., Suzumura, 1983), a similar condition has been defended as embodying *nonpaternalism*: when i compares two options by mentally occupying j's position, he should react exactly the same conclusions as does j himself. This interpretation of **(PA)** is suitable for other philosophical contexts, but would miss the essential point here. Axiom **(PA)** expresses Harsanyi's conviction that personal utilities u_j can be reconstructed from knowledge of j's type t_j. In (1977a) he assumes in effect that there are sufficiently precise psychological laws, as well as sufficiently widespread knowledge of these laws, to make the deduction of the u_j function unproblematic for any observer i. This (very strong) assumption is metaphysical rather than ethical in character. It might well involve similar practical consequences to, but does not follow from, the purely ethical attitude of nonpaternalism.

Our third axiom, to be called *Fundamental Preference*, says that extended preference comparisons are the same from one individual to the other. Conceptually, we are only interested in the individual's preferences over extended *alternatives*. But for reasons that the proof below will make clear, we state Fundamental Preference as the requirement that preference comparisons between extended *lotteries* be uniform across individuals.

Axiom FP

$$\forall\, p, p' \in \Delta_s(X \times T_N), \forall\, i, j \in N, v_i(p) \geq v_i(p') \text{ iff } v_j(p) \geq v_j(p').$$

That some uniformity assumption is needed to derive Harsanyi's conclusion of a *unique* moral utility function has been recognized by his commentators.[113] There is no doubt about the mathematical point. Whether **(FP)**, or similar strong uniformity assumptions, can be justified conceptually is another matter. When the objects of extended preference are pairs of initial alternatives and *individualities* (rather than types), there is no reason why uniformity of judgement should prevail. Remember La Fontaine's fable: the poor shoemaker fancies himself happier in the role of the successful moneymaker, whereas the latter holds the exactly opposite view. The whole point of introducing the type notion, and Harsanyi's strong assumption about objectively known laws, is precisely to make uniformity of judgements plausible. But even in this interpretation, the argument for **(FP)** remains open to doubts.[114]

Proposition 6.1 (Harsanyi's Impartial Observer Theorem) *Assume that the personal utilities $u_1, \cdots, u_n$ and moral utilities $w_1, \cdots, w_n$, are VNM functions on X, and that the extended utilities $v_1, \cdots, v_n$ are VNM functions on $\Delta_s(X \times T_N)$. If* **(EC)**, **(PA)** *and* **(FP)** *hold, there is a common function w such that each w_j is a positive affine transformation of w, and:*

$$\forall\, x \in X,\ w(x) = 1/n \sum_{i=1}^{n} u'_i(x),$$

where $u'_1, \cdots, u'_n$ are positive affine transformations of $u_1, \cdots, u_n$, respectively.

Notice carefully that without **(FP)**, the conclusion would simply be this: each individual moral function w_j is some positive affine transformation (PAT) of $1/n \sum_{i=1}^{n} u'_{ij}(x)$, where u'_{ij}

[113] See Pattanaik (1968), Sen (1970), Kaneko (1984), MacKay (1986), and Suzumura (1994).

[114] The intricacies of the argument are illustrated by Broome's (1993) discussion of Harsanyi's and Kom's claims that extended preference judgements must be made in the same way by different individuals. Broome rejects this claim even in Harsanyi's version. His main point is that extended preference constructions confuse with each other the two notions of an *object* of preference and of a *cause* of preference. Kolm (1994) denies that this constitutes a relevant objection.

is a PAT of u_i which depends *on the particular* j. Note also that if **(FP)** were restricted to preference comparisons between extended alternatives, the argument would not carry through. Extended lotteries are needed if one is to apply the VNM machinery. To make the argument straightforward, we have assumed that extended preferences are defined on the set of *all* extended lotteries, but this assumption could be weakened.[115]

Contrary to the Aggregation Theorem, which leads to nontrivial variants, the Impartial Observer Theorem has little technical interest. But it provides an important argument for *mean rule* utilitarianism,[116] as well as a framework in which a number of conceptual issues can be addressed. The first objection raised against the other theorem, to the effect that it does not state the relevant axiom of interpersonal comparisons, would be equally inappropriate here. The above axiomatic decomposition makes it clear that Harsanyi wants to assume the *general* possibility of interpersonal utility comparisons: this much is implies by the conjunction of **(PA)** with the technical assumption of a well-defined extended preference. Exactly like the Aggregation Theorem, the Impartial Observer Theorem derives the *specific* utilitarian form of utility comparisons from the assumption that they are possible in general, and various other assumptions. The second objection against the other theorem, to the effect that Harsanyi should not have exclusively relied on VNM utility representations of moral and personal preferences, applies here with equal force. We refer the reader to the relevant discussion of the Sen-Weymark critique in Section 5.3. This discussion can be reproduced word for word, except for the words "individual" and "social", which should now be replaced by "personal" and "moral", respectively.

Some further philosophical problems are specific to the Impartial Observer Theorem. Most writers in the extended preference literature have followed Arrow (1977) in defining this concept ordinally. Accordingly, some of these writers, like Kolm (1972), have taken Harsanyi to task for *cardinalizing* extended preferences by defining them on a domain of VNM lotteries. Their criticism appears to be based on the view that preference (in whatever context) is an exclusively

[115] Karni and Weymark (1996) derive the Impartial Observer Theorem for a more parsimonious domain assumption.

[116] Although, of course, this rule is equivalent to sum utilitarianism when the population size is fixed.

ordinal concept. We have argued in Section 2.3 against the latter view: it is too restrictive, because "ordinalism" does not analytically follow from the definition of preference. Hence, we do not see cardinality of extended preferences as constituting a problem *per se*. That Harsanyi's specific procedure for cardinalizing extended (as well as other kinds of) preferences by a VNM device may not be appropriate is a different issue, with which part of the Sen-Weymark critique was precisely concerned.

When reviewing early variants of the Impartial Observer theorem, we suggested that there is room for disagreements in the analysis of the individual's position, and of what in his position should be assumed to be unknown. Indeed, part of Rawls's (1971) critique of Harsanyi reflects a disagreement of this sort. Another part has to do with the modelling of ignorance and the way of coping with it. In particular, **(EC)** is questionable. Rather than maximizing expected utility with respect to an equiprobable lottery, Rawls contends, agents should apply the alternative model of *maximin*, which implies extreme caution on their part. That component of the Rawls-Harsanyi debate has already surfaced in the framework of social welfare functionals (see Section 4.4). The *maximin* approach does not make use of subjective probabilities, which has led to severe criticisms on the Bayesian writers' part. We said earlier that they do not have to endorse **(EC)**, but this line of thinking has attracted little attention.[117]

Further discussions of the Impartial Observer Theorem relate in effect to axioms **(FP)** and **(PA)**, and mostly to the former. We mentioned the misgivings caused by Harsanyi's claim that extended preference judgements are the same from one individual to another. Pattanaik (1968) suggests applying Harsanyi's Aggregation Theorem in order to amalgamate VNM extended utility functions that do not represent the same fundamental preferences. Using a classical framework of social choice theory, Suzumura (1994) shows that aggregation of extended preference judgments, as formalized by binary relations, lead to difficulties of the Arrow type. These

[117]It is possible to devise subjective probability variants of the Impartial Observer Theorem. One simple method is to exploit the formal analogies between Harsanyi's construction of the original position and Anscombe and Aumann's construction of SEU in the single individual case (see Section 3.3). In the resulting model, lotteries will be replaced by uncertain prospects, while individuals will be formally analyzed in the same way as are states of the world in the AA context.

two writers are in effect exploring the consequences of dispensing with axiom **(FP)**.

6.4 Rawls's "Original Position" and "Veil of Ignorance"

As is well known, Rawls's device of putting the individual behind the "veil of ignorance" is meant to explain why it is rational to accept his "two principles of justice". They are:

1. Each person has an equal right to the most extensive basic liberty compatible with the same liberty for others.

2. Social and economic inequalities must be (a) to the greatest benefit of the least advantaged members of the society, and (b) attached to positions open to all.

Rawls has repeatedly claimed that the first principle is to have priority over the second, and part (b) of the second principle over part (a) (which he labels the difference principle). For some time, economists knew Rawls primarily through the formal reconstruction of Section 4, and therefore tended to ignore the broader perspective of his philosophy, which is primarily concerned with basic liberties and equal chances of promoting the individual's own conception of good. That is to say, one should be careful not to emphasize the "economic" part of the doctrine, even when restricting attention to the second principle. Apart from this overemphasis, the early reading involved the already mentioned procedure of discussing Rawls in terms of standard utility functions without properly justifying this restatement.

All this has come to be recognized, and a perhaps more genuine interchange between Rawls and the economists has recently taken place. In particular, it addresses Rawls's notion of primary goods and the problems raised by its formalization. Rawls has described primary goods as "all-purpose" means for the person's promoting his own conception of the good. "Primary goods are things which it is supposed a rational man wants whatever else he wants" (1971, p. 92). The way to define them is closely related to the statement of the two justice principles. The principles are best understood as regulating the distribution of primary goods among citizens at the ideal founding stage of the "well-ordered society". More precisely, Rawls (1982, 1988) defines them in terms of the following list:

1. The basic liberties, in particular freedom of thought, freedom of association, and political liberties.

2. Freedom of movement and of choice of occupation.

3. Powers and prerogatives.

4. Income and wealth.

5. "The social bases of self-respect".

The relative importance of these goods is roughly conveyed by their numbering, and mirrors the respective importance of each principle, or subprinciple, of justice. Because of his lexicographic ranking, Rawls is not prepared to consider trade-offs between all the five groups: the fundamental liberties should be equally distributed, and there should be fair equality of opportunities; only the last three groups are susceptible of being balanced against each other. Rawlsl (1971, pp. 93–95) appears to recognize that weights should be defined for at least the non-prioritized goods, if the notion of an "index of primary goods" is to make sense at all, but neither in *A Theory of Justice* nor in his later articles (e.g., 1982, 1988) does he come out with a satisfactory formulation. He has himself come to recognize the full difficulty of the indexing problem, as will be explained shortly.

The two principles of justice – or equivalently, the distribution of primary goods they prescribe – are supposed to result from every person's choice under the "veil of ignorance". Notice carefully that the Rawlsian choice is concerned with principles, not with social states, a point which elementary presentations sometimes overlook. A primary good distribution is not a social state in the ordinary sense, since it does not specify the individuals' circumstances entirely. This feature of Rawls's theory reflects the claim that the members of society's make entirely personal use of their primary goods assignments. It makes a significant difference from Harsanyi's theory, which entails an assessment of principles *only indirectly*, i.e., as a result of initially assessing the states. Harsanyi's impartial observer is utilitarian not because he chooses to be so, but because his choices among states turn out to satisfy the utilitarian criterion. The philosophical

demarcation here can be understood in terms of the initial contrast of this section between impartial-observer and state-of-nature theories. The Rawlsian "original position" is a state of nature, at least in the following sense: it (ideally) precedes any social state. The impartial observer's position is just distinct from, but not conceptually prior to, the social states. As a state-of-nature theory, Rawls's makes it possible to determine only the society's mot general principles, whereas Harsanyi's impartial observer theory supposedly provides a ranking of all social states. Another relevant contrast is that between a mostly deontological and an exclusively teleological analysis: Rawls does not want to make the notion of justice hinge on an evaluation of consequences, whereas Harsanyi wants to do precisely that. All in all, the philosophical differences are so overwhelming that it seems to be less important than has often been suggested to compare Harsanyi and Rawls in terms of the particular choice-theoretic features of their ideal positions.

The Rawlsian veil of ignorance is thick. It obscures not only the individual's place in society (typically, his wealth and income) and his defining characteristics (such as his talents and handicaps), but also his own conception of good. The individual does not know either in what kind of society he will live; he has no idea of the property regime, the income and wealth statistics, etc. On the other hand, he is aware of the basic facts of human nature and has extensive knowledge of a general kind. Under these conditions, Rawls argues, the individuals will tend to form their expectations in terms of an identical primary good index. It will function as "a publicly recognized measure" (1971, p. 95). When revisiting the issue, Rawls (1982) makes it clear that the choice of the index reflects the informational constraints of the original position, and that it has to be common knowledge between the participants. As a further step, he claims that the safe-playing attitude that is inevitable under the veil must lead to choose the distribution of primary goods recommended by the two justice principles. It is here that the famous maximin "analogy" (1971, p. 152) comes in; it is just an analogy, because, again, the relevant unit of evaluation is the index, not the individuals's utility.

Even leaving aside the basic liberties and conditions of fair opportunity, the commensurability problem raised by Rawls's index is an acute one. Rawls (1971, p. 94) initially attempted

to circumscribe the problem by requiring a weighting of primary goods only for the least advantaged, but this is unsatisfactory, as he came to recognize. How could one identify the least favoured group? There seems to be no way of identifying it, except in terms of an index defined for all in the society. Besides, the indexing problem arises not only because one should balance the last three groups of goods against each other, but already because one has to aggregate goods within two of these groups, namely "powers and prerogatives" and "the social bases of self-respect". For lack of a better solution, Rawls (1982) has eventually resigned himself to taking monetary wealth as a proxy for the index.

We shall sketch an alternative argument here. *To assume that the index can be constructed is, formally, to assume that there is a utility function.* This function is essentially unknown, but it will inherit at least some recognizable properties from the preceding philosophical argument. First, it must be unique, since, by assumption, all individuals share the same index. Second, it must be strictly increasing in each of its arguments, which are quantities of (eminently desirable) primary goods. More specific properties than these two do not follow as a matter of logic. But quasi-concavity (diminishing rates of substitution along indifference curves) and other standard microeconomic properties are not logically excluded either.

We can go one step further and inquire whether Rawls's philosophical framework can agree with *technical welfarism*, i.e., the formal notion of welfarism defined in Section 4. There would be little philosophical sense in introducing a multi-profile setting of utility functions. Rather, Rawls's argument points towards a single-profile setting (in the special case in which the n individual utility functions are identical). In order to reap the benefits of the Welfarism Lemma, it is enough to be able to accept the Unrestricted Profile and Relative Neutrality assumptions that make it possible to extend technical welfarism to the single-profile framework (see Section 4.6). The former is a richness-of-domain assumption, which must now be assessed in terms of a primary goods interpretation. The latter (conceptually more crucial) assumption says in essence that the social evaluation of any primary goods allocation will depend only on the distribution of numerical values taken by the index. Both assumptions can be defended in Rawlsian terms. Once the conclusion of the Welfarism Lemma is granted, the discussion can proceed along the

lines of SWFL theory. Comparability holds by assumption, and the general philosophical argument favours Ordinality. Thus, we have gone a long way towards the earlier reconstruction of the economic part of the "difference principle" in Section 4, but there remain two differences: for one, the individuals share the same utility function, and for another, the case for leximin (as opposed to any other rule compatible with Ordinality and Comparability) has not been settled. In view of Proposition 4.3 above, to accept leximin at this stage is tantamount to accepting Anonymity and Separability, while rejecting Leximax.

6.5 Alternative notions of fairness

Choice theorists have often rejected maximin, as well as leximin, on the grounds that being non-probabilistic, these criteria clash with the Bayesian axioms of rational choice. A related point against maximin, which can be found in Arrow (1973a), is that it is just a limiting case of an expectational approach to uncertainty: it amounts to giving an infinite weight to the worst outcome. These criticisms fully apply to the welfarist reconstruction of Rawls's original position, but do not lose their force when redirected against alternative construals of the primary goods index, such as Rawls's use of a monetary proxy. Another strand of criticisms relates to the Ordinality and Comparability assumption. The recent work in normative economics has often revived the assumption of Ordinality and Non-Comparability that Arrow had emphatically endorsed in *Social Choice and Individual Values*. Despite this heavy informational restriction, the recent work shares much in common with Rawls's project of founding a notion of distributive justice, and significantly borrows Rawlsian expressions such as *fairness or equity* to refer to the looked-for foundation. To dismiss Ordinality and Comparability has the effect of shaking the logical basis of maximin comparisons. Thus, at least implicitly, the new social choice rules take the earlier choice-theoretic critique of Rawls into account.

Among these new rules, we shall emphasize the *no-envy criterion* and its cognates. It was mentioned by Tinbergen, but first explored by Foley (1967) and Kolm (1972; see also 1995), who turned it into a general notion of distributive justice. Varian (1974) is very clear in arguing that envy-free efficient allocations (which he initially labelled *fair* allocations) constitute an

alternative to the Rawlsian outcome of the original position. These three writers, and most of their followers, count as a theoretical strength the weak informational demand made by these concepts.[118]

Suppose that there are n individuals and a given vector $\overline{\omega} = (\overline{\omega}_1, \cdots, \overline{\omega}_m)$ of goods is to be distributed between them ($\overline{\omega}_h > 0, h = 1, \cdots, m$). By assumption, individual utility functions $u_i(x_i)$ are ordinal and non-comparable. The set of feasible allocations is:

$$X = \{x \in \mathbb{R}_+^{mn} : \sum_{i=1}^{n} x_{ih} = \overline{\omega}_h, h = 1, 2, \cdots, m\}.$$

A feasible allocation x is said to be *envy-free* if it satisfies the condition that:

$$u_i(x_i) \geq u_i(x_j), \quad \forall\, i, j, i \neq j.$$

In words, every individual is at least as well-off with his own allocation as he would be with any other individual's. Whenever the reverse strict inequality holds for two individuals i, j, we shall say that i *envies* j. The equal endowment allocation, i.e., $\omega = (\omega^1, \cdots, \omega^n)$, with $\omega^i = \overline{\omega}/n$ for $i = 1, \cdots, n$, is clearly envy-free. A feasible allocation will be said to be *efficient* if it is Pareto-optimal in the strong sense; that is to say, if no other feasible allocation satisfies the antecedent clause of the Strict Pareto condition (see Sections 4 and 5). In the special (Rawlsian) case in which the individuals' utility functions are identical, the equal endowment allocation is not only envy-free, but also efficient. More generally, it might lack the efficiency property.

Consider now an exchange economy, with initial endowments $\omega = (\omega^1, \cdots, \omega^n)$, and suppose that it satisfies the following assumptions for the existence of a competitive equilibrium (x, p): preferences are continuous orderings, hence representable by utility functions $u_1, \cdots, u_n$ and are strictly increasing and convex, so that these utility functions are strictly increasing in each argument and quasi-concave (see Section 2.1). From the definition of a competitive equilibrium, x is a feasible allocation, and p is a price vector, such that:

$$x_i \in \text{argmax}_{x_i'}\{u_i(x_i') : px_i' \leq p\omega^i\} \text{ for all } i.$$

[118] For further relevant contributions to the issue of no-envy, see Thomson (1982), Thomson and Varian (1985), and the references discussed in Ansperger's (1994) and Thomson's (1994) extensive surveys. See also Moulin's (1995) discussion and further elaboration.

If we now make the assumption that the initial endowments ω^i are equal, it immediately follows that the competitive equilibrium allocation x is envy-free. (Suppose it is not; then, there are i, j such that $u_i(x_i) < u_i(x_j)$; from the equilibrium property of x_i, it must be the case that x_j exceeds i's income; but since incomes must be equal at the equilibrium, it also exceeds j's income – a contradiction with the equilibrium property of x_j). It can also be proved that the equilibrium allocation x is efficient. This strong version of the "first fundamental theorem of welfare economics" follows here from the assumption that preferences are strictly increasing. Thus, we have just proved the existence of an envy-free efficient allocation:[119]

Proposition 6.2 *Consider an exchange economy with equal initial endowments and such that individual utility functions are strictly increasing in each argument, continuous, and quasi-concave. This economy has a competitive equilibrium which is envy-free and efficient.*

(Under the same assumptions, the leximin allocation with identical utility functions would also be envy-free efficient. This holds because it would ensure equal utility amounts to each individual.)

The simultaneous realization of envy-freeness and efficiency is not a robust microeconomic property. Following an informal argument due to Foley (1967), Pazner and Schmeidler (1974) have shown that the existence conclusion of Proposition 6.2 depended on assuming an exchange economy. An example in their paper illustrates the lack of existence in a production economy. We state another example in the (Rawlsian) particular case of a common utility function:

Example 6.3. There are two individuals who produce one good out of their labour. For $i = 1, 2$, denote i's consumption and labour by x_i and ℓ_i, respectively. Suppose that the two individuals have identical Cobb-Douglas utility functions, i.e.,

$$u(x_i, \ell_i) = x_i^{\alpha} \ell_i^{(1-\alpha)}, 0 < \alpha < 1, \ i = 1, 2,$$

[119] The reasoning of this paragraph echoes Varian (1974, Theorems 2.2 and 2.3). The assumption of strictly increasing preferences was introduced by Varian for convenience reasons.

and that their production function is given by:

$$x_1 + x_2 = (1 - \ell_1) + \varepsilon(1 - \ell_2), \text{ with } 0 < \varepsilon < 1.$$

This equation implies that 2's productivity is lower than 1's. Now, we equalize the marginal rate of substitution with the marginal rate of transformation for each individual. Hence the following two equations that an efficient allocation should satisfy in addition to the production equation:

$$x_1 = \alpha/(1-\alpha)\ell_1, \quad x_2 = \varepsilon\alpha/(1-\alpha)\ell_2.$$

The corresponding utility values are:

$$u(x_1, \ell_1) = (\alpha/(1-\alpha))^{\alpha}\ \ell_1, \quad u(x_2, \ell_2) = \varepsilon^{\alpha}(\alpha/(1-\alpha))^{\alpha}\ell_2.$$

Because $\varepsilon^{\alpha} < 1$, we see that for every efficient allocation:

$$u(x_1, \ell_1) > u(x_2, \ell_2).$$

Hence, utility values can never by equalized (even when $\ell_2 = 1$). Individual 2 is worse off because of his low productivity. There exists no envy-free efficient allocation

Considering how severe the existence problem is, Pazner and Schmeidler (1978) proposed the novel concept of an *efficient egalitarian equivalent allocation* (EEEA). This concept involves introducing a generally fictitious state, which would be perfectly egalitarian. More precisely, an allocation is said to be *egalitarian-equivalent* if there exists a state in which individual consumptions – including leisure – are equal, and each individual enjoys the same utility as in the actual allocation. To define an EEEA, let X denote the set of feasible allocations in the actual economy, as generated by the set $Z \subset \mathbb{R}^m$ of productive production plans z (where positive and negative coordinates denote outputs and inputs, respectiely) and by a vector of aggregate initial resources $\overline{\omega} \in \mathbb{R}^m_+$, i.e.,

$$X = \{x \in \mathbb{R}^{mn}_+ : \sum_i x_i \in \{\overline{\omega} + Z\}\}.$$

An allocation x is said to be EEEA if it is efficient and if, for some consumption bundle $x_0 \in \mathbb{R}^M_+$, not necessarily in $\{\overline{\omega} + Z\}$,

$$u_i(x_i) = u_i(x_0) \text{ for all } i.$$

Pazner and Schmeidler (1978) prove the following existence result:

Proposition 6.4 *Assume that each utility function u_i is continuous and strictly increasing in each argument. Assume also that the set X of feasible allocations is compact and comprehensive (i.e., $x' \leq x \in X \Rightarrow x' \in X$), and has non-empty interior. Then, there exists an efficient egalitarian-equivalent allocation.*[120]

6.6 Equality of Resources and Welfare

The EEEA concept has an interest of its own, but the existence problem of the initial no-envy criterion remains unsolved. This difficulty derives from an even deeper reason than the role of production, to wit, the *non-exchangeability of some individual characteristics*. Handicaps could have played the role of the individuals' productivities in Example 1. In relation to the non-exchangeability problem, Dworkin (1981) has introduced the distinction between "external resources" and "internal resources", and claimed that both should be taken into account in solving the problem of distributive justice. In effect, Rawls considers only external resources. Whether the original position device succeeds in establishing principles of justice crucially depends on how resources are delineated.

In order to equalize resources, both internal and external, across individuals, Dworkin has proposed two famous procedures, each of which is based on a distinctive notion of the original position.[121] The starting point for Dworkin's first procedure is the property underlying Proposition 6.2: in any "well-behaved" exchange economy (i.e., when all resources are tradeable and there are no externalities), any competitive equilibrium implying an equal income distribution passes the "envy test" successfully. At the equilibrium, no one would like to exchange his final bundle of resources with someone else's bundle. One way of extending this conclusion to inter-

[120]For further discussion and a generalization of the EEEA solution, see Moulin (1995, 1996). Moulin also introduces several alternative ordinal criteria, such as the "stand alone test", which prescribes that each individual's utility should not exceed the utility he would obtain if he were the only agent in the economy.

[121]These mechanisms have been extensively investigated, in particular by Roemer (1985, 1994, 1996), Varian (1985) and van Parijs (1990).

nal resources is to make them tradeable *notionally*, that is to say, to give every individual equal rights on all the members of society's external *and* internal resources. As long as individual i's internal resources have the effect of increasing only i's utility, efficient allocations will automatically imply that these resources are not consumed or used as inputs by any other individual. From the "second fundamental theorem of welfare economics", we know that under standard assumptions, any efficient allocation can be decentralized by some competitive equilibrium. In the particular context, this means that there will be implicit prices for both internal and external resources in the society. The reasoning of Proposition 6.2 may now be applied to the fictitious exchange economy just constructed.

To illustrate, let us go back to Example 6.3. The two individuals have different productivities (1 and ε respectively), a case of differing internal resources. Suppose that "rights to consume leisure" are created, one for the leisure of 1, the other for the leisure of 2, and that the two individuals receive equal shares of these rights. Suppose also that these rights are tradeable. A competitive equilibrium will involve three prices, i.e., the price of the produced good (which we normalize to 1) and the prices of both types of leisure. From the equilibrium condition, the latter should be equal to the respective productivities. Does this equilibrium ensure that individual 2 will be compensated for his low productivity? The equilibrium solutions are:

$$x_1 = x_2 = \alpha(1+\varepsilon)/2,\ \ell_1 = (1-\alpha)(1+\varepsilon)/2,\ \ell_2 = (1-\alpha)(1+\varepsilon)/2\varepsilon.$$

Intuitively, individual 2 is now *overcompensated* for his lower productivity. A similar consequence had been noted in the different context of optimal taxation (Mirrlees, 1974, 1982). In Dworkin's opinion, this "slavery of the talented" is highly undesirable, which leads him to explore an alternative way of equalizing resources.

Dworkin's second method is to construct "an hypothetical insurance market, which assumes equal initial assets and equal risk". Individuals will be able to hedge against the lack of sufficient internal resources. Formally, it seems best to understand Dworkin's second mechanism in terms of a three-stage procedure, which involves a novel concept of the original position. At the first stage – the original position proper – the individuals do not know what their own type will be

and insure against the implied risk. At the second stage, any individual i's is associated with a type, say t_k, in the set T_N of available types; this fixes i's endowment of internal resources. By assumption, types are drawn from an equiprobable lottery, i.e., $p(t_k) = 1/n$. At the third stage, the resulting competitive equilibrium is computed. Due to the presence of internal resources, the individual equilibrium external resources $x_1, x_2, \cdots, x_n$ will not be equal, and hence the corresponding utility values $u(x_1, t_1), u(x_2, t_2), \cdots, u(x_n, t_n)$ will not be equal either. It is against the risk of eventually getting a low utility value that individuals insure in the first place. Under the VNM assumptions, this three-stage mechanism leads to an additive rule which is formally similar to Harsanyi's. This analogy is noted in Roemer's (1985 and 1994) analysis of Dworkin's two mechanisms, to which the reader is referred for more detail.

Dworkin's project should be assessed against the background of the ethically relevant distinction between two groups of individual characteristics: those describing "tastes", for which the individuals should bear responsibility and there should be no insurance; and those describing either "talents" or "handicaps", which should be insured against. Such a partitioning is intended to take care of the problems inhering in the use of "subjective preferences" in judgements of justice. Like Rawls (see Section 2.5), Dworkin rejects differences in tastes as being a fair basis for unequal resource allocations. Rawls goes very far by eliminating *all* individual characteristics and introducing the supposedly objective index of primary goods. In some sense, Dworkin adopts an intermediate position between Rawls's and the welfarists', since he is willing to include *some* of the individual characteristics, while excluding the others. By doing so, he becomes open to criticisms from both the Rawlsian and the welfarist points of views. Roemer has complained that he assumes clarity where there is none: "Where does one draw the line on this slippery slope, which separates those traits of a person which should properly be deemed part of his preferences, from those which are part of his resource endowment?" (1985, in 1994, p. 146). When discussing expensive tastes and handicaps in Section 2.5, we suggested that sometimes they should be treated the same, sometimes not. There is no obvious rule to overcome this indeterminacy. This failure of human reason to implement an ethically important distinction suggests that there is perhaps no stopping point between complete admission of individual

characteristics, leading to some form of welfarism, and complete rejection of them, leading to strong antiwelfarism.

7 Concluding comments

This chapter has reviewed a number of constructions in social ethics, which share the common feature of applying the apparatus of utility theory to a collective (though not an interactive) setting. We have classified them into three broad categories. The constructions in the first group belong to that part of social choice theory which takes individual utility functions (rather than preference relations) as primitives, i.e., the theory of social welfare functionals. Interpersonal comparisons of utility are the main topic of this theory, which has provided illuminating axiomatizations of utilitarianism and egalitarianism (in the sense of the Rawlsian leximin). In the second group, which is exemplified by Harsanyi's and his followers' work, choice-theoretic restrictions imposed on the observer's utility functions play – roughly speaking – the role of interpersonal comparison assumptions in the first group. While discussing these choice-theoretic assumptions at length, we have emphasized the formal connection between Harsanyi's expected-utility approach to utilitarianism and earlier results in the theory of social welfare functionals. The contributions in the third group typically derive the observer's rule from constructing an ideal ethically-loaded reference position. Philosophically, the latter refers to either the Impartial Observer's vision of the society, or to a State of Nature which (logically, if not factually) precedes society; this distinction is the bequest of 17th and 18th century writers to contemporary political and moral philosophy. The reference position is typically analyzed as a state of relative ignorance among individuals, and thus becomes amenable to standard utility-theoretic methods. We have lumped with the last group some recent notions of "fairness", such as the no-envy concept, because they can be viewed as alternatives to Rawls's use of leximin to resolve the ignorance prevailing in the State of Nature.

Even granting our initial scope restriction to *social* ethics, the present survey is far from being exhaustive. An important topic for both utility theory and social ethics is the role of

preference externalities, in particular of benevolent and malevolent preferences. It has indeed been extensively discussed within the first group of theories in relation to Sen's Paretian Liberal Paradox. Another topic which has recently attracted the normative economists' interest is population ethics, and more generally the dynamic side of the more familiar social choice rules, i.e. leximin and utilitarianism. We have only touched on these issues. While laying special emphasis on the second and third groups of theories, we said nothing about the following relevant questions. Would it make ethical sense to extend Harsanyi's aggregative approach (or Rawls's and Harsanyi's analyses of the reference position) by assuming non-expected or non-probabilistic utility theories rather than the von Neumann-Morgenstern and subjective expected utility theory? Among the various models of bargaining which have found their way into social ethics, which one is the most appropriate philosophically, and what does the bargaining approach add to the standard aggregative approach?

Despite these – and other – missing topics, we hope that there is enough in this chapter to document the conflicting positions taken among both philosophical and economic circles on "welfarism". After sketching the pros and cons in the welfarism debate in Section 2, we have found that utility theory remains a relevant tool of analysis in social ethics, *provided that it receives an appropriate interpretation.* The first and perhaps most attractive interpretation results from this simple observation: there is a tight connection between the individual's well-being and his considered or improved preferences. The latter notion needs clarifying, and may or may not exhaust the former. This is a topic for serious philosophical discussion, but there is little doubt about the broad fact that the two notions are linked to each other. To the extent that they represent rational, well-informed, and (on some construals) self-interested preferences, utility functions can *also* represent the individual's well-being. Given that the alternative, purely objective accounts of well-being are still at the tentative stage, it would be methodologically objectionable to dispense with this utility-based approach to well-being. This is not to say that non-utility information is irrelevant. It should play a role for two reasons, i.e. (i) well-being is *not* all what the social good, let alone social ethics, is about; (ii) even restricting oneself to a teleological framework in which well-being is the only notion of good, non-utility information,

typically information on handicaps, talents, and acquisition of preferences, might be necessary in order to derive the observer's rule. Our brief analysis of Dworkin's theory suggests that to combine utility with non-utility information in one and the same model is not an easy task; however, the difficulties encountered here might lie with the implementation, rather than the substance, of the programme. Our tentative conclusions do not preclude work along the lines of minimal and moderate antiwelfarism (in the terminology of Section 2.5). Elsewhere, we reached a more determinate result of a negative sort. We claimed that the approach typically followed by welfare economists equivocates between several interpretations of utility that cannot easily be reconciled. The improved preference interpretation *cannot* coincide with the actual preference interpretation which welfare economists have borrowed from positive microeconomics. Put bluntly, if welfarism can be salvaged philosophically, traditional welfare economics cannot.

Second, we have introduced a notion of *technical* welfarism as against the (mostly polemical) philosophical definitions in current use. Technical welfarism is perhaps best understood as that branch of measurement theory which is concerned with comparisons between personal characteristics. Well-being comparisons are one example, but there are also comparisons of other sorts. An easy (perhaps too easy) application of technical welfarism to a philosophically non-welfarist doctrine is our discussion of Rawls's "primary goods". The analysis of "justice as fairness" in terms of a utility-like index is not so widely off the mark as was claimed by some antiwelfarists. We are not returning to the initial state of the discussion of leximin versus utilitarianism, but rather emphasizing this simple contrast: the primary goods index does not represent well-being, but nonetheless shares several properties in common with a utility-function index of well-being.

Some readers will perhaps be dissatisfied with the semantic manoeuvre that consists in preserving the constructions of utility theory while referring them to a special (i.e., rational and well-informed) kind of preferences. These readers have a point, which leads us to our third and last comment. The formal part, if perhaps not the substantial part, of the rationality concept, can be made explicit *within utility theory itself* under the guise of well-known axioms. In other words, the chosen interpretation can influence the formal constructions. Methodologically,

this internalization of semantics appears to be desirable. One application in this chapter was the analysis (in Sections 5 and part of 6) of aggregative issues under strong choice-theoretic restrictions imposed on the individuals' preferences. That impossibility theorems follow from systematizing this approach remains a challenge to our understanding of collective rationality.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'Ecole Américaine. *Econometrica*, 21, 503–546.

Allais, M. and O. Hagen (eds.) (1994). *Cardinalism*. Dordrecht: Kluwer.

Anscombe, F. and R. Aumann (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, 34, 199–205.

Aristotle. The Nichomachean ethics. In Barnes, J. (ed.), *The Complete Works of Aristotle: The Revised Oxford Translation*. Princeton: Princeton University Press, 1984.

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

Arneson, R.J. (1990a). Liberalism, distributive subjectivism, and equal opportunity for welfare. *Philosophy and Public Affairis*, 19, 159–194.

Arneson, R.J. (1990b). Primary goods reconsidered. *Nous*, 24, 429–454.

Arnsperger, C. (1994). Envy-freeness and distributive justice. *Journal of Economic Surveys*, 8, 155–186.

Arrow, K.J. (1951). *Social Choice and Individual Values*. New Haven: Yale University Press, 2nd revised edition, 1963.

Arrow, K.J. (1973a). Some ordinalist-utilitarian notes on Rawls's theory of justice. *Journal of Philosophy*, 70, 245–263. Reprinted in K.J. Arrow (1984a), chapter 8.

Arrow, K.J. (1973b). Rawls's principle of just saving. *Swedish Journal of Economics*, 75, 323–335. Reprinted in K.J. Arrow (1984a), chapter 10.

Arrow, K.J. (1977). Extendes sympathy and the possibility of social choice. *American Economic Review, Papers and Proceedings*, 67: 219–229. Reprinted in K.J. Arrow (1984a), chapter 11.

Arrow, K.J. (1984a). *Collected Works, Volume 1: Social Choice and Justice*. Harvard University Press, Cambridge (Mass).

Arrow, K.J. (1984b). *Collected Works, Volume 3: Individual Choice under Certainty and Uncertainty*. Harvard University Press, Cambridge (Mass).

d'Aspremont, C. (1985). Axioms for social welfare orderings. In L. Hurwicz, . D. Schmeidler and H.F. Sonnenschein (eds.), *Social Goals and Organization*, 19–76. Cambridge University Press, Cambridge.

d'Aspremont, C. (1995). Economie du bien-être et utilitarisme. In L.A. Gérard-Varet and J.C. Passeron (eds.), *Le modèle et l'enquête. Les usages du principe de rationalité en sciences sociales*, 217–241. Editions de l'E.H.E.S.S., Paris.

d'Aspremont, C. and L. Gevers (1977). Equity and the informational basis of collective choice. *Review of Economic Studies*, 44, 199–209.

d'Aspremont, C. and L.-A. Gérard-Varet (1991). Utilitarian fundamentalism and limited information. In J. Elster and J.E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, 371–385. Cambridge University Press, Cambridge.

Barry, B. (1989). *Theories of Justice*. Harvester, London.

Barry, B. (1995). *Justice as Impartiality*. Clarendon Press, Oxford.

Beccaria, C. (1764). *Dei delitti e delle pene.* Rizzoli, 1994, Milan.

Bentham, J. (1776). A fragment on government. In J. Bowring (ed.) (1838), *The Works of Jeremy Bentham*, volume 1, 221–295. Reprinted by Russel and Russel, New York, 1968.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation, The Hafner Library of Classics.* MacMillan, 1948, New York.

Bergson, A. (1954). On the concept of social welfare. *Quarterly Journal of Economics*, 68: 233-252.

Binmore, K. (1994). *Game Theory and the Social Contract*, volume 1. The M.I.T. Press, Cambridge (Mass).

Blackorby, C., Bossert, W. and D. Donaldson (1995a). Income inequality measurement. To appear in J. SIlbert (ed.), *Income Inequality Measurement: From Theory to Practice.* Dordrecht, Kluwer.

Blackorby, C. Bossert, W. and D. Donaldson (1995b). Interpersonal population ethics: critical-level utilitarian principles. *Econometrica*, 63:1303–1320.

Blackorby, C. and D. Donaldson (1978). Measures of relative equality and their meaning in terms of social welfare. *Journal of Economic THeory*, 18: 59–80.

Blackorby, C. and D. Donaldson (1984). Social criteria for evaluating population changes. *Journal of Public Economics*, 25:13–33.

Blackorby, C., D. Donaldson and J. Weymark (1984). Social choice with interpersonal utility comparisons: a diagrammatic introduction. *International Economic Review*, 25:327–356.

Blackorby, C., Donaldson, D. and J. Weymark (1996). Harsanyi's social aggregation theorem for state-contingent alternatives. Discussion Paper n° 96-26, Department of Economics, University of British Columbia, Vancouver.

Blaug, M. (1980). *Economic Methodology.* Cambridge University Press, Cambridge.

Boadway, R.W. and N. Bruce (1984). *Welfare Economics.* Blackwell, Oxford.

Bouyssou, D. and J. Vansnick (1990). Utilité cardinale dans le certain et choix dans le risque. *Revue Economique*, 6: 979-1000.

Bowring, J. (ed.) (1838). *The Works of Jeremy Bentham*, volume 3. Reprinted by Russell and Russel, New York, 1962.

Brandt, R. (1959). *Ethical Theory.* Prentice-Hall, Englewood Cliffs, N.J.

Brandt, R. (1979). *A Theory of the Good and the Right.* Clarendon, Oxford.

Brandt, R. (1992). *Morality, Utilitarianism and Rights.* Cambridge University Press, Cambridge.

Broome, J. (1978). Choice and value in economics. *Oxford Economic Papers*, 30: 313–333.

Broome, J. (1989). SHould social preferences be consistent? *Economics and Philosophy*, 5, 7–17.

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

Broome, J. (1990-91). Fairness. *Proceedings of the Aristotelian Society*, 91: 87–102.

Broome, J. (1991a). *Weighing Goods.* Blackwell, Oxford.

Broome, J. (1991b). Utility. *Economics and Philosophy*, 7:1–12.

Broome, J. (1992). *Counting the Cost of Global Warming.* WHite Horse, Cambridge.

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

Broome, J. (1996). The welfare economics of population. *Oxford Economic Papers*, 48: 177–193.

Buchanan, J.M. (1954). Individual choice in voting and the market. *Journal of Political Economy*, 62:334–343.

Canto-Sperber, M. (ed.) (1996). *Dictionnaire d'éthique et de philosophie morale.* Presses Universitaires de France, Paris.

Chipman, J.S., Hurwicz, L., Richter, M.K. and H.F. Sonnenschein (eds.), 1971. *Preferences, Utility and Demand.* Harcourt Brace Jovanovich, New York.

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

Collard, D. (1978). *Altruism and Economy.* Martin Robertson, Oxford.

Coulhon, T. and P. Mongin (1989). Social choice theory in the case of von Neumann-Morgenstern utilities. *Social Choice and Welfare*, 6:175–187.

Cowen, T. (1989). Normative population theory. *Social Choice and Welfare*, 6: 33–43.

Daniels, N. (ed.) (1989). *Reading Rawls.* Stanford University Press, Stanford.

de Finetti, B. (1937). La prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68.

De Meyer, B. and P. Mongin (1995). A note on affine aggregation. *Economics Letters*, 47:177–183.

Debreu, G. (1960). Topological methods in cardinal utility theory. In K.J. Arrow, S. Karlin and P. Suppes (eds.), *Mathematical Methods in the Social Sciences*, 16–26, Stanford University Press, Stanford.

Deschamps, R. and L. Gevers (1978). Leximin and utilitarian rules: a joint characterization. *Journal of Economic Theory*, 17: 143–163.

Diamond, P.A. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: a comment. *Journal of Political Economy*, 75: 765–766.

Dérathé, R. (1951). *Jean-Jacques Rousseau et la science politique de son temps*. Vrin, Paris.

Dworkin, R. (1981). What is equality. I: Equality of welfare, and What is equality. II: Equality of resources. *Philosophy and Public Affairs*, 10, 185–256 and 283–345.

Edgeworth, F.Y. (1881). *Mathematical Psychics*. Kegan Paul, London. Reprinted by A.M. Kelley, New York, 1967.

Ellsberg, D. (1954). Classic and current notions of 'measurable utility'. *Economic Journal*, 64: 528–556.

Elster, J. and J.E. Roemer (eds.) (1991). *Interpersonal Comparisons of Well-Being*. Cambridge University Press, Cambridge.

Epstein, L.G. and U. Segal (1992). Quadratic social welfare functions. *Journal of Political Economy*, 100: 691–712.

Fagot-Largeault, A. (1991). Réflexions sur la notion de qualité de la vie. *Archives de philosophie du droit*. Droit et science, 26: 135–153. English translation in L. Nordenfelt (ed.), *Concepts and Measurement of Quality of Life in Health Care*, Dordrecht, Kluwer, 1992, 135–160.

Fine, T. (1973). *Theories of Probability*. Academic Press, New York.

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

Fishburn, P.C. (1982). *The Foundations of Expected Utility*. D. Reidel, Dordrecht.

Fishburn, P.C. (1984). On Harsanyi's utilitarian cardinal welfare theorem. *Theory and Decision*, 17:21–28.

Fishburn, P.C. (1986). The axioms of subjective probability. *Statistical Science*, 1: 335–358.

Fishburn, P.C. (1988). *Nonlinear Preference and Utility Theory.* The John Hopkins University Press, Baltimore.

Fishburn, P.C. (1989). Retrospective on the utility theory of von Neumann and Morgenstern. *Journal of Risk and Uncertainty*, 2: 127–158.

Fleurbaey, M. (1995). Equal opportunity or equal social outcomes. *Economics and Philosophy*, 11: 25–55.

Fleurbaey, M. (1996). *Théories économiques de la justice.* Economica, Paris.

Foley, D. (1967). Ressource allocation and the public sector. *Yale Economic Essays*, 7: 45–98.

Gaertner, W. and M. Klemisch-Ahlert (1992). *Social Choice and Bargaining Perspectives on Distributive Justice.* Springer, Berlin.

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

Gérard-Varet, L.A. and J.C. Passeron (eds.) (1995). *Le modèle et l'enquête. Les usages du principe de rationalité en sciences sociales.* Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Gevers, L. (1979). On interpersonal comparability and social welfare orderings. *Econometrica*, 47: 75–89.

Gibbard, A. (1974). A Pareto-consisstent libertarian claim. *Journal of Economic Theory*, 7: 388–410.

Gibbard, A. (1979). Disparate goods and Rawls' difference principle: a social choice theoretic treatment. *Theory and Decision*, 11: 267–288.

Glannon, W. (1995). Equality, priority, and numbers. *Social Theory and Practice*, 21: 427–455.

Goodin, R.E. (1986). Laundering preferences. In J. Elster and A. Hylland (eds), *Foundations*

of Social Choice Theory, 75–101. Cambridge University Press, Cambridge.

Gorman, W.M. (1968). The structure of utility functions. *Review of Economic Studies*, 35: 367–390.

Graaff, J.d.V. (1957). *Theoretical Welfare Economics*. Cambridge University Press, Cambridge.

Griffin, J. (1986). *Well-Being*. Clarendon Press, Oxford.

Halévy, E. (1901-1904). *La formation du radicalisme philosophique*, 3 volumes. Félix Alcan, Parix, Paris. New edition, Paris, Presses Universitaires de France, 1995.

Hammond, P.J. (1976). Equity, Arrow's conditions and Rawls's difference principle. *Econometrica*, 44, 793–804.

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

Hammond, P.J. (1982). Utilitarianism, uncertainty and information. In A.K. Sen and B. Williams (1982), 85–102.

Hammond, P.J. (1983). Ex-post optimality as a dynamically consistent objective for collective choice under uncertainty. In P. Pattanaik and M. Salles (eds.), *Social Choice and Welfare*, 175–205. North-Holland, Amsterdam.

Hammond, P.J. (1987). On reconciling Arrow's theory of social choice with Harsanyi's fundamental utilitarianism. In G. Feiwel (ed.), *Arrow and the Foundations of the Theory of Economic Policy*, 79–221. New York University Press, New YorK

Hammond, P.J. (1988a). Consequentialist foundations for expected utility theory. *Theory and Decision*, 25, 25–78.

Hammond, P.J. (1988b). Consequentialist demographic norms and parenting rights. *Social*

Choice and Welfare, 5, 127–145.

Hammond, P.J. (1991). Interpersonal comparisons of utility: why and how they should be made. In J. Elster and J.E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, 200–254. Cambridge University Press, Cambridge.

Hammond, P.J. (1995). Social choice of individual and group rights. In W.A. Barnett, H. Moulin, M. Salles and N.J. Schofield (eds.), *SOcial Choice, Welfare, and Ethics*, 55–77, Cambridge University Press, Cambridge.

Hammond, P.J. (1996). Consequentialist decision theory and utilitarian ethics. In F. Farina, F. Hahn and S. Vanucci (eds.), *Ethics, Rationality and Economic Behaviour*, 92–118, Clarendon Press, Oxford.

Hare, R.M. (1976). Ethical theory and utilitarianism. In H.D. Lewis (ed.), *Contemporary British Philosophy*. Reprinted in A.K. Sen and B. Williams (1982), 23–38.

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. Repritned in J.C. Harsanyi (1976), chapter 1.

Harsanyi, J.C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63, 309–321. Reprinted in J.C. Harsanyi (1976), chapter2.

Harsanyi, J.C. (1958). Ethics in terms of hypothetical imperatives. *Mind*, 67: 305–316. Reprinted in J.C. Harsanyi (1976), chapter 3.

Harsanyi, J.C. (1967-1968). Games with incomplete information played by 'Bayesian' players. *Management Science*, 14: 159–182, 320–334, and 486–502.

Harsanyi, J.C. (1975). Nonlinear social welfare functions: Do welfare economists have a special exemption from Bayesian rationality? *Theory and Decision*, 6:311-332. Reprinted in J.C. Harsanyi (1976), chapter 5.

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

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

Harsanyi, J.C. (1977b). Morality and the theory of rational behavior. *Social Research*, 44. Reprinted in A.K. Sen and B. Williams (1982), 39–62.

Harsanyi, J.C. (1977c). Rule utilitarianism and decision theory. *Erkenntnis*, 11, 25–33.

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

Harsanyi, J.C. (1992). Games and decision-theoretic models in ethics. In R. Aumann and S. Hart (eds.), *Handbook of Game Theory*, Volume 1, chapter 19, 669–707. North-Holland, Amsterdam.

Hausman, D. and M. McPherson (1994). Preference, belief, and welfare. *American Economic Review. Papers and Proceedings*, 84: 396–400.

Hausman, D. and M. McPherson (1996). *Economic Analysis and Moral Philosophy.* Cambridge University Press, Cambridge.

Herstein, I.N. and J. Milnor (1953). An axiomatic approach to measurable utility. *Econometrica*, 21, 291–297.

Hicks, J.R. (1939). *Value and Capital.* Oxford University Press, Oxford. 2nd edition, 1946.

Hicks, J.R. (1956). *A Revision of Demand Theory.* Clarendon Press, Oxford.

Hild, M., R. Jeffrey and M. Risse (1997). Problems of preference aggregation. In M. Salles and J. Weymark (eds.), *Justice, Political Liberalism and Utilitarianism: Themes from Harsanyi and Rawls.* Cambridge University Press, Cambridge.

Hobbes, T. (1651). *Leviathan.* Blackwell, Oxford.

Howson, C. and P. Urbach (1989). *Scientific Reasoning. The Bayesian Approach.* Open Court, Peru, Ill. Revised edition, 1993.

Hume, D. (1736). Treatise on human nature. Reprinted 1896 by Clarendon Press, Oxford.

Hurwicz, L, D. Schmeidler and H.F. Sonnenschein (eds.) (1985). *Social Goals and Social Organization.* Cambridge University Press, Cambridge.

Hutcheson, F. (1725). *On the Nature and Conduct of the Passions and Affections* Foulis, Glasgow.

Hylland, A. and R. Zeckhauser (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.* University of Chicago Press, Chicago. 2nd revised edition, 1983.

Jevons, S. (1871). *The Theory of Political Economy.* Macmillan, London.

Kagan, S. (1992). The limits of well-being. In E. Paul, F. Miller and J. Paul (eds.), *The Good Life and the Human Good*, 169–189. Cambridge University Press, Cambridge.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47, 263–291.

Kalai, E. (1975). Solutions to the bargaining problem. In L. Hurwicz, D. Schmeidler and H.F. Sonnenschein (1985), chapter 3.

Kalai, E. and M. Smorodinsky (1975). Other solutions to Nash's bargaining problem. *Econometrica*, 47, 1623–1630.

Kaneko, M. (1984). On interpersonal utility comparisons. *American Economic Review. Papers and Proceedings*, 1, 165–175.

Kant, I. (1785). Grundlegung zur Metaphysik der Sitten. English translation by H.J. Paton, *The Moral Law*. Hutchinson, London, 1953.

Kant, I. (1788). Kritik der Praktischen Vernunft. English translation by L.W. Beck, *Critique of Practical Reason*. Bobbs-Merril, Indianapolis, 1977.

Karni, E. (1996). Social welfare functions and fairness. *Social Choice and Welfare*, 13, 487–496.

Karni, E. and J. Weymark (1998). An informationally parsimonious impartial observer theorem. *Social Choice and Welfare*, 15(3), 321–332.

Kelsey, D. (1987). The role of information in social welfare judgments. *Oxford Economic Papers*, 39, 301–317.

Kolm, S.C. (1972). *Justice et Equité.* Editions du Centre National de la Recherche Scientifique, Paris.

Kolm, S.C. (1974). Sur les conséquences économiques de principes de justice et de justice pratique. *Revue d'Economie Politique*, 84, 80–107.

Kolm, S.C. (1984). *La Bonne Economie. La Réciprocité Générale*. Presses Universitaires de France, Paris.

Kolm, S.C. (1993). The impossibility of utilitarianism. In P. Koslowski and Y. Shionoya (eds.),

The Good and the Economical, 30–66. Springer, Berlin.

Kolm, S.C. (1995). *Modern Theories of Justice*. The M.I.T. Press, Cambridge (Mass).

Kolm, S.C. (1998). Chance and justice: Social policies and the Harsanyi-Vickrey-Rawls problem. *European Economic Review*, 42(8), 1393–1416.

Krantz, D.H., Luce, R.D., Suppes, P. and A. Tversky (1971). *Foundations of Measurement*, volume 1. Academic Press, New York.

Lange, O. (1942). The foundations of welfare economics. *Econometrica*, 10, 215–228.

Lauwers, L. (1995). Time-neutrality and linearity. *Journal of Mathematical Economics*, 24, 347–351.

Lerner, A. (1944). *The Economics of Control*. Macmillan, London.

Levi, I. (1990). Pareto unanimity and consensus. *Journal of Philosophy*, 87, 481–492.

Little, I.M.D. (1950). *A Critique of Welfare Economics*. Clarendon Press, Oxford.

Little, I.M.D. (1952). Social choice and individual values. *Journal of Political Economy*, 60, 422–432.

Locke, J. (1690). Second treatise of civil government. In P. Laslett (ed.) (1960), *Two Treatises of Government*, Cambridge University Press, Cambridge.

Loomes, G. and R. Sugden (1982). Regret theory: An alternative to rational choice under uncertainty. *Economic Journal*, 92, 805–824.

Luce, R.SD and H. Raiffa (1957). *Games and Decisions*. Wiley, New York.

Lyons, D. (1965). *Forms and Limits of Utilitarianism*. Clarendon Press, Oxford.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, 60, 151–158.

Mas-Colell, A., Whinston, M. and J. Green (1995). *Microeconomic Theory*. Oxford University Press, New York.

Maskin, E. (1978). A theorem on utilitarianism. *Review of Economic Studies*, 45, 93–96.

McClennen, E.F. (1990). *Rationality and Dynamic Choice*. Cambridge University Press, Cambridge.

McKay, A.F. (1986). Extended sympathy and interpersonal utility comparisons. *Journal of Philosophy*, 83, 305–322.

Mill, J.S. (1859). On liberty. In H.B. Acton (1972). *Utilitarianism, Liberty, Representative Government*, 65–170. Dent Dutton, London.

Mill, J.S. (1863). Utilitarianism. In H.B. Acton 1972). *Utilitarianism, Liberty, Representative Government*, 1–61. Dent Dutton, London.

Mirrlees, J.A. (1974). Notes on welfare economics, information and uncertainty. In M. Balch, D. McFadden and S. Wu (eds.)., *Essays on Economic Behavior under Uncertainty*, 243-258. North-Holland, Amsterdam.

Mirrlees, J.A. (1982). The economic use of utilitarianism. In A.K. Sen and B. Williams (1982), 63–84.

Mongin, P. (1984). Modèle rationnel ou modèle économique de la rationalité? *Revue Economique*, 35, 9–64.

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

Mongin, P. (1994b). L'optimisatio est-elle un critère de rationalité individuelle? *Dialogue*, 33: 191–222. Reprinted in L.-A. Gérard-Varet and J.C. Passeron (1995), 279–307.

Mongin, P. (1995a). Consistent Bayesian aggregation. *Journal of Economic Theory*, 66, 131–351.

Mongin, P. (1995b). L'utilitarianisme originel et le développement de la théorie économique. Postface to the new edition of E. Halévy (1901–1904). Also published in *La Pensée Politique*, 3, 341–361.

Mongin, P. (1996). The paradox of the Bayesian experts and state-dependent utility theory. *Journal of Mathematical Economics*, 29(3), 331–361, 1998.

Monroe, D.H. (ed.) (1972). *A Guide to the British Moralists*. Fontana, London.

Moore, G.E. (1903). *Principia Ethica*. Cambridge University Press, Cambridge.

Morris, S. (1995). The common prior assumption. *Economics and Philosophy*, 11, 227–253.

Moulin, H. (1988). *Axioms of Cooperative Decision Making*. Cambridge University Press, Cambridge.

Moulin, H. (1995). *Cooperative Microeconomics*. Princeton University Press, Princeton.

Moulin, H. (1996). Stand alone and unanimity tests: a reexamination of fair division. In F. Farina, F. Hahn and S. Vanucci (eds.), *Ethics, Rationality and Economic Behaviour*, 121–142. Clarendon Press, Oxford.

Myerson, R.B. (1981). Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica*, 49, 883–897.

Nagel, T. (1970). *The Possibility of Altruism*. Oxford University Press, Oxford.

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

Nelson, M. (1991). Utilitarian eschatology. *American Philosophical Quarterly*, 28, 339–347.

Nozick, R. (1974). *Anarchy, State and Utopia*. Basic Books, New York.

Pareto, V. (1896-97). *Cours d'économie politique*. In *Oeuvres Complètes*, volume 1. Droz, 1964, Genève.

Pareto, V. (1909). *Manuel d'économie politique*. In *Oeuvres Complètes*, volume 7. Droz, 1966, Genève.

Pareto, V. (19717-1919). *Traité de sociologie générale*. In *Oeuvres Complètes*, volume 12. Droz, 1968, Genève.

Parfit, D. (1984). *Reasons and Persons*. Clarendon Press, Oxford.

Pattanaik, P.K. (1968). Risk, impersonality, and the social welfare function. *Journal of Political Economy*, 76, 1152–1169.

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

Pivacet, E. (1996). *Choix rationnel et vie publique*. Presses Universitaires de France, Paris.

Pigou, A.C. (1920). *The Economics of Welfare*. Macmillan, London, 4th revised, 1932 edition.

Plato. Republic. English translation by Grube (1974). Hackett, Indianapolis

Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32, 122-136.

Ramsey, F.P. (1931). Truth and probability. In R.B. Braithwaite (ed.), *The Foundation of Mathematics and other Logical Essays*, 156–198, Harcourt and Brace, New York.

Rawls, J. (1958). Justice as fairness. *Philosophical Review*, 67, 164–194.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Cambridge, Mass.

Rawls, J. (1980). Kantian constructivism in moral theory. *Journal of Philosophy*, 77, 515–572.

Rawls, J. (1982). Social unity and primary goods. In A.K. Sen, and B. Williams (1982), 159–185.

Rawls, J. (1988). The priority right and ideas of the good. *Philosophy and Public Affairs*, 17, 251–276.

Richter, M. (1971). Rational choice. In J.S. Chipman, L. Hurwicz, M.K. Richter and H.F. Sonnenschein (1971), 29–58.

Riley, J. (1988). *Liberal Utilitarianism. Social Choice Theory and J.S. MIll's Philosophy.* Cambridge University Press, Cambridge.

Robbins, L. (1932). *An Essay on the Nature and Significance of Economics.* MacMillan, London, 2nd revised edition, 1937.

Roberts, K.W.S. (1980a). Interpersonal comparability and social choice theory. *Review of Economic Studies*, 47, 421–439.

Roberts, K.W.S. (1980b). Social choice theory: the single and multi-profile approaches. *Review of Economic Studies*, 47, 411–450.

Roberts, K.W.S. (1980c). Possibility theorems with interpersonally comparable welfare levels. *Review of Economic Studies*, 47, 409–420.

Roberts, K.W.S. (1995). Valued opinions and opinionized values: The double aggregation problem. In K. Basu, P. Pattanaik, and K. Suzumura (eds.), *Choice, Welfare and Development*, 141–165. Oxford University Press, Oxford.

Roemer, J.E. (1985). Equality of talent. *Economics and Philosophy*, 1, 151–181. Reprinted in Roemer (1994), essay 6.

Roemer, J.E. (1986). The mismarriage of bargaining theory and distributive justice. *Ethics*, 97, 88–110. Reprinted in Roemer (1994), essay 9.

Roemer, J.E. (1994). *Egalitarian Perspectives.* Cambridge University Press, Cambridge.

Roemer, J.E. (1996). *Theories of Distributive Justice.* Harvard University Press, Cambridge (Mass).

Rousseau, J.J. (1755). Discours sur l'origine de l'inégalité. In Dérathé, R. et al. (eds.), *Oeuvres complètes*, volume 3. NRF, Bibliothèque de la Pléiade, Paris. English translation in "The Social Contract and Discourses", by G.D.H. Cole (1950). Dutton, New York.

Rousseau, J.J. (1761). Du contrat social. In Dérathé, R. et al. (eds.), *Oeuvres complètes*, volume 3. NRF, Bibliothèque de la Pléiade, Paris. English translation in "The Social Contract and Discourses", by G.D.H. Cole (1950), Dutton, New York.

Rubinstein, A. (1984). The single profile analogues to multi-profile theorems: Mathematical Logic's approach. *International Economic Review*, 25, 719–730.

Samuelson, P.A. (1967). Arrow's mathematical politics. In S. Hook (ed.), *Human Values and Economic Policy*, 41–50. New York University Press, New York.

Savage, L.J. (1954). *The Foundations of Statistics.* Wiley, New York, 2nd revised, 1972 edition.

Scanlon, T.M. (1975). Preference and urgency. *Journal of Philosophy*, 72, 655–669.

Scheffler, S. (1982). *The Rejection of consequentialism.* Oxford University Press, Oxford.

Scheffler, S. (ed.) (1988). *Consequentialism and its Criticis.* Oxford University Press, Oxford.

Schervish, M.J.,Seidenfeld, T. and J.B. Kadane (1990). State-dependent utilities. *Journal of the American Statistical Association*, 85, 840–847.

Schervish, M.J., Seidenfeld, T. and J.B. Kadane (1991). Shared preferences and state-dependent utilities. *Management Science*, 37, 1575–1589.

Schumpeter, J. (1954). *History of Economic Analysis.* Macmillan, London.

Screpanti, E. and S. Zamagni (1993). *History of Economic Thought*. Oxford UniversityPress, Oxford.

Seidenfeld, T. Kadane, J.B. and M.J. Schervish (1989). On the shared preferences of two Bayesian decision makers. *Journal of Philosophy*, 86, 225–244.

Sen, A.K. (1970). *Collective Choice and Social Welfare*. North Holland, Amsterdam.

Sen, A.K. (1973a). Behaviour and the concept of preference. *Economica*, 40, 241–259. Reprinted in Sen (1982a), chapter 2.

Sen, A.K. (1973b). *On Economic Inequality*. Clarendon Press, Oxford.

Sen, A.K. (1974). Rawls versus Bentham: An axiomatic examination of the pure distribution problem. *Theory and Decision*, 4, 301–309.

Sen, A.K. (1976). Liberty, unanimity and rights. *Economica*, 43, 217–245. Reprinted in Sen (1982a), chapter 14.

Sen, A.K. (1979). Utilitarianism and welfarism. *Journal of Philosophy*, 76, 463–480.

Sen, A.K. (1980). *Equality of What? Tanner Lectures on Human Values*, volume 1. Cambridge University Press, Cambridge. Reprinted in Sen (1982a), chapter 16 and in Sen (1992), 12–30.

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

Sen, A.K. (1982a). *Choice, Welfare and Measurement*. Blackwell, Oxford.

Sen, A.K. (1982b). Rights and agency. *Philosophy and Public Affairs*, 11, 3–39. Reprinted in S. Scheffler (1988).

Sen, A.K. (1985). *Commodities and Capabilities*. North Holland, Amsterdam.

Sen, A.K. (1986). Social choice theory. In K.J. Arrow and M.D. Intriligator (eds.), *Handbook of*

Mathematical Economics, volume III, 1073–1181. North Holland, Amsterdam.

Sen, A.K. (1987). *On Ethics and Economics*. Blackwell, Oxford.

Sen, A.K. (1992). *Inequality Reexamined*. Clarendon Press, Oxford.

Sen, A.K. and B. Williams (eds.) (1982). *Utilitarianism and Beyond*. Cambridge University Press, Cambridge.

Sidgwick, H. (1884). The mehtod of ethics. MacMillan, London, 7th revised edition, 1907.

Sikora, R. and B. Barry (eds.) (1978). *Obligations to Future Generations*. Temple, Philadelphia.

Singer, P. (ed.) (1991). *A Companion to Ethics*. Blackwell, Oxford.

Smart, J.J.C. (1974a). An outline of a system of utilitarian ethics. In J.J.C. Smart and B. Williams (eds.), *Utilitarianism: For and Against*, 3–74. Cambridge University, Cambridgie.

Smith, A. (1759). The theory of moral sentiments. Revised 1790. New edition by D.D. Raphael and A.L. Macfie. Oxford University Press, Oxford.

Stephen, L. (1901). The English utilitarians. Reprinted 1968 by A.M. Kelley, New York.

Stigler, G.J. (1965). *The History of Economics*. The University of Chicago Press, Chicago.

Strasnick, S. (1976). Social choice and the derivation of Rawls' difference principle. *Journal of Philosophy*, 73, 184–194.

Suppes, P. (1966). Some formal models of grading principles. *Synthese*, 6, 284–306. Reprinted in P. Suppes (1969). *Studies in Methodology and Foundations of Science*, Dordrecht, Reidel.

Suppes, P. (1981). *Logique du probable*. Flammarion, Paris.

Suppes, P. and M. Winet (1955). An axiomatization of utility based on the notion of utility differences. *Management Science*, 1, 259–270.

Suzumura, K. (1983). *Rational Choice, Collective Decisions and Social Welfare*. Cambridge University Press, Cambridge.

Suzumura, K. (1994). Interpersonal comparisons of the extended sympathy type and the possibility of social choice. Discussion Paper No. 295, Institute of Economic Research, Hitotsubashi University. Published in K.J. Arrow, A.K. Sen and K. Suzumura (eds.) (1996). *Social Choice Re-Examined*, volume 2, 202–229. London, MacMillan.

Temkin, L. (1987). Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16, 138–187.

Temkin, L. (1993). *Inequality*. Oxford University Press, New York.

Thomson, W. (1982). Equity in exchange economies. *Journal of Economic Theory*, 18, 217–244.

Thomson, W. (1994). L'absence d'envie: une introduction. *Recherches Economiques de Louvain*, 60, 43–61.

Thomson, W. (1995). Population monotonic allocation rules. In W.A. Barnett, H. Moulin, M. Salles and N.J. Schofield (eds.), *Social Choice Welfare and Ethics*, 79–124. Cambridge University Press, Cambridge.

Thomson, W. and H. Varian (1985). Theories of justice based on symmetry. In L. Hurwicz, D. Schmeidler and H.F. Sonnenschein (1985), 107–129.

Vallentyne, P. (ed.) (1991). *Contractarianism and Rational Choice*. Cambridge University Press, Cambridge.

Vallentyne, P. (1994). Infinite utility and temporal neutrality. *Utilitas*, 6, 193–199.

Van Liedekerke, L. and L. Lauwers (1997). Sacrificing the patrol. *Economics and Philosophy*, 13, 159–174.

Van Parijs, P. (1990). Equal endowments as undominated diversity. *Recherches Economiques de Louvain*, 56, 327–355.

Van Parijs, P. (1991). *Qu'est-ce qu'une société juste?* Le Seuil, Paris.

Varian, H. (1974). Equity, envy and efficiency. *Journal of Economic Theory*, 9, 63–91.

Varian, H. (1985). Dworkin on equality of resources. *Economics and Philosophy*, 1, 110–125.

Verneaux, R. (1971). *Le vocabulaire de Kant.* Aubier-Montaigne, Paris.

Vickrey, W. (1945). Measuring marginal utility by reaction to risk. *Econometrica*, 13: 319–333.

Vickrey, W.S. (1960). Utility, strategy, and social decision rules. *Quarterly Journal of Economics*, 74, 507–535.

Viner, J. (1949). Bentham and J.S. Mill. *American Economic Review*, 39, 360–382.

von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior.* Princeton University Press, Princeton. 2nd edition, 1947, 3rd edition, 1953.

Wakker, P. (1989). *Additive Representation of Preferences.* Kluwer, Dordrecht.

Wasserman, D. (1996). Let them eat chances: probability and distributive justice. *Economics and Philosophy*, 12: 29–49.

Weymark, J. (1991). A reconsideration of the Harsanyi-Sen debate on utilitarianism. In J. ELster and J.E. Roemer (eds.) (1991), 255-320.

Weymark, J. (1993). Harsanyi's social aggregation theorem and the weak Pareto principle. *Social CHoice and Welfare*, 10: 209–221.

Williams, B. (1973). A critique of utilitarianism. In J.J.C. Smart and B. Williams (eds.), *Utilitarianism: For and Against*, 77–150. Cambridge University Press, Cambridge.

Yaari, M.E. (1981). Rawls, Edgeworth, Shapley, Nash: theories of distributive justice re-examined. *Journal of Economic Theory*, 24:1–39.

Yaari, M.E. and m. Bar-Hillel (1984). On dividing justly. *Social Choice and Welfare*, 1, 1–24.

Zhou, L. (1997). Harsanayi's utilitarianism theorems: general societies. *Journal of Economic Theory*, 72, 198–207.