Rational and moral action: a critical survey of rational choice theory

de Jonge, J.P.R.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
1. Introduction

In chapter VI I introduced utilitarian ethics and I said that the sentiment to which utilitarianism appeals is generalized benevolence. This expresses the notion that an agent counts himself neither more nor less than any other person. As benevolent people, utilitarians want as many people as possible, now and in the future, to be as happy as possible, or to enjoy as much utility as possible. To that end, the preferences of each individual should count equally. The moral choice would be to choose that social state in which social utility is maximized. Social states must be compared and ranked and to that end social welfare functions must be constructed that are based on the ethical preferences of all individuals.

So far, so good; but in developing this idea several problems arise, such as the possibility of interpersonal comparisons and
the question whether, besides looking at average well-being, also the distribution of well-being matters and, last but not least, whether in Harsanyi's approach there is any relation at all between preferences and well-being. These issues will be discussed in the sections to follow. In the sections 3 and 4 I shall discuss extended preferences and the impartial observer theorem, subjects that are directly related to the problem of interpersonal comparisons. In section 5 the question of the dispersion of utilities and the inequality of well-being is considered. In section 6 the question is raised whether Harsanyi's approach is still concerned with well-being. And in section 7 I present my conclusions. But I first introduce Harsanyi's construction of a social welfare function.

2. Harsanyi's social welfare function

In this section I will present Harsanyi's version of utilitarian ethics. (see Harsanyi; 1953, 1955, 1975, 1977, 1982, 1985) According to Harsanyi the updated version of classical utilitarianism (neo-utilitarianism) is the only theory that treats moral behavior as a special form of rational behavior.

The emergence of modern decision theory has made ethics an organic part of the general theory of rational behavior, which relies on Bayesian rationality postulates and on Pareto optimality. Harsanyi believes that if one takes the rationality postulates of Bayesian theory seriously, then one can obtain a clear and unambiguous solution for the problem of defining an adequate social welfare function. It can be shown that the social welfare function must be a linear function of all individual utilities; this means that it must be defined as the arithmetic mean of the utility levels of all individuals.

323 For a detailed discussion of Harsanyi's social welfare function I refer the reader to the article by Mongin and d' Aspremont “Utility Theory and Ethics” (1998).
324 "Ethics is a theory of rational behavior in the service of the common interests of society as a whole." (Harsanyi, 1982, 43)
In using individual welfare functions for collective choice, there are at least three problems: a) the measurability of individual welfare; b) the possibility of interpersonal comparisons and c) the form of a function, which will specify a social preference relation. Or, to formulate this differently, is the function related to any relevant conception of utility, i.e., does it relate to well-being? In order to characterize the relationship between the social welfare function of a given individual i and the utility functions of all individuals, Harsanyi introduced in his earlier articles implicitly two models. In the 1953 article he formulated the Impartial Observer Theorem", and in his 1955 article he stated his "Aggregation Theory". One can regard these models as delivering the moral respectively the mathematical argument for the linearity of the social welfare function.

In the first model Harsanyi explicates the notion of social welfare through the welfare judgments of individuals. The problem of aggregating these individual social welfare functions he did not undertake, because he thought this to be superfluous. He assumed that the social welfare functions would be the same for each individual if full factual information were available to the evaluating individual about other people's utilities.

The Aggregation model is based on the following argument: there is a group of individuals each of whom has a preference relation on the lottery L. Based on these individual preferences (U_i), a collective or social preference relation (W) is determined. The individual and the social preferences are related by a Pareto principle. Each Pareto principle is a formalization of the idea that unanimity should be respected. When both individual and social preferences satisfy the axioms of expected utility theory, then it is possible to represent each individual’s preference relation U_i by a vNM utility function V_i, and it is possible to represent the social preference relation W by a vNM utility function V. Given Pareto Indifference the social utility function must be a weighted combination of the individual utility functions. This is Harsanyi’s Aggregation Theorem that is based on three
postulates: a) the personal preferences of i can be represented by a vNM utility function; b) the moral preferences of i can be represented by a vNM utility function and c) if i’s welfare is greater in social state x than in social state y and if all other individuals are personally indifferent between the two social states, then social welfare is greater in social state x than in social state y. This Pareto condition provides the link between the individual utility functions and the social welfare function of i.

\[ V(p) = \sum_{i=1}^{n} a_i V_i(p) + b \]

The three postulates together imply that i’s social welfare function will be a linear combination of individual utilities. I already discussed this in chapter VI.

In absence of any objective criteria to compare the utilities of each individual with each other, an individual’s social welfare function will take the form of a weighted sum of all individual utility functions with more or less arbitrary weights chosen according to his own value judgements. People’s preferences may differ and people may react differently to similar situations and this does present a real difficulty in comparing the utilities enjoyed by different people. Harsanyi refers to this as the psychological difficulty. But he is confident that the more complete people’s information and the more value judgements are shared, the more the individual’s social welfare functions will converge toward the objective quantity, namely the unweighed sum of all individual utilities. (Harsanyi, 1955) Harsanyi did not really aggregate individual preference relations; instead of this he took one of those preference relations, constituted not by actual preferences but by moral preferences, as representing the social preference relation. In doing this Harsanyi departed from the Aggregation model and switched to the Impartiality model. This is the model of the impartial observer, which has two variants.
In his 1953 article Harsanyi assumed that the impersonality of an individual’s judgement concerning social welfare requires that the individual be completely ignorant about his or her relative position within actual society. The impersonal judgement of preferences amounts to a choice between equiprobable prospects made by an outside observer. This construction personalizes the ethical problem by setting the observer as judge over alternative social states and inquiring as to the principles he might use to reach his judgement. Impartiality in this context implies that the observer must show equal concern for each individual and, therefore, equal respect for the preferences of each individual. Harsanyi uses the intrapersonal comparison of the ideal observer, whose preferences take account of the cardinal information built into the utility functions of each individual, to solve the problem of the interpersonal comparison of utility. It follows that the ideal observer's utility function has exactly the same form as the utilitarian welfare function. But treating society as a person begs all the important questions.325

Therefore Harsanyi took another road by replacing the outside observer by an inside observer. In his 1955 article the requirement of impartiality is equated with the idea that one has the chance of occupying anyone else's social position, including his or her subjective features. The impartial judgement is obtained by complete empathetic identification. The inside observer perspective relies on the similarity postulate, that is the assumption that "once proper allowances are made for empirically differences in tastes, education, etc, between me and other persons, then it is reasonable to assume that our basic psychological reaction to any given alternative will be otherwise much the same". (Harsanyi quoted by Fontaine, 2001, 400) In the context of the original position it is very

325 Rawls observed that the consequence of this view is to extend to society the principles of choice for one man. Utilitarianism does not take seriously the distinction between persons. (Rawls, 1971, 27) His criticism focuses also on the inappropriateness of the analogy between individual choice and the choice of a social ethic. Moreover, since all persons have been collapsed into one, so to speak, distribution drops out as a value.
tempting to use this apparent paradox. Is it not the purpose of the veil of ignorance to remove everything that makes people different? If so, then one can count on the moral preferences held by those in the original position to be the same. The psychological difficulty has, in Harsanyi’s opinion, been solved. From this similarity postulate the equivalence of intra- and interpersonal comparisons can be deduced.326

This conclusion, however, has met with severe criticism. And I shall investigate Harsanyi’s Impartial Observer Theorem in more detail in the following two sections. To solve the problem of interpersonal comparisons Harsanyi uses two devices: extended preferences and complete identification.

3. Extended preferences

A conclusion drawn by ordinalists is that we cannot compare one person’s good with another’s. Person i’s preference will inform us whether one alternative is better for i than another. Person j’s preference will tell whether an alternative is better for j than another. But no one’s preferences will tell us whether one alternative is better for i than another is for j. Extended preferences are invented as a basis for interpersonal comparisons. "An extended preference is a preference between extended alternatives and an extended alternative consists of a way of life paired together with particular personal characteristics". (Broome, 1999c, 31) The idea is that everyone will have the same extended preferences, provided that the extended alternatives are construed widely enough. Harsanyi calls a utility function that represents extended

326 The similarity hypothesis asserts that rational individuals placed in the same situation will respond identically: "if Peter had Paul’s biological make up, had Paul’s life history behind him, and were currently subject to Paul’s environmental influences, then he would presumably have the same personal preferences as Paul has and would ascribe the same utility as Paul does now to each particular situation’. (my italics, Harsanyi quoted by Broome, 1999c, 32) But when Peter had Paul’s life history behind him, he wouldn’t be Peter but Paul.
preferences an extended utility function.\textsuperscript{327} Let ’s investigate an extended preference function (I follow Broome, 1999c). Let R stand for the causal variables (her upbringing, her friends, education and so on) that determine the form of a person’s preferences. A person who is subject to the causal influences R will have preferences that can be represented by a utility function U(R). Any person subjected to the same influences R will have the same preferences. Thus, everyone with a way of life A will have the same utility function U (A; R). This function U is a universal function, the same for everybody. It represents the preferences that a person subjected to causal influences R will have about ways of life A. Harsanyi calls Vi(Aj, Rj) an extended utility function which is Broome’s U (A; R). Harsanyi asserts that it represents extended preferences over ways of life A and causal variables R together. But Broome denies this.

Harsanyi hoped to construct universal extended preferences as being the same for everyone. But in Broome’s opinion he overlooked a complication. The things that have a causal influence over people’s preferences may also be things that people have preferences about. For instance, people have preferences over friendship and their friends, at the same time, influence their preferences. Harsanyi’s separation between ‘objective position A’ and causal variables R is too sharp. Let A be a variable that stands for ways of life. Let P stand for personal characteristics. Then (A, P) is an extended alternative, a way of life combined with particular personal

\textsuperscript{327} Mongin, and also Weymark, noticed that when Harsanyi argues about extended preferences, he replaces them with extended utility assessments. The utility values measure, in the opinion of Harsanyi, the individual satisfaction based on the realization of preferences, given the objective and subjective variables. This delivers uniformity of utility amounts (under the assumption of full comparability), but not uniformity of extended preferences. Mongin concludes that there is no way of bridging the gap between Harsanyi ’s argument in terms of utilities and his argument of the uniformity of extended preferences. (see, Weymark, 1991 and Mongin, 2001)
characteristics. People have extended preferences about these extended alternatives when we make sure that A and P are defined broadly enough to include anything that anyone has a preference about. Now there are causal variables R that determine people’s extended preferences. A person who is subject to causal preferences R will have extended preferences V (A, P; R). V is a universal function, for anyone subjected to the same causal variables will have the same extended preferences. R contains the causal variables that a person is actually subjected to. (A, P), on the other hand, contains the values of the variables that the person contemplates as objects of her preference. V (A, P; R) is a universal function. But it does not represent universal extended preferences. "It does not represent any preference at all over A, P and R taken together. It represents preferences over A and P, but many different preferences, one set of preferences for each value of the causal variable R". (Broome, 1999c, 36) The causal argument was offered as a demonstration that everyone will have the same extended preferences. It fails, Broome concludes.

There is a suggestion that extended preferences could be universal when everyone had the same information. Let us, for instance, ask how we should solve the question whether it would be better to live in conditions x having the characteristics of i than to live in conditions y having the characteristics of j? The preference satisfaction theory would tell me that I (!) would be better off in conditions x with the characteristics of i than in conditions y with the characteristics of j if and only if I prefer (x,i) to (y,j). But this preference does not have to be universal. It does not say that i thinks that it would be better for her to live in conditions x than it would be for j to live in condition y, just because I myself prefer (x,i) to (y,j). The well-being that I derive for having my preference satisfied, does not have to be shared by i and j. A possible solution would be to say that only rational and well-informed extended preferences coincide (this is what Harsanyi actually believes). But even this conclusion is implausible. For it implies that my preference (x,i) > (y,j) only
differs from another person’s preference \((y, j) > (x, i)\) because of some irrationality or lack of information on the part of one of us (which Harsanyi also believes). But it could be the case that our preferences differ because we have different values. (see Broome, 1993) 

4. The impartial observer theorem

The important thing to notice about ‘hypothetical choice theories’ is that what hypothetical persons would choose is determined by the conditions to which they are subjected. (see Mongin, 2001) Therefore, the interpretation of the observer’s ignorance is very important. What is he allowed to know about i or j: a) their names; b) their preferences; or c) the causal factors, which determine their preferences? And what is the proper way to understand these causal factors? Do they bring about that i has the preferences he has? And is it implied that he could have had other preferences and still be himself? And, thus, could it be that j could be subjected to the same causal factors as i, and thus have the same preferences, while still being j? And if so, then j’s and i’s identities can be defined independently of what their preferences are. (Mongin, 2001, 156) Harsanyi clearly differentiates between an individual’s subjective attitudes (including his preferences) and all the objective causal variables needed to explain his subjective attitudes. His formal notion of extended preferences should be read with interpretation c) in mind. Mongin calls this version the causal one. Following the causal interpretation;

\[(x, k) \in^i (y, j)\] should be read as either:

328 Broome thinks that the preference satisfaction condition is false for extended preferences. He thinks that Arrow and Harsanyi accept it because they endorse a monistic theory of good. They believe that the goodness of life consists in one thing -happiness, satisfaction- and each life delivers a particular quantity of this thing. Consequently, it is only a matter of objective fact which of the two lives is better. (Broome, 1993, 67)
1e individual \(i\) prefers to be in state \(x\) and under the influence of the factors determining \(k\) ‘s preferences than to be in state \(y\) under the influence of the factors determining \(j\) ‘s preferences.

The more standard formulation is:

2e individual \(i\) prefers to be \(k\) in social state \(x\) to being \(j\) in social state \(y\).

The second formulation requires that the individual, in the position of observer \(i\), identifies himself with \(k\) ‘s and \(j\) ‘s personality or, at least, manages to reproduce in himself their preferential attitudes. The word empathy is used to cover these psychological experiences. By contrast, the first formulation does not require the observer entering a particular mental state. Here the process of identifying oneself with the other gives way to the process of deducing what one’s own preferences would be under certain ideal conditions. (Mongin, ib, 160) To decide whether it is better to be rich and old than being poor and young, the observer will use his knowledge of psychological laws and deductive ability. Harsanyi follows the first formulation.

The benefits that Harsanyi hoped to reap from the causal account have, in Mongin ‘s view, exclusively to do with the principle of Uniformity of Extended Preferences (UEP). Mongin ‘s objection to the EUP is not based on his belief that the psychological laws are too imprecise to deliver accurate predictions on individual ‘s dispositions. His objection is connected with the way Harsanyi mixes up preferences and utilities (see note 329 and section 6). He proposes to formulate a well-being interpretation of extended comparisons in order to contrast this with the interpretation of causal factors:

3e ‘\(k\) in social state \(x\) is better off than \(j\) in social state \(y\)’

Essentially, the argument for this conclusion consists in saying that well-being is an objective concept, and, therefore, may be uniform among observers. But we have already rejected this interpretation and it does not fit Harsanyi either,
because he is a preference utilitarian. "It would be nice if comparisons of well-being could be deduced from some universally shared set of preferences, but there is no argument available to provide the desired set of preferences." (Mongin, 2001, 166)

The impossibility of confirming the impartial observer theorem along the causal way has led many writers to elaborate the possibility of complete empathetic identification. "When I put myself in another person’s shoes in complete empathetic identification, I imagine myself in double counterfactual circumstances. I imagine myself first, not in my own actual objective circumstances (occupation, residence, family, income), but in that of the other person. Second, I imagine myself in that situation not as having my own personality, subjective outlook, and response, but with that of the other person. (..) The aim of that manoeuvre is to induce in the investigator some experience or mental state that will ‘represent’ that of the other person. That is, by imagining myself in another’s objective situation with his personality traits and subjective responses, I expect to have an experience just like the experience he is presumably undergoing. I might perform this manoeuvre once on Smith and once on Jones. Since, by hypothesis, these induced experiences replicate those that Smith and Jones have in their actual situation, I can use the resources of my self-knowledge

---

329 Even if we should replace the extended preference functions by extended utility functions, it could not be taken for granted that we could compare two extended utility functions. How do we compare \( U_j(y) \) and \( U_i(x) \). The extent to which preferences are satisfied is the position they have in a preference ranking. We could employ the zero-one rule, in which we assign the value of 1 to the highest ranked preference and the value of 0 to the lowest ranked. But the question is whether \( j \)’s bottom is on par with \( x \)’s bottom. (Hausman, 1995a, 480) (Another complicating factor is that the zero-one rule does not exclude expensive tastes.) The conclusion, therefore, is that individuals who are on the same location in their preference ranking are not necessarily equally well-off. Not only does this conclusion trivialize the comparison, it also leads us to deny that well-being is unproblematic related to the satisfaction of preferences.
to compare their experience”. (MacKay, 1986, 309) MacKay calls this the induced-experience approach.

Another point of view is that I expect that complete empathetic identification "gives me the position to make counterfactual judgements to the effect that I would have such and such experience if I were in Jones's objective situation, with his personality and subjective response. And, knowing that, I can then compare his experience (that I would then have had) with others." (Ib, 310) This time, ‘I wonder what I would feel if... compares with what I would feel if...' MacKay call this the counterfactual judgement approach.

The logical basis of the interpersonal utility comparison along this way is the similarity postulate - the idea that all human beings have much the same basic biological and psychological needs and, therefore, have much the same basic desires for all these valuable things. The similarity postulate reduces the interpersonal to the intra-personal by replacing a multiperson, categorical, comparison by a single person, counterfactual comparison. Does this really solve the problem of interpersonal utility comparison? This postulate says something hypothetical: if I were to satisfy some conditions, I would have a certain reaction. But to imagine satisfying those conditions is not actually satisfying them. Therefore, how do I know I have succeeded in imagining what it is like to be him (her) and be leading his or her life? And, even more urgent, how could I possibly know how rewarding life has been for another person? Any plausible hypothesis must be embedded in a complex psychological picture of the nature and possibilities of human experience and of the causes of differences in human experiences. (Gibbard, 1986) One cannot circumvent these difficulties with the similarity postulate.

Gauthier has raised similar doubts regarding the possibility of complete identification. For Harsanyi's argument to survive, he argues, we must assume that whenever utilities are interpersonally comparable, then $U_a(x) = U_i(a,x)$, for all
persons a and i in all situations x. This need not be so. It holds if and only if the extent to which person’s i’s preferences are fulfilled by being person a in situation x is the same as the extent to which a’s preferences are fulfilled by being in situation x. In evaluating the utility of a, i takes into account not just the preferences a has, and the extent to which they are fulfilled, but his own preferences for having certain personal characteristics, and the extent to which they would be fulfilled. Harsanyi leaves this out of account. (Gauthier, 1986, 242)

Full impartiality requires that one be ignorant, not only of the characteristics one now has, but also of the characteristics of the persons in the social states. One must assume equiprobability and impartiality, that is, one must assume being anybody. But in not knowing who she is, she cannot express a single set of preferences, to be represented by a single utility function. Gauthier claims that the average-utilitarian fails to understand the human individual. To suppose that a person should choose between social states solely by maximizing the average level of preference fulfillment is to suppose, not that he should choose impartially, but that he should exhibit no concern for who he is, that he should care not what capacities and preferences he has, but only that his capacities enable him to fulfill his preferences. A choice is rational only to those able to identify with it, and impartial only if all are able to identify with it. But the choice required by average utilitarianism affords no basis for this identification. Whereas Rawls said that utilitarianism does not take serious the distinction between persons, Gauthier argues "Utilitarianism does not take seriously the individuality of persons." (Gauthier, 1986, 245)

When Arrow suggested solving the problem of the interpersonal comparison of utility by formulating extended preferences he was cautious to formulate some difficulties. When I step in someone’s shoes and the satisfaction that the

\[ U_a(x) \] is person a’s utility for being in state x, \[ U_i(a, x) \] is person i’s utility for being person a in x.

---

330
The requirement that average utility be maximized may be thought equivalent to the idea that the social state is best in which a citizen selected at random is best off. This argument, however, is invalid for there can be two social states with population and aggregate utility equal in both, but where the probability of picking at random a citizen whose utility falls below the average is much greater in the one than in the other. The same average utility is compatible with both a great and a small dispersion of utility. Utilitarianism is focused on the aggregate outcome, and the extent to which this outcome is divided more or less equally among the population is of no or minor importance.

Diamond envisaged a hypothetical society consisting of two individuals, where the government has a choice between two alternative policies. One policy would yield the utility vector (1,0) with certainty. The second would yield the two utility vectors (1,0) and (0,1) with equal probabilities. According to the axioms of Bayesian decision theory the two policies should be assigned the same value. But, in Diamonds
opinion, the second is socially preferable because it would yield both individuals a fair chance. (Diamond, 1967) Harsanyi rejected Diamond's argument for logical and moral reasons. First, it would conflict with Bayesian's postulates, and therefore, conflict with rationality. But Diamond questioned whether the Bayesian postulates could be transferred from individual to social choices. In particular he doubted whether the independence axiom is valid in the case of social choice. Harsanyi next argued that Diamond is mistaken, because it would mean that the utility of a lottery ticket does not only depend on its expected value, but also on some measure of risk, i.e., some measure of dispersion among its possible utility outcomes. This would be an inadmissible double counting for the risk-aversion of a person is already included in the utilities he assigns to various possible lotteries. We observe that Harsanyi suddenly moves the discussion from a concern with (in)equality to a concern with risk-aversion. A move that does not come as a surprise because the vNM functions deal with choice-behavior in relation with lotteries and in this context there is a relation between risk-acceptance and the attitude towards inequality.

331 The question is not whether the independence axiom makes sense, but whether it can be regarded as a universal rule (as Harsanyi believes). This is questioned by Diamond in the case of ethical matters (matters of fairness). I believe that the arguments in favour of the independence axiom in these matters do not hold against the objections (see chapter III and XVIII).

332 The relation between risk-aversion and (in)equality can be elucidated as follows. Suppose there are two prospects A and B that deliver two equiprobable events I and II.

<table>
<thead>
<tr>
<th>event I</th>
<th>event II</th>
</tr>
</thead>
<tbody>
<tr>
<td>prospect A</td>
<td>1,1</td>
</tr>
<tr>
<td>prospect B</td>
<td>0,2</td>
</tr>
</tbody>
</table>

The numbers that indicate the wealth that will result from a particular prospect if a particular event is shown, first for person 1 then for person 2. Both prospects deliver the same expected total value. Someone who subjectively thinks that the difference between 1 and 0 is greater than the difference between 2 and 1, will choose prospect A. This person is risk-averse. Someone who is concerned about equality would make the same
To do justice to Diamond’s argument, however, we better discuss it in terms of fairness, for instance, in the context of social preferences (in Pattanaik’s sense). (The following argument is adapted from Broome, 1989). Consider the following prospects:

<table>
<thead>
<tr>
<th>event 1</th>
<th>event 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>prospect A</td>
<td>2,2</td>
</tr>
<tr>
<td>prospect B</td>
<td>3,0</td>
</tr>
</tbody>
</table>

Person 1 attaches a probability of 0.7 to event 1 and 0.3 to event 2. Person 2 attaches a probability of 0.3 to event 1 and 0.7 to event 2. For both persons the expected utility in prospect B is 2,1 and consequently both prefer B to A. The Paretian condition demands that also the social preference should indicate the preference B > A.

What does a social preference exactly mean? In this case it is linked with the aggregation of individual preferences. "B is socially preferred to A" means that both persons individually think that B is better for him/her than A. Alternatively, a social preference is linked with social choice. "B is better than A" means that if the two persons had to vote for a social policy, they would choose one that would have as outcome prospect B. In this interpretation the statement "B is socially preferred to A" does not mean that B is better than A, but that it should come about. This is the democratic argument. (Broome, 1989, 10) Now, assume that in the democratic choice. Harsanyi assumes that all persons have the same attitude to risk, i.e., they are risk-neutral.

There are three distinct concepts: personal preferences of the individual, 'moral' preferences of the individual (Harsanyi's social preferences, based on an individual’s value judgements concerning social welfare) and Arrow's social (collective) preferences. Arrow derived his social preference (choice) directly from the personal, subjective preferences. Pattanaik’s suggestion, that is followed here, is that they should be derived from the personal 'moral' preferences. (Pattanaik (1967)) It is, of course, assumed that personal 'moral' preferences differ. Thus, the similarity postulate is repudiated.
argument much weight is attached to an equal distribution of wealth, and that A is chosen instead of B. This is a contradiction, for it would imply that the persons do not follow their individual preferences when they vote for a social policy. How could this contradiction be solved? To begin with, it must be acknowledge that, when both persons decide individually, they will choose B. But, though they both prefer B to A, it cannot be better for both. Fortune decides who will be the winner and who will be the loser. However, they could show some consideration for each other and try to judge impartially what is best for the two of them. Thus, impartial considerations could turn the balance.

Of course, the case I have presented here in defense of Diamond's point of view is entirely hypothetical (and does not represent Broome's view). People could just as well base their votes on their self-interested subjective preferences. All I wish to emphasize is, that from a moral point of view one could easily sympathize with people who decide to reflect on moral constraints and act accordingly.

Sen also argued with Harsanyi on the question whether the dispersion of utility should be taken into account when measuring social welfare. Sen would make social welfare depend, not only on the mean, but also on some measure of inequality, i.e., of dispersion among the different individuals' utility levels. This means that the social welfare function should be a non-linear function of the individuals' utilities. (Sen, 1973, 18) Harsanyi's counter argument is that the dispersion argument would be perfectly valid if it referred to money values of the possible outcomes instead of their utilities. It makes good sense to assume a decreasing marginal utility for money, but it would make no sense whatever to assume a law of decreasing marginal utility for utility. Harsanyi regards Sen's dispersion argument as "an illegitimate transfer of a mathematical argument from money amounts, for which it does hold, to utility levels, for which it does not hold". (Harsanyi, 1977, 321) The basic reason why the utility dispersion argument is invalid is similar to the
reason he advanced against Diamond, namely that it would involve double counting because the individual utility functions already reflect a concern for inequality (in terms of diminishing marginal utility). The social welfare function should treat different individuals' utility functions in a similar manner, and should assign the same weight to each of them.\footnote{Sen's first reaction was that the utility-dispersion argument, to which Harsanyi made extensive reference was, in that form, a figment of Harsanyi's imagination. He never made any proposal to define a non-linear welfare function on vNM functions. (Sen, 1977b, 279) In a second reaction (1977c), Sen questioned Harsanyi's justification for the linearity of the social welfare function. What Sen brings up for discussion is what the \( W_i \) values exactly are. In his opinion they are mere accounting values for predicting choice under uncertainty. They need not coincide with any concept of utility that has an independent meaning, such as happiness or preference-fulfilment. The numbering system \( W_i \) works the way it does precisely by incorporating attitudes to risk in the \( W_i \) values themselves. The fact that \( W \) is the sum of the \( W_i \) values does not imply sum ranking of the utility values \( U_i \). This procedure does not yield any evaluative conclusions concerning social states as such, only linearity. Because "(..) the sum-of-utilities theorem is really a theorem about the representability of social preferences, and not directly about the proper way to evaluate social states, (..)." (McClellan, 1981, 601)}

The justification of a linear social welfare function is also based on the impartiality argument. People have to choose between social states without knowing their own position in them, assuming equiprobability. Sen admits that, if in this 'as if' position, people are assumed to behave according to the axioms of expected utility theory, then they must end up

\footnote{Nunan thinks that Harsanyi "makes a serious mistake by substituting an empirical law of human psychology for a moral principle". (Nunan, 1981, 589)}
maximizing expected utilities from the different lotteries, and since the prizes are being particular individuals this would amount to maximizing the mean utility of all individuals taken together. (Sen, 1977c, 300) But even if the independence axiom is accepted and so the entire vNM framework for social choice involving risks and even if one believes that the utilitarian rule is acceptable, what is its force? In the view of Sen the linear form it yields is trivial: \[ W = \sum_i (1/n)W_i \] follows from the fact that \( W \) is the utility of the lottery and \( W_i \)'s are the prizes with equal probability. But this linear form asserts very little, since \( W_i \) is simply the value of the lottery of being person \( i \) with \( 1/n \) probability in state \( x \). The result is quite non-controversial. (..) But its non-controversiality arises essentially from its non-assertion'. (Sen, 1977c, 300/1)

6. Harsanyi and his critics

Harsanyi's construction of a social welfare function has been criticized for very different reasons. In Rawls' theory of justice the alternatives to choose between are not social states but alternative conceptions of justice to regulate the basic structure of society. All that is known in the original position is that each of the conceptions of justice available to the parties has consequences superior to general egoism. Rawls emphasizes that "Neither our situation in other societies, nor in a state of nature, have any role in assessing conceptions of justice." (Rawls, 1993, 279) Rawls, moreover, rejects the use of the Bayesian decision theory in the original position. The use of probabilities in the original position is pointless for two reasons. First, because it is characterized by radical uncertainty; not only are the parties unable to conjecture the likelihood of the various possible circumstances, they cannot say much about what the possible circumstances (social situations) are. Consequently, they have no knowledge over alternative social systems at all. Second, there is no objective ground for assuming that one has an equal chance of turning out to be anybody. He also rejects the use of the vNM-utility
function, because it basically expresses people's attitudes toward risk-taking, i.e., toward gambling, and these attitudes have no moral significance. Moreover, it is inappropriate to equate individual choice with the "choice" of a social ethic. Furthermore Rawls objects to the way in which the problem of interpersonal utility comparison is 'solved', namely, by assuming that individuals have similar preferences. (see Rawls, 1971, 167ff)

Gauthier argues that only a non-utilitarian ethical theory is compatible with the general theory of rational behavior. It is generally impossible to satisfy two maximization requirements -the maximization of individual utility and the maximization of welfare- simultaneously, as the prisoner's dilemma illustrates. Therefore, if ethics is part of the general theory of rational behavior, then it must be shown that it is rational, at least in some contexts, to constrain preferences. (Gauthier, 1982, 148) Gauthier also rejects Harsanyi's claim that an impartial choice is necessarily a choice based on the average utility principle. In choosing among prospects without knowing what one's characteristics and circumstances will be, one rationally maximizes not the average utility received by the various persons in their situations, but the average of one's own utilities for being each of those persons. Impartial choice can only be rescued by violating individuality. (Gauthier, 1982, 159) And he concludes, that a choice reflecting no point of view, but which awards equiprobability to each point of view, has lost both ethical and rational appeal.

Harsanyi is very confident that the utilitarian rule of choosing that social state, of those available, which maximizes the preferences, in sum, of all individuals, considered impartially, is the only rule that is justifiable. And doesn't he have two kinds of justifications: a logical and a moral one? The first is developed in his Aggregation Theorem and the second in his Impartiality Theorem. But are his arguments as valid as Harsanyi thinks they are? Harsanyi equates 'not knowing' (judgements made in a situation of hypothetical ignorance) both with impartiality and with
‘equal chance’. How does he proceed from here to the utilitarian rule? There are two steps in the argument: one from ignorance (not knowing) to equal chance and the other from equal chance to impartiality. The first step amounts to the application of Laplace ‘s principle of insufficient reason. Complete ignorance should be modelled as equiprobability. But Bayesian writers have always been divided on the significance of Laplace ‘s principle. There are plausible alternatives within this doctrine, which replace ‘equal chances’ with the assumption of subjective probabilities, one for each observer. (see Mongin, 2001). In this alternative approach impartiality is understood epistemically as complete ignorance. Once this line is taken it is quite naturally that observer i assesses social states in terms of some probabilistic measure p(i). The conclusion is that “a genuinely utilitarian formula, that is with equal weights and no observer dependence is out of reach..” (Mongin, 2001, 174) The observer’s dependence follows from the observer ‘s reliance on subjective probabilities and the weights are unequal because they are determined by these unequal probabilities. When unequal probabilities are accepted, more arguments become available which throw doubts on

335 In Harsanyi’s philosophy ignorance and risk are important ingredients of a moral position. The statement ‘slavery is good’ would be a moral judgement when the speaker says it even though he does not know what his position is going to be (because his statement is irrespective of his own position, it is universalizable), or to be more precise, even though he knows, so to speak, that he has an equal chance of being slave or slave holder. Risk and universalizability together give us Harsanyi ‘s concept of an impersonal, impartial preference. (see Pattanaik, 1967)

336 This principle says that equal probabilities should be assigned to different events unless some reason exists to suppose that one is more likely than the other. Harsanyi reasons that in the original position the veil of ignorance conceals all information about each position and so Laplace’s principle tells us to assign equiprobabilities to every event. Significant paradoxes arise only when Laplace’s principle is applied to situations that make it necessary to evaluate an infinite number of possible states of the world
Harsanyi’s enterprise. Sen has raised doubts concerning the question whether Harsanyi’s theory can be called a utilitarian theory at all. This criticism is based on two distinct objections. First, in Harsanyi’s theorems utility is only used to represent preferences and has no other independent basis. Therefore Harsanyi's theorems can only be regarded as social preference representation theorems. Second, when an independent concept of utility is used, utility functions may not be vNM functions, even if the expected utility axioms are satisfied and social welfare functions could be non-linear and hence, obviously, non-utilitarian.

If utility and preferences have an independent basis (utility is for instance measured as a mental state -happiness-, or as desire fulfilment), then the preference-ordering Ri and the utility function Ui may be congruent. When utility does not have a meaning independent of preferences, then utility is the derived concept and then Sen is right in saying that the utility function is merely a representation of a preference ordering. And then the numbers generated by a vNM function aren't' a measure of anything other than preferences and certainly not of something called ‘utility’ that could be identified with ‘welfare’ which is supposed to underlie the preferences. Then utility is only an ordinally measurable unit and noncomparable between individuals. And neither Harsanyi’s aggregation theorem, nor his impartial observer theorem could stand the test. Furthermore, with this concept of utility

337 Has Mongin undermined Harsanyi’s moral justification? When the assumption of equiprobability is suspended, then also the logical justification comes under attack. Broome argues that the Aggregation Theorem is only feasible if the vNM conditions and the Pareto conditions on social preferences are mutually consistent and they are only consistent when the individuals agree on the probability of every social state. (Broome, 1991b) The Aggregation Theorem will also not be feasible when people differ in risk attitudes. When people differ in risk attitudes not only will their ordering of uncertain prospects differ, but also their ethical preferences will not be the same and social welfare functions need not to be linear.
it is arbitrary to consider only vNM utility functions. If, on the other hand, utility does not simply represent preference, then Harsanyi's impartial observer theorem can be used to justify the axiomatization of utilitarianism, provided: 1) well-being is cardinally measurable and fully comparable and 2) each person's well-being, including that of the impartial observer, is measured by vNM utility functions and 3) the principle of welfare identity is satisfied. (Weymark, 1991, 310)

What does this third condition entail? As we have seen, the impartial observer in Harsanyi's theory could in fact be one of the individuals in society who has temporarily adopted an impartial perspective from which to judge the relative desirability of alternative outcomes. Harsanyi is in effect saying that all interpersonal comparisons can be reduced to intra-personal comparisons. He regards everyone as fundamentally the same (the similarity postulate), and thus any individual could in principle determine the well-being of any other individual by simply imagining how well-off person i would be with person's j alternative (lottery) and j's causal variables. This is the "principle of welfare identity". This principle is easy to satisfy, for according to Harsanyi, "..., it amounts to much the same thing as assessing the utilities I would derive from various commodity's baskets if

Harsanyi believes that: "utility is merely a convenient measure of a person's preferences but does not explain his preferences. It is rather the other way around: a person's preferences explain and determine his utilities." (Harsanyi, 1987, 340)

Harsanyi holds to the vNM representations because they preserve the risk-attitudes of the decision-maker. He believes that there is a strong connection between the amount of risk someone is willing to take in order to possibly receive x, and his preference for x. (Harsanyi, 1982, 52/3) The more higher-ranked preferences are realized, the more satisfied an individual is with his state of affairs. In this way he still connects preferences with 'utility'and well-being. I think this a legitimate argument when he is prepared to admit that a vNM utility is more than a number and that it also serves as an emotional index. But this is just what vNM indexes are not supposed to be.
my personal situation and personal characteristics underwent a specific change." (Harsanyi, 1977, 327) But Harsanyi has never provided satisfactory justifications of these critical assumptions. He restricts his attention to vNM utility functions because he mistakenly assumes that their use is required by the adoption of the axioms of expected utility theory.

Following Weymark the debate can be summarized as follows: if utility has no independent basis Sen is right in denying Harsanyi’s theorems a utilitarian status. "Furthermore, no significance should be attached to the linearity or non-linearity of the social welfare functions, as the curvature of this function depends solely on whether or not vNM representations are used, and the use of these representations is arbitrary. If utility does have an independent status, Harsanyi’s impartial observer argument can be interpreted as an axiomatization of utilitarianism, provided certain conditions are fulfilled." (Weymark, ib, 315) The similarity postulate and the principle of welfare identity trivialize however this result.

7. Conclusion

Social preferences are, in contrast to individual preferences, pure theoretical constructs, as I hope this chapter has made clear. The problematic aspect, even as a theoretical construct, is the interpersonal comparison of utilities. Harsanyi circumvents this problem by equating the intra-personal comparison and the inter-personal comparison. He feels supported in the permissibility of this item of his construction by using the construction of an "original position" in which individuals have to choose between alternative social systems from behind "a veil of ignorance". 339 They know the whole

339 Though it seems that the utilitarian ethics puts the individual central, it is a mistake to believe so. Utilitarianism is concerned with states of affairs, i.e., with aggregate welfare. In the political domain the utilitarian view is impersonal. Most utilitarians believe that benefits and burdens should be
history of the world, but they do not know what their personal position will be under either system. They would have the same probability of taking the place of the worst-off or of the best-off position or any position in between: this is the equiprobability assumption. He asks himself: ‘which decision rule would rational individuals use in this model when they would apply a moral point of view.’ His answer is that each individual will act as choosing in circumstances of uncertainty and that she shall evaluate each system in terms of its institutional arrangements and each arrangement in terms of the arithmetic mean of the vNM utility function of each individual. This is the utilitarian principle. The proof of this principle depends heavily on complicated concepts as universal extended preferences and complete empathetic identification and auxiliary assumptions of which the similarity assumption or the assumption of welfare identity is the most crucial. This assumption does not stand the test of criticism. This means that interpersonal comparisons of well-being are not value free. And when they are not value free, different weights can be attached to different individual utility functions. Exit the utilitarian principle?

Of course, an outside observer with a utilitarian character still could decide to give equal weights to all utility functions. Which moral (and non-moral) considerations would make him do that? What moral principles might require of any person is to do for each person affected by his actions what he wishes were done for him in the hypothetical circumstances that he was precisely in their situation. Hare argues that this way of putting the matter emphasizes that he has to give the

freely weighted against each other, even if they come to different people. Utilitarianism merges individual welfare into overall totals or averages and takes the improvement of these totals or averages as desirable quite apart from the decisions of any individual that it is. In short: utilitarianism is not concerned with people but with states of affairs. The right act maximizes the good. The goal is not to respect people but rather to respect the good to which people may or may not be useful contributors. (see chapter VI)
same weight to everybody’s interests, including his own. It may be surprising to find that a Kantian rule has to ground utilitarian ethics, but this is part of Hare’s defence of utilitarianism. Hare stipulates that in moral reasoning we are required to prescribe universally for cases of a given sort. The utilitarian doctrine prescribes that we choose that alternative, of those available to us, which maximizes the preference-satisfaction, in sum, of all those affected by our action, considered impartially. Hare believes that this is the only categorical imperative in ethics. (Hare, 1989, 108) This is so because only by giving equal positive weight to the equal preferences of all individuals can we find universal prescriptions that are most acceptable to us. Kymlicka does not think that the idea of equal concern automatically ends in maximizing good. We may as well show our concern by guaranteeing each an adequate level of resources and liberties. Or we might do what is best for the least well off. (Kymlicka, 1990) In the next chapter I will show that Rawls on Kantian grounds formulates such a general rule.