Nr. 21

John C. Harsanyi

Nonlinear Social Welfare Functions

or

Do Welfare Economists Have a Special Exemption from Bayesian Rationality ?

July 1974

Nonlinear Social Welfare Functions
or
Do Welfare Economists have a Special Exemption From Bayesian Rationality?

by

John C. Harsanyi
University of California, Berkeley,
and
University of Bielefeld

Abstract.

It is argued that Bayesian decision theory is a solution of an important philosophical problem, viz. the problem of how to define rational behavior under risk and uncertainty. The author has shown in earlier papers that if we take the Bayesian rationality postulates seriously, and take an individualistic point of view about social welfare, then our social welfare function must be a linear function of individual utilities: indeed, it must be their arithmetic mean. The paper criticizes Diamond's contention that one of the Bayesian postulates (viz. the sure-thing principle) does not apply to social decisions, even though it does apply to individual decisions. It also criticizes Sen's proposal of making social welfare a nonlinear concave or quasi-concave function of individual utilities. The social welfare function proposed by the author depends on interpersonal utility comparisons. The use of such comparisons is defended. It is also argued that anybody who feels that the utilitarian(i.e., linear) form of the social welfare function is not egalitarian enough, should reject the author's individualism axiom,instead of trying to reject the Bayesian rationality axioms. However, this would be equivalent to giving egalitarian considerations a priority in many cases over humanitarian considerations. Finally, the paper discusses the reasons why even full agreement on the mathematical form of the social welfare function would not give rise to a utopian state of moral consensus: moral controversies arising from disagreements about what predictions to make about future empirical facts would still remain.

SYNOPSIS

1. Introduction.

Besides problems of empirical fact, and of formal (logical and mathematical) validity, a third important class of theoretical problems are problems of finding a rigorous scientific concept α (usually, but not necessarily, a concept defined by formal axioms) as a possible replacement for a vague and unclear concept β of prescientific commonsensical discourse [Carnap, 1950, pp. 3-8].[1/] I shall describe problems of this kind as conceptual or philosophical problems. They play a major role in philosophical discussions. But they also arise in economics and in the other social sciences. Examples are the problems of how to define rational behavior under certainty, risk, and uncertainty, as well as in game situations; how to define the concept of public interest - -, or, more exactly, how to define a social welfare function that would adequately formalize our intuitive notion of public interest; how to define social power, or social status, etc. (As everybody familiar with modern physics knows, philosophical problems of this type also arise in the natural sciences, though their substantive focus will be obviously different.)

The Bayesin theory of rational behavior under risk and uncertainty is one of the few cases where such an essentially philosophical problem has found a very specific and unambiguous solution, accepted as correct by an increasing convergence of expert opinion. (The problem of how to define rational behavior under certainty has been solved already by classical economic theory, whereas the problem of what the precise definition of rational behavior should be in game situations is still largely an open question, though perhaps some progress is now being made towards its solution [see, e.g., Harsanyi, 1974; cf. also Harsanyi, 1966] .

I propose here to briefly summarize the arguments in support of the Bayesian position: 1. The axioms (rationality postulates) of Bayesian theory are intellectually very compelling for anybody who has taken the trouble of properly understanding them; and no argument has yet been proposed that would cast serious doubt on these axioms as sensible criteria for rational behavior. 2/ 2. Bayesian theory does not seem to lead to any counterintuitive implications. 3. The main rival definitions of rational behavior under risk and uncertainty are known to have highly counterintuitive implications [Radner and Marschak,1954].

In two earlier papers [Harsanyi, 1953, and 1955] , I have argued that, if we take the rationality postulates of Bayesian theory seriously, then we can obtain a clear and unambiguous solution also for the time-honored philosophical problem of defining an adequate social welfare function. In fact, it can be shown that the social welfare function must be a linear function of all individual utilities - - or, more exactly, it must be defined as the arithmetic mean of the utility levels of all individuals in the society.

In recent years my theory has been quoted with approval by some distinguished economists [see, e.g., Theil, 1968, p.336]. But it has also come under criticism [Diamond, 1967; Rawls,1971; Sen, 1970 and 1973].[3/] The most specific criticism was Diamond's, who at least clearly recognized that my theory can be rejected only if one rejects one or more of its axioms. (Sen, as well as some other advocates of nonlinear social welfare functions, have never made it clear which particular axiom(s) of my theory, if any, they wish to deny.) Diamond himself has chosen to reject the sure-thing principle as applied to social decisions. Below, I shall consider Diamond's argument in some detail. I shall try to show that 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; and I shall try to outline some of the curious implications of Diamond's point of view.

I shall also argue that Sen's proposal [1973, pp. 20 and 52] of making our social welfare function dependent, not only on the mean value of individual utilities, but also on their variance, is open to the same objections as the view that the utility of a lottery ticket should depend, not only on its expected utility, but also on its utility variance.

It is, of course, clear enough why some economists are unhappy with the utilitarian theory entailed by using a linear social welfare function. While Robbins [1938] once objected to utilitarianism because apparently he felt that it would have all too egalitarian implications, in our own age most objections are likely to come from economists who find utilitarianism *not* to be egalitarian *enough*. I shall argue that an economist who takes the latter position should not try to temper with the rationality postulates of Bayesian theory: this could lead only to highly counterintuitiv results. Instead, he should reject the *individualism postulate* of my theory [Harsanyi, 1955, Postulate c on p. 313], which makes social preferences fully determined by individual preferences.

This means that anybody who wants to adopt a moral position more egalitarian than the utilitarian position already is, must admit that the well-being of the individual members of society is *not* his ultimate moral value, and thathe *is* willing in certain cases to sacrifice humanitarian considerations to egalitarian objectives when there is a conflict between the two. This is the real moral issue here; and this important moral issue should not be obscured by superficially attractive, but really quite untenable, objections to the sure-thing principle or to any other rationality axiom.

2. Welfare economics and the sure-thing principle.

For our purposes it is sufficient to consider the sure-thing principle as stated for *risky* situations (where all probabilities are known to the decision maker). In this case the principle asserts that, other things being equal, it is always preferable to have a chance of winning a *more highly valued prize* with a given positive probability, to having the chance of winning a less highly valued prize with the same probability. A variant of this principle is the substitution principle: it makes no difference whether one has the chance of winning one prize or another with a given probability, if one regards these two prizes as being equally valuable. Both forms of the sure-thing principle are intellectually highly compelling requirements for rational behavior, and it is very hard to envisage any situation in which a rational individual could feel justified in violating either of them.

Diamond admits that the sure-thing principle is a sensible rule for individual choice behavior, but denies it any validity for social choices. Even on the face of it, it would be very surprising if this view were correct. Surely , when we act on behalf of other people, let alone, when we act on behalf of society as a whole, we are under an obligation to follow, if anything, higher standards of rationality than when we are dealing with our own private affairs. If common prudence requires private individuals to follow the sure-thing principle, then government officials who are supposed to look after our common interests can hardly be absolved from doing the same. Nor can welfare economists reasonably advise these public officials against doing so. Of course, this is not a conclusive argument: after all, there could be some very special reasons why public officials and welfare economists would not be bound by the Bayesian standards of rationality. In order to draw firmer conclusions, we have to consider Diamond's argument in more specific terms.

Diamond envisages 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 other would yield the two utility vectors (1,0) and (0,1) with equal probabilities. Assuming that society would attach equal weight to the interests of the two individuals, it is easy to see that, according to the sure-thing principle (substitution principle), the two policies should be assigned the same value from a social point of view. But, in Diamond's own opinion, the second policy is in fact socially strictly preferable to the first because it would yield both individuals a "fair chance" while the first would not. According to Diamond this shows that the sure-thing principle has no validity for social decisions.

In order to gain a better understanding of Diamond's argument, I propose to apply it to a couple of more specific hypothetical situations. For example, let us imagine two societies, A and B. Society A has an extremely unequal income distribution, so extreme in fact that even politically rather conservative observers find it absolutely revolting. Moreover, it has virtually no social mobility, and certainly no mobility based on what could be described as individual merit. Society B is exactly like society A, except for the following difference. By old social custom, all babies born in B during any given calender month are randomly redistributed by government officials among all families who had a baby during that period, so that every baby born in that month will have the same chance of ending up in any given family.

(I shall assume that all families fully accept this social custom, and treat the babies randomly allocated to them completely as their own.)

Should we now say that society B would be morally less objectionable than society A, because in B all individuals would have a "fair chance" of ending up in a rich family and, therefore, in a privileged social and economic position? By assumption, B is a society with an income distribution just as unfair as A is. In both societies, any individual's social and economic position has nothing to do with personal merit, but rather is completely a matter of "luck". In A it depends wholly on the accident of birth - - on the "great lottery of life" which decides who is born into what particular family. In contrast, in B it depends wholly on a government-conducted lottery. Why should we assign higher moral dignity to a lottery organized by government bureaucrats than we assign to the "great lottery of life" which chooses a family for each of us without the benefit of government intervention? Why should a bureaucratic lottery be regarded as being a "fairer" allocative mechanism than the great biological lottery produced by nature?

Indeed, suppose we would obtain reliable information to the effect that the families we are born into are always chosen literally by a huge heavenly lottery. Can anybody seriously assert that this metaphysical information would make the slightest difference to our moral condemnation of hereditary social and economic inequalities?

Next, let us consider another example. Suppose that the government has a choice between two policies. The first policy would consist in abolishing an obsolete and, by now, economically very harmful protective tariff. This would benefit all citizens (many of them quite substantially), except for a small group of workers and employers in the hitherto protected industry, who would suffer moderate economic losses. The second policy would result in the same vector of individual utilities as the first, except that the individual components of this vector would be randomly permuted, because now the gainers and the losers would be chosen by a government-conducted lottery. To fix our ideas, we shall assume that the second policy would actually consist in implementing the first policy (i.e., removal of the protective tariff), followed by a random redistribution of income on the basis of a lottery.

Once more, would it make any sense to assert that the second policy would be morally preferable to the first? Under the first policy, the losers would be the members of one particular industry, who presumably have entered this industry by family association or by other accidents of personal life history. Thus, being a member of the loser group would be just as much a matter of personal "bad luck" as would be under the second policy, where the losers would be selected literally by a lottery. Again, what would make such a lottery a morally superior allocating mechanism to those historical accidents which make people enter a particular industry?

In fact, if it cannot be avoided (e.g., by compensating the losers) that some people should suffer net losses as a result of an otherwise desirable governmental policy, then, it seems to me, it is surely fairer to "let the chips fall where they may" - - instead of trying to reallocate these losses in a wholly arbitrary manner.

It would be easy to adduce many more examples to corroborate the same conclusion. Diamond's suggestion that economic and social privileges allocated by government policies with random components are morally more acceptable than social and economic privileges allocated by the accidents of birth and of personal life history - - i.e., the suggestion that the first kind of personal "luck" is morally superior to the second kind - - is wholly without merit. Therefore, his claim that social choices are not subject to the sure-thing principle falls to the ground.

Let me add that the same conclusion applies to the allocation of biological qualities, such as intelligence, scientific and artistic talent, health, beauty, physical strength, etc. Under appropriate safeguards, techniques of genetic engineering could no doubt benefit mankind by improving the biological endowment of the average individual of future generations. But a mere redistribution of the same unimproved biological qualities by means of an official lottery (what a horrible thought!), even if it were technologically feasible, would certainly not represent any improvement over the existing situation from a moral point of view.

3. Mean utility vs. utility variance.

Whereas my own theory would make the social welfare function the arithmetic mean of individual utilities, Rawls [1958 and 1971] has proposed a social welfare function based on the maximin principle, and always measuring the welfare level of society by the utility level of the worst-off individual. This means mathematically that Rawls's social welfare function would always assign infinitely more weight to the interests of the poorest, or otherwise least fortunate, members of society than it assigns to the richer, or otherwise more fortunate, members. In contrast, my own social welfare function would always assign the same weight to equally urgent needs of different individuals, regardless of their social or economic positions. But of course since, typically, poor people have many more unfilled urgent wants than rich people do, in practice in most cases my own social welfare function will lead to similar policy decisions to Rawls's, because it will give much higher priority to poor people's needs. Only in those, rather special, cases where some rich people may have even more urgent needs than some poor people do, will my social welfare function lead to opposite policy decisions. However, precisely in these cases I feel that my social welfare function would lead to the morally right decisions while Rawls's would lead to morally wrong ones.

For example, if a philanthropist has to decide whether to give $ 100 in cash, or the same value in food, or in clothing, etc., to a poor man or to a millionaire, it is clear that he should choose the poor man since the former will have a much greater need for extra cash, or for extra food, or for extra clothing, etc. But if he has to choose between giving a life-saving drug in short supply to a poor man or to a millionaire, then the only relevant consideration must be who needs it more, i.e., who could derive the greater medical benefit from it. Surely, it would be highly immoral discrimination against the millionaire to refuse him a life-saving drug, even though he has the best claim to it from a medical point of view, merely on the ground that he happens to be a millionaire. Yet, Rawls's social welfare function would force us precisely to engage in such immoral discriminatory practices in some, rather special, but perhaps not — all — too — rare, cases. 4/.

In his interesting book On Economic Inequality, Sen [1973] proposes an intermediate view between the two "extreme" positions taken by Rawls and by myself. He feels that poor people's interests should be given more weight, but only a finite number of times more weight, than rich people's interests should. To accomplish this, he suggests that the social welfare function should be a concave, or at least a quasi-concave, function of individual utilities (pp. 20 and 52). Whereas my own theory would make social welfare depend only on the mean value of the different individuals' utility levels, Sen's theory would make it depend also on some measure of inequality (dispersion) among different people's utility levels, such as the variance of individual utilities.

Unfortunately, it is not always the case that the truth lies somewhere in the middle between two extreme positions. Sometimes, it will actually lie with one of these two. Finding a middle ground may be the key to good international diplomacy, but it may not be the most effective way of finding the best solution to a theoretical problem. In fact, Sen's theory would give rise to unfair discrimination against people enjoying relatively high utility levels, much the same way as (though less often than) Rawls's theory does.

Moreover, Sen's theory shows far-reaching formal similarities to what I shall call the lottery-variance argument, viz. to the view that the utility assigned to a lottery ticket should depend, not only on its expected utility, but also on its utility variance(and perhaps also on the higher moments of the probability distribution over possible utility outcomes). More particularly, so the argument runs, a decision maker averse to risk-taking should assign a lottery ticket a utility below its expected utility - - making the difference an increasing function of the variance in possible utility outcomes when other things are kept constant. Therefore, for such a decision maker, the utility of a lottery ticket cannot be a linear function of the utilities of the various prizes if all probabilities are kept constant; rather it must be a concave or a quasi-concave function of these utilities.

Yet, the lottery-variance argument is known to be mistaken [see Luce and Raiffe, 1957, p.32, "Fallacy 2"; or see any textbook on decision theory]. To be sure, a similar argument would be correct if all references to the utilities of the various prizes were replaced by references to their money values. Let L be a lottery ticket yielding $ x as expected money gain. Then, it will be no doubt true that a decision maker with risk aversion will assign L a utility below the utility of obtaining $ x with certainty. Moreover, other things being equal, the difference between these two utilities will tend to be an increasing function of the variance in the money values of the various prizes.5/ It is equally true that, for such a decision maker, the utility of a lottery ticket cannot be a linear function of the money values of the various prizes when all probabilities are kept constant; rather it must be a concave function of the latter.

Why can this argument not be extended from the money values of the various prizes to their utilities? The basic reason is that the decision maker's von Neumann-Morgenstern utility function already makes an appropriate allowance for his attitude towards risk-taking. Thus, if, e.g., he has a negative attitude towards risk, then this fact will already be fully reflected in the utilities he assigns to the various prizes and, therefore, also in the expected utility associated with the lottery ticket.

Accordingly, it would represent unnecessary double counting if we made an allowance for the decision maker's risk aversion for a second time, and made his utility for the lottery ticket dependent on the variance in utility, or if we made this utility a nonlinear concave or quasi-concave function of the utilities of the various prizes. Indeed, such a procedure would be not only _unnecessary_; it would also be logically _inadmissible_. For, it can be shown that, if the decision maker follows the rationality postulates of Bayesian theory, then he _must_ assign a lottery ticket a utility equal to its expected utility (assuming that his utility is measured in von Neumann-Morgenstern utility units). This means that he simply _cannot_ make this utility a function of the variance in utility, or a nonlinear function of the utilities of the various prizes.

I now propose to show that Sen's theory succumbs to the same objection as the lottery-variance argument does; it is an illegitimate transfer of a mathematical relationship from money amounts, for which it does hold, to utilities, for which it does not hold. It is certainly true that social welfare cannot be equated with (average) real income per head, and cannot be even a function of the latter variable _alone_. Rather, if real income per head is kept constant, then, other things being equal, social welfare will tend to increase with a more equal distribution of income. This means that social welfare is a nonlinear concave or quasi-concave function of individual incomes.

This is so because, if we can costlessly redistribute income, and can transfer $ 100 from a rich man to a poor man, then presumably the damage done to the former will be considerably less than the benefit accruing to the latter. This follows from the law of decreasing marginal utility for money for each individual [6/] (together with the assumption that different individuals will have reasonably similar utility functions for money, so that the law of decreasing marginal utility will affect them in a similar way).

Yet, even if social welfare is a nonlinear concave or quasi-concave function of individual _incomes_, it does not follow at all that it is also a similar function of individual _utilities_. Of course, it makes no sense to discuss how much weight the social welfare function should give to utility increments or decrements for different individuals, unless we are willing to assume that such utility increments or decrements for different individuals are _comparable:_ otherwise these weights cannot be defined at all in an unambigous way. [7/] I shall come back to the problem of interpersonal utility comparisons below (Section 4).

Once, however, we assume the possibility of interpersonal comparisons, then it becomes immediately clear that the argument establishing the concavity or quasi-concavity of the social welfare function in individual incomes does not carry over at all to individual utilities. If we decrease a rich man's utility level by 100 utility units and simultaneously increase a poor man's utility level by 100 utility units, then (assuming that we have measured utility in equal units in both cases) the utility loss suffered by the former will be exactly the _same_ as the utility gain accruing to the latter. It makes good sense to assume a law of decreasing marginal utility for _money_ (or for commodities); but it would make no sense whatever to assume a law of decreasing marginal utility for _utility_. (It would be surely nonsensical to assert that a utility decrease from 1,000,100 units to 1,000,000 units of utility is a "smaller" utility change than a utility decrease from 300 units to 200 units. By describing both as 100-unit changes we are automatically committed to assuming their equality.)

What is at stake here, of course, is not a (rather trivial) point in simple mathematics. Rather, as I have already argued in discussing Rawls's theory, we are dealing with an important moral issue. When we are assigning the same quantitative measure to utility changes affecting two different individuals (e.g., when we call both of them 100-unit changes), then we are implicitly asserting that these utility changes for both individuals involve human needs of equal urgency. But, this being so, it would be highly unfair discrimination to claim that, as a matter of principle, satisfaction of one man's needs should have lower moral priority than satisfaction of the other's. Though I have already illustrated this point by an example (involving a life-saving drug), I shall propose a second example in order to consider some other aspects of the problem.

Suppose there are two five-year old boys in my neighborhood. One of them, A, is a child of very lucky temperament, who seems to be very happy most of the time, and who can derive great joy from minor presents. The other boy, B, has a rather unlucky temperament. He looks unhappy most of the time, and minor presents seem to give him only little satisfaction. I happen to have a little present in my pocket. Which boy should I give it to?

Utilitarian theory supplies a clear answer to this question: The present should go to that boy who is likely to derive more utility from it. Presumably, this means that it should go to A, who can be expected to get more immediate enjoyment out of it. (But this conclusion would have to be reversed, should I feel there was reasonable hope that receiving presents and other signs of attention might have a large enough beneficial long-run effect on B's unfavorable personality.)

In contrast, Rawls's theory would <u>always</u> favor giving the present to B, who is obviously the less fortunate of the two boys. Finally, Sen's theory would suggest that it should be given either to A or to B, depending on the actual distance between the two boys' utility levels: If B's utility level is <u>not</u> very much below A's, then A's higher marginal utility for the present will be the deciding factor, and the little present should go to A; whereas if B's utility level <u>is</u> very much below A's, then this difference in their utility levels will be the deciding factor, and the little present should go to B.

Once more, the issue is this. In case I come to the conclusion that, everything considered, I could create more human happiness by giving the present to A, am I permitted to do so? Or, am I required under certain conditions to discriminate against A and give the present to B, even though I know that A would make better use of it - - merely because A is already a pretty happy fellow?

As things are, A has a rather high utility level while B has a rather low one. This situation is not of my own making: it had already existed when I first appeared on the scene. The question is what obligations this state of affairs imposes on me. Sen and myself agree that one obligation I have is to create as much human happiness as I can in this situation. But Sen seems to hold that I am also under a second obligation (which in some cases may override the first) of compensating B for his low utility level as such. In my opinion, the question of compensating B does not arise at all. Of course, I would certainly owe B a fair compensation if his low utility level were a result of my own culpable actions in the past. But, under our assumptions, this is not the case. Therefore, unwarranted guilt feelings about B's low utility level should have no influence on my behavior. (Nor should my behavior be influenced, of course, by any irrational resentment against people who, like A, are lucky enough to enjoy high utility levels, achieved without any recourse to morally objectionable activities.) Rather my only obligation in this situation is to pursue the basic goal of all morally good actions, viz. to create as much happiness as possible in this world.

To conclude, Sen's proposal of making social welfare a nonlinear concave or quasi-concave function of individual utilities suffers from much the same difficulties as the lottery-variance argument does. It is quite unnecessary to make social welfare a nonlinear function of these utilities: this is so because the concavity (or near-concavity) of people's utility functions in money and in commodities already ensures that, normally, poor people's needs will receive much higher priority- - even if the social welfare function itself is linear in individual utilities. Indeed, introducing nonlinearities in the social welfare function would be not only unnecessary; it would be morally and logically inadmissible. It would be morally inadmissible because it would amount to unfair discrimination against people who happen to enjoy rather high utility levels. It would also be logically inadmissible, at least for people believing in an individualistic-humanistic moral philosophy, because, as I have shown [Harsanyi, 1955, pp. 312-314], the Bayesian rationality postulates, together with the individualism axiom [called "Postulate c" in my paper just quoted] ,logically entail the use of a social welfare function linear in individual utilities.

4. Interpersonal utility comparisons.

The Bayesian rationality postulates and the individualism axiom only imply that our social welfare function must be linear in individual utilities: but they say nothing about the weights we should give to the various individuals' utility functions.

Yet, it is natural to supplement this theory by adding a symmetry axiom, which requires that the social welfare function should treat different individuals' utility functions in a similar manner, and should assign the same weight to each of them.

This requirement, in turn, implies that our social welfare function must be based on _interpersonal utility comparisons_. For, in order to assign the same weight to the various individuals' utility functions, we must be able to express all of them in the same utility unit. Of course, when we first define a von Neumann-Morgenstern utility function for each individual in the usual way, we shall normally choose an independent utility unit for each individual. But, then, we must engage in interpersonal utility comparisons in order to estimate _conversion ratios_ between the different individuals' utility units. (For example, we may first choose, for each individual, as utility unit his utility for an extra $ 1. But then we must try to estimate how different individuals' marginal utilities for an extra $ 1 compare with one another.)

This dependence of linear social welfare functions, and of utilitarian theory in general, on interpersonal utility comparisons has given rise to a good deal of misunderstanding. Most of this misunderstanding could have been avoided if more attention had been paid to the close similarity between the role that subjective probabilities play in Bayesian decision theory and the role that interpersonal utility comparisons (interpersonal utility conversion rations) play in utilitarian moral theory.

More specifically, since in uncertain situations expected utility is defined in terms of subjective probabilities, Bayesian decision theory requires us to assign subjective probabilities to alternative contingencies. Obviously, this requirement is not based on assuming that human decision makers are necessarily very good at assessing these probabilities. Indeed, in situations where they have insufficient information to guide their probability judgements, these judgements are bound to be of rather poor quality. Nevertheless, Bayesian theory suggests that we should make our decisions on the basis of such probability assessments, even in situations of very insufficient information, because:

1. If our behavior is consistent with the Bayesian rationality postulates, then we simply cannot avoid making such probability assessments, at least implicitly. For example, suppose I have to choose between two bets. Bet A would make me win $ 100 if candidate X wins a particular election, and would make me lose $ 100 if he loses; whereas bet B would exactly reverse these two contingencies. Then, by choosing either bet, I will make an implicit probability judgement about the chances of his winning or losing the election. I simply cannot avoid making such a judgement.

2. Since we have to make such probability judgements at least implicitly, we shall be better off if we make them explicitly: this will enable us to avoid damaging inconsistencies in our probability judgements, and will enable us also to make the fullest possible use of the information actually available

to us - - however much or however little this information may be in any given case.

In the same way, utilitarian theory requires us to use a social welfare function based on interpersonal utility comparisons. However, this requirement in no way presupposes that human decision makers are particularly good at making such comparisons. Presumably, when we know two individuals (or two groups of individuals) reasonably well, and have good knowledge of the situations they are in, then we can compare the utilities they would derive from given commodity baskets (or even from less tangible benefits), with a tolerable accuracy. (For example, I think I can tell with some assurance which friend of mine would derive the highest utility from a Mozart opera, or from a good dinner at a French restaurant, etc.) On the other hand, utility comparisons between two individuals with social and cultural backgrounds unfamiliar to us must be subject to wide margins of error. But, in any case, utilitarian theory suggests that we should make our moral decisions on the basis of such interpersonal utility comparisons, because:

1. If we follow the axioms of utilitarian theory (and, in everyday life, all of us behave as utilitarians, at least some of the time), then we simply cannot avoid making interpersonal utility comparisons, at least implicitly. For example, when we have to decide whether we want to give a particular present to A or to B, then one of the important considerations (though perhaps not the only important consideration)

will be whether A or B is likely to derive a higher utility from it.

2. Since we have to make such interpersonal utility comparisons anyhow, we are under a moral obligation to make them with the greatest possible care, and with the fullest use of all relevant information available to us - - at least when they will serve as a basis for an important moral decision. Any attempt to avoid such interpersonal comparisons, owing to some philosophical prejudice against them, can only lead to careless and irresponsible moral decisions.

As I have argued elsewhere [Harsanyi,1955, pp.316-321; and 1973, pp. 23-26] , there is nothing mysterious in our undeniable ability to make such comparisons, with various degrees of accuracy, depending on the situation. Any such comparison is logically equivalent to a prediction of what our own choice behavior and our own emotional reactions would be in certain hypothetical situation - - possibly in situations very different from anything we have experienced so far. Trying to assess the utilities that another individual of a very different personality and social background would derive from various commodity baskets is not very different from trying to assess the utilities I myself would derive from various commodity baskets if my income, social position, personal situation, or emotional attitudes, underwent a major change.

Nobody claims that such assessments are always very reliable. All we claim is that in many cases we are under a moral obligation to make such assessments and, indeed, to make them as carefully and as knowledgeably as we possibly can - - just as, in many cases, we are under a prudential (pragmatic) obligation to make probability judgements, and to make them as carefully and as knowledgeably as we possibly can.

5. Super-egalitarian social welfare functions.

Owing to the great cogency of the Bayesian rationality postulates, it seems to me that it is a rather hopeless undertaking to build up a moral theory on a rejection of these postulates either for individual or for social choices [i.e., on a rejection of my Postulate a or b in Harsanyi, 1955, p.313] . Any such theory could only give rise to highly counterintuitive implications. Rather, anybody who wanted to construct a *super-egalitarian* theory, i.e., a theory more egalitarian than the utilitarian theory already is, would have to deny, or at least substantially weaken, my individualism axiom [Postulate c, op.cit.].

One possible approach would be this. Suppose that individual i wants to construct a social welfare function for society X. I shall here assume that i himself is an outside observer and is himself not a member of society X. (The model can be easily extended to the case where i is in fact a member.) Then, i may start with specifying his own political preferences about the income and utility distribution he would like to see in society X. For easier reference, I shall describe these as i's egalitarian preferences. (These preferences may depend on the variance of individual utilities and/or individual incomes in society X, or may depend on some more complicated function of individual utilities and/or incomes.) In view of my preceeding argument, it will be desirable to assume that i's egalitarian preferences satisfy the Bayesian axioms of rationality (e.g., Marschak's [1950] postulates). Then, i may replace my Postulate c by a less individualistic Postulate c^* : "If two alternatives are indifferent from the standpoint of each individual j in society X, and are also indifferent from the standpoint of my (i's) egalitarian preferences, then they are indifferent from a social standpoint as well."

By a slight modification of the proof given in my paper [Harsanyi,1955, pp.312-314], it is easy to verify that these axioms imply the following theorem:

Theorem. The social welfare function W of individual i must be of the following mathematical form

$$W = t \sum_{j=1}^{n} a_j U_j + (1-t)V_i ,$$

where $U_1,\ldots, U_j,\ldots, U_n$ are the utility functions of the various individual members of society X, V_i is a von Neumann-Morgenstern utility function defined in terms of i's egalitarian preferences, whereas t, and $a_1,\ldots,a_n$ are constants.

If i so desires, then he can choose t=0, and can choose to make V_i quite independent of the utility functions $U_1, \ldots, U_n$: in this case, of course, W will lose all connection with individual preferences. But, even if he does not go quite as far as this, the mathematical form of W will clearly indicate that the well-being of the individual members of society X is not his overriding consideration, and that he is willing to sacrifice their well-being, at least in some cases, to his own egalitarian preferences when the two conflict with each other.

Personally, I would find such a theory of ethics or of welfare economics highly objectionable from a <u>moral</u> point of view. To my mind, humanitarian considerations should never take a second place to any other considerations, including egalitarian ones. Indeed, at a deeper level of philosophical analysis, I think such a theory would be highly <u>irrational</u>, because individual i cannot have any rational motive for wanting to impose his own political preferences on the members of society X (whether he himself is a member of society X or not). But rationality is a concept of many dimensions: the theory I have described would at least be consistent with the most conspicuous requirements of rationality, by conforming to the Bayesian rationality postulates - - even if it violated some standards of rationality at deeper levels.

6. Linear social welfare functions and the problem of moral consensus.

I have argued that there are rather compelling reasons for using social welfare functions linear in individual utilities - - at least for people believing in an individualistic-humanistic moral philosophy. Among the people accepting this point of view, one source of moral disagreements will disappear: they will agree at least on the proper mathematical form of the social welfare function.

Of course, another rather obvious source of possible disagreements will remain: such people may still differ on how to compare the utilities of different individuals. Yet, one should not exaggerate the likely practical importance of this problem. In this respect, conventional treatments of welfare economics give a rather misleading impression. A little reflection will show that very few real-life policy controverties actually arise from disagreements about interpersonal utility comparisons.

This of course does not mean that agreement on the mathematical form of the social welfare function (even if this were accompanied by reasonably close agreement on interpersonal utility comparisons in all important cases) would bring us much nearer to a utopian state of moral consensus - - just as full acceptance of Bayesian decision theory would not bring us much nearer to a consensus on purely pragmatic policy problems (i.e., on those involving no controversial moral issues of any significance).

In my opinion, the most important sources of moral disagreements about what conditional and unconditional predictions - - whether deterministic or probabilistic predictions - - to make about future empirical facts. For example, we cannot expect a moral consensus in our society so long as we strongly disagree on what the likely effects of the conventional anti-inflationary policies will be in somewhat unusual economic situations, for which no real historical precedents may exist; or on what the long-run effects of continuing inflation will be on our social institutions; or how our children will be affected in the long run by an increased emphasis on "creativity", and a decreased emphasis on "intellectual excellence", in our schools; or on whether a greater pressure on certain despotic regimes will speed up or rather slow down their hoped-for liberalization and democratization, etc. No doubt, further progress in human psychology, in the social sciences, and, in some cases, in the natural sciences, will give clearer answers to some of these problems than now are available to us. But open problems of this type will always remain with us - - partly because the very solution of some of these problems will give rise to new ones. By solving some of these problems a radically new historical situation is created, and even if we have learned how to make predictions in the old situation, this may not help us very much in making predictions in the new one. Keynesian economics has enabled us to make much better predictions about the effects of various economic policies in conditions of mass unemployment. By this means, it has also enabled us to eliminate these very conditions, and to create a completely novel economic situation of continuing high employment, in which Keynesian predictions may no longer work. [8/] [9/]

Nevertheless, even if an agreement on the mathematical form of our social welfare function will not cure our most important moral disagreements - - as long as we keep on disagreeing in our predictions about future empirical facts - - such an agreement is an objective very much worth striving for. What a good formal theory of the social welfare function can do for us in the field of moral, political, and economic decisions is much the same as what Bayesian decision theory can do for us in the field of purely pragmatic decisions. It can help us in organizing our analysis of the situation, in clarifying what we do know and what we do not know, and what the implicit assumptions we are making are; and, most important of all, it can help us in bringing a large number of - - often quite heterogeneous - - pieces of information together in one coherent and systematic decision-making process.

In actual fact, even if we disregard its possible practical uses, it seems to be a philosophically rather interesting proposition, should it be true, that our basic criteria of rationality, together with an individualistic-humanistic moral philosophy, leave us no other option but defining our social welfare function as a linear combination of individual utilities and, indeed, as their arithmetic mean.

REFERENCES

Rudolf Carnap: Logical Foundations of Probability. University of Chicago Press: Chicago,1950.

Peter Diamond: "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility: A Comment", Journal of Political Economy, 75,(1967), 765-766.

Milton Friedman and Leonard J.Savage: "The Utility Analysis of Choices Involving Risk", Journal of Political Economy, 56 (1948), 279-304.

John C.Harsanyi: "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking", Journal of Political Economy, 61 (1953), 434-435.

John C.Harsanyi: "Cardinal Welfare, Individualistic Ethics,and Interpersonal Comparisons of Utility", Journal of Political Economy, 63 (1955), 309-321.

John C.Harsanyi: "A General Theory of Rational Behavior in Game Situations", Econometrica,34 (1966), 613-634.

John C.Harsanyi: "Can the Maximin Principle Serve as a Basis for Morality: A Critique of John Rawls's Theory", Working Paper CP-351 (May 1973), Center for Research in Management Science, University of California, Berkeley, Ca. 94720. To appear in the American Political Science Review.

John C.Harsanyi: The Tracing Procedure: A Bayesian Approach to Defining a Solution for n-Person Noncooperative Games", Parts I-II, Working Papers 15-16 (May 1974), Institute of Mathematical Economics, University of Bielefeld, Schloss Rheda, 484 Rheda, West Germany.

R.Duncan Luce and Howard Raiffa: Games and Decisions. John Wiley & Sons: New York, 1957.

Jacob Marschak: "Rational Behavior, Uncertain Prospects, and Measurable Utility", Econometrica, 18(1950), 111 - 141.

Roy Radner and Jacob Marschak: "Notes on Some Proposed Decision Criteria", in *Decision Processes* (Robert M. Thrall *et al.*, editors). John Wiley & Sons: New York, 1954.

John Rawls: "Justice as Fairness", *Philosophical Review*, 67(1958), 164-194.

John Rawls: *A Theory of Justice*, Harvard University Press: Cambridge, Mass., 1971.

Lionel Robbins: "Interpersonal Comparisons of Utility", *Economic Journal*, 48 (1938), 635-641.

Amartya K.Sen: *Collective Choice and Social Welfare*. Holden-Day: San Francisco, 1970.

Amartya K.Sen: *On Economic Inequality*, Clarendon Press: Oxford, 1973.

Henri Theil: *Optimal Decision Rules for Government and Industry*. North Holland & Rand McNally: Amsterdam & Chicago,1968.

FOOTNOTES

[1] The author wishes to thank the National Science Foundation for supporting this research through Grant GS-3222 to the Center for Research in Management Science, University of California, Berkeley. Thanks are due also to the Institute of Mathematical Economics, University of Bielefeld, where the author has been Visiting Professor in 1973-74; as well as to Professor Reinhard Selten, of the University of Bielefeld, to Professor Amartya K.Sen, of the London School of Economics, and to Professor Louis Gevers and Robert Deschamps, both of the University of Namur, for helpful discussions.

[2] The postulates of Bayesian theory have a very clear meaning in all situations where the utility of each "prize" is independent of the probability of attaining it. (Virtually all situations important for welfare economics belong to this category.) Difficulties of interpretation arise only when this is not the case, e.g., when part of the attraction of mountain climbing lies in its inherent danger (i.e., in the fact that the probability of safe arrival at the top is less than one [cf. Marschak,1950]). But even in such cases the postulates retain their formal validity if the "prizes" are properly defined (e.g., mountain climbing must be defined so as to include the presence of danger, or, more exactly, the joy of danger successfully overcome). However, this means that in such cases the behavioral implications of the Bayesian postulates will become somewhat ambiguous unless these postulates can be supplemented by a psychological theory predicting how the utilities of the prizes will quantitatively depend on their probabilities.

3. In this paper I shall mainly discuss the criticisms of Professors Diamond and Sen, since I have already considered Professor Rawls's theory elsewhere [Harsanyi, 1973] .

4. For a much more detailed discussion of Rawls's view, the reader is referred to Harsanyi [1973] . Copies are available on request from the Center for Research in Management Science, University of California, Berkeley, California 94720.

5. We need an "other things being equal" clause because, in general, this difference will depend, not only on the variance, but on the higher statistical moments as well.

6. The force of this argument will not significantly decrease even if we accept Friedmann and Savage's [1948] contention that people's utility functions may have some - - relatively short - - ranges of increasing marginal utility for money. Globally, these functions will still display an overall tendency to decreasing marginal utility for money.

7. Only Rawls's social welfare function escapes this requirement because it gives infinitely more weight to the utilities of some individuals than it gives to the utilities of other individuals.

8. On deeper analysis, disagreements about interpersonal utility comparisons, also, are disagreements about predictions. in certain hypothetical situations - - such as disagreements about what the preferences and the emotional reactions of certain individuals would be if their incomes, social positions, education levels, cultural attitudes, and even their personalities, where different from what they actually are.

9. Most contemporary philosophers divide declaratory statements in two main classes, viz. logical - mathematical statements, and empirical statements. Of course, this leaves no proper room for philosophical(conceptual)statements, which cannot be really subsumed unter either category (see Section 1 above). Moreover, this classification also pays insufficient attention to the special problems connected with statements about future empirical facts, which obviously cannot be verified in the same way as statements about present or past empirical facts can, but which are major ingredients to some of our most important value judgements.