Rawls and Utilitarianism

Holly Smith Goldman
UNIVERSITY OF ILLINOIS
AT CHICAGO CIRCLE

One of the major polemical concerns of John Rawls' *A Theory of Justice*—perhaps *the* major polemical concern—is to provide a satisfactory alternative to the utilitarian account of social justice (15, 22, 166; all parenthetical citations are to John Rawls' *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971)).

In its most general form, utilitarianism is the theory that objects of moral appraisal, such as actions, social institutions, moral codes, or traits of character, can be evaluated strictly in terms of their impact on general human welfare. If an action (or institution, code, etc.) has better consequences for human welfare than those of its rivals, then it is morally acceptable; otherwise, it must be rejected. Utilitarian thought, which flowered in the writings of the classical utilitarians, Jeremy Bentham, John Stuart Mill, and Henry Sidgwick, has been the dominant conception in Anglo-American moral and social philosophy for roughly the past two centuries.[1]

The appeal of utilitarianism is clearcut. First and most important, it identifies effects on human welfare as *the* criterion to use in assessing social phenomena. It is impossible to deny that human welfare is relevant to such assessments, and it is difficult at least initially to imagine that anything else could possibly be relevant. Second, utilitarianism presents us with a *single rule* which covers all decision-making. This is one of its major advantages over what Rawls terms "intuitionistic theories," theories which present us with a plurality of rules to use in making decisions, but which typically fail to guide us in bal-

ancing the importance of these rules when they conflict (51). Thus one such theory tells us both to do good and also to treat people equally, but it does not tell us what to do when, for example, treating people equally would produce less good than treating them unequally.[2] Utilitarianism, which employs only one criterion, can never be faced with such a problem. Finally, utilitarianism promises to provide us with a *precise formula* for making decisions, one which resolves every dilemma by a process of calculating the effect on human welfare which is relatively invulnerable to the whims and biases of all-too-human decision-makers. Here again, intuitionist theories fall short, for their application typically relies on decision-makers' intuitions about the weights to be assigned the various conflicting considerations, intuitions whose moral basis is uncertain, and which are likely to be distorted by personal interest in the case.

For these kinds of reasons, utilitarianism has seemed to many to be the only serious contender in our search for an adequate moral theory. Nevertheless severe criticisms have been brought repeatedly against utilitarianism in its turn. In the resulting stand-off, Rawls' theory appeared in 1971 as a long-awaited alternative which managed to avoid the defects of both utilitarian and previous intuitionistic views. Despite the criticism which Rawls' book has attracted, many critics agree with A.M. MacLeod in holding that one of its prime merits is that it "succeeds brilliantly in displaying the inadequacy of a utilitarian theory of justice."[3] In this article I will attempt to assess the truth of this claim, by summarizing Rawls' main arguments in *A Theory of Justice* against utilitarianism, and then exploring and evaluating the responses which may be made in defense of utilitarianism. Since many of Rawls' arguments involve comparing his principles of justice with utilitarianism, of necessity much of the following discussion will have the same character. I will assume that the reader is familiar with the general structure of Rawls' theory and the argument for it.

EVALUATION OF RAWLS' EXTRA-CONTRACTARIAN ARGUMENTS

The Contrast

Since Rawls' theory of justice and utilitarianism are viewed as *rival* theories, we must first be clear on what the character of the two

rivals is. This is problematic on both sides. Rawls' theory is a general theory of social justice which includes not simply principles of justice, but also an elaborate edifice supporting those principles, an edifice which has no counterpart in utilitarianism proper. Utilitarianism, on the other hand, appears in many forms, and we must narrow the focus of inquiry down to the one which is the most clearcut rival to Rawls' system. This is consonant with Rawls' own practice, since his discussion focuses on one version of utilitarianism which he takes as representative of the best utilitarian thinking, even though he states that his theory represents an alternative to utilitarian thought generally (22).

Strictly speaking, utilitarianism is a normative theory: a theory to be used in evaluating human phenomena and deciding among them. This theory may be, and has been, argued for in many different ways. For example, some have argued that utilitarianism provides the best explanation of our common moral beliefs, others have argued that "right" simply *means* "conducive to the greatest good," while others have argued that utilitarianism provides us with a rational decision procedure which is better than any alternative. We may call such arguments "meta-ethical" justifications for a normative theory. The major *normative* components in Rawls' theory are his principles of social justice. Thus it is these which are to be compared with utilitarianism. Other salient aspects of the theory—the argument from reflective equilibrium, the idea of an original position, the derivation of principles of justice from an original position—are to be viewed as primarily meta-ethical devices used in arguing for these principles. Thus these aspects are not necessarily in contention between utilitarianism and Rawls' theory of justice, since potentially they may be, as Rawls himself notes, used to argue for utilitarian principles rather than the ones Rawls proposes (121).

Rawls takes the subject of principles of justice to be what he calls the "basic structure of society," that is, the ways in which the major social institutions distribute fundamental rights and duties, and determine the division of advantages from social cooperation. Major social institutions are such things as the political constitution, and the principal economic and social arrangements (7). On his view, principles used in assessing this basic structure must satisfy what he calls the "formal constraints on right"—conditions such as universality and publicity. Since he concedes that utilitarianism can meet these requirements, we need not investigate them in more detail (130-131; but see discussion in "The Strains of Commitment: Psychological Stability" below).

Rawls proposes two different conceptions of social justice which he calls the "general" and the "special" conceptions of justice. The general conception is the more basic one and is expressed as follows:

> All social primary goods—liberty and opportunity, income and wealth, and the bases of self-respect—are to be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored. (303)

This conception of justice applies when the social wealth is low enough that the basic liberties cannot be effectively established or exercised for all citizens (152, 542-543). As the level of civilization improves, a special case of this conception comes into play, the famous two principles of justice which comprise Rawls' "special conception." For our purposes, the following will serve as a statement of these principles:

> First: each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others (60).
>
> Second: social and economic inequalities are to be arranged so that they are both (a) to the greatest benefit of the least advantaged and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.[4] (83)

As is well-known, the two principles of the special conception are "lexically ordered," so that society concerns itself first with satisfying the first principle, and only then with satisfying the second.

We now have an account of Rawls' two conceptions of justice which form his side of the contrast. Let us turn to the utilitarian side. As noted above, the utilitarian criterion may be applied to many different objects of moral appraisal. In this century it has perhaps been most common to view it as a principle for appraising the individual acts of human agents. This version of utilitarianism contrasts with what Rawls terms the system of natural duties and obligations for individuals (333-335). But what we are interested in here is the version of utilitarianism which applies, as Rawls' principles of justice do, to the basic structure of society. Such a version states that the basic structure is just if and only if its consequences for human welfare are at least as good as those of any alternative structure.[5]

This statement leaves two important questions open. In assessing the consequences of a social system for human welfare, do we consider the *total* human welfare, or the *average* human welfare? Classical utilitarianism concentrated on total welfare, and indeed it makes no difference which is chosen in cases where changes in population

size are not an issue. But if total population can be increased in such a way that total welfare increases, even though the average welfare falls, then the classical view advocates increasing the population, whereas the average view does not. Rawls concurs with the utilitarians who have taken the average view on this problem to be more satisfactory, and therefore takes the average view as the rival to which his principles of justice are to be compared (161-164).[6] Although there are problems with the average view, for the sake of argument, I shall do the same.

The second question to be addressed is how the notion of "human welfare" is to be interpreted. In the history of utilitarianism, many accounts have been given, the predominant ones identifying human welfare with happiness, pleasure, or the satisfaction of desire. Rawls addresses himself to a form of utilitarianism which defines the good as the satisfaction of desire, or more accurately the satisfaction of rational desire (25). A desire is satisfied when the state of affairs desired obtains, whether or not the person desiring it is aware that it obtains. The notion of "rational" desire has received many interpretations; for our purposes, we can say that a desire is rational just in case it is based on true beliefs concerning the matter at issue.[7]

One of the greatest contributions of Rawls' book is his development of the theory of the good, but he states that the main idea is that "a person's good is determined by what is for him the most rational long-term plan of life given reasonably favorable circumstances....To put it briefly, the good is the satisfaction of rational desire" (92-93). Thus the theory of the good is "not in dispute between the contract doctrine and utilitarianism" (92). What is in dispute is the relation between justice and the good. Our concern then is the contrast between Rawls' general and special conceptions of justice on the one hand, and on the other hand, a version of utilitarianism which states that the basic structure of society is just if and only if the average satisfaction of rational desire which it produces is at least as great as that which would be produced by any alternative structure.[8]

The First Extra-Contractarian Argument: Reflective Equilibrium

Let us now turn to the arguments by which Rawls attempts to show that this version of utilitarianism is inferior to his own principles of justice. These arguments can be divided into two categories. Some

are designed to show that utilitarianism is inadequate without invoking all of Rawls' contractarian apparatus, that is, the original position, the veil of ignorance, and so forth. I shall call these arguments "extra-contractarian" arguments. Others are designed to show that parties to the original contract would not adopt utilitarianism by preference to Rawls' principles of justice. I shall call these the "contractarian" arguments. In some cases, an argument may fall into both of these categories, or it may not be clear which category it is intended to fit. In such cases I shall classify the argument for expositional convenience. The remainder of this section will be concerned with four major extra-contractarian arguments: (i) the argument that utilitarianism generates prescriptions which violate our considered moral judgments concerning what is just and unjust; (ii) the argument that the reasoning in favor of utilitarianism illegitimately "merges persons"; (iii) the argument that utilitarianism requires us to make interpersonal comparisons of utility, or welfare, which have no scientific basis; and (iv) the argument that utilitarianism prescribes the satisfaction of desires which themselves are the product of possibly unjust institutions. Only the second of these arguments is novel, but for many readers, Rawls has given them a statement which is especially forceful and illuminating.

This section will be devoted to a consideration of the first of these extra-contractarian arguments. Let us note initially that Rawls must regard it as one of the most fundamental available to him, for one of the major justifications (perhaps *the* major justification) he offers for the entire contractarian apparatus he describes is the fact that it generates principles and prescriptions for individual cases which are in "reflective equilibrium" with our considered judgments about what is just and what is not. The final court of appeal is to these judgments, or "intuitions," and a showing that utilitarianism violates them would provide Rawls with what he views as the definitive argument against it. Since the contractarian apparatus is designed precisely to ensure that principles which are in reflective equilibrium are chosen by the parties in the original position, any arguments against utilitarianism derived from that apparatus are logically secondary to the extra-contractarian arguments, and specifically to the first of them (19-21; § 9, § 87).[9]

In arguing that utilitarianism generates prescriptions which violate our considered judgments on justice, Rawls advances two related claims. Both arise from the fact that utilitarianism is fundamentally unconcerned with how welfare, or the satisfaction of desires, is dis-

tributed over the population. First, he claims that utilitarianism requires some individuals to suffer lesser life prospects simply so that others may enjoy a greater sum of advantages (4, 14, 26, 177-178, 180). Second, he claims, as a special case of greater concern, that utilitarianism requires that the liberties of some be sacrificed for the sake of greater goods for others (3-4, 26, 176). Thus it has sometimes been held that utilitarianism could justify either slavery or serfdom, or other serious infractions of liberty, for the sake of greater social benefits (156).[10] According to Rawls, our considered judgments of justice reject such social systems: "[according to] our intuitive conviction... [e]ach person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override. For this reason justice denies that the loss of freedom for some is made right by a greater good shared by others. It does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many" (3-4; also 27-28). It may appear that Rawls' theory cannot condone such inequalities, for the first principle of the special conception of justice prohibits liberties from being traded for any gain in other social goods, and part (a) of the second principle, called the "Difference Principle," only allows inequalities in economic goods when they benefit the least well-off, not merely when they maximize the average welfare. Rawls claims in fact that such inequalities must benefit *everyone* (80, 102-104, 178-179).

What response can the defender of utilitarianism make to this argument? The first thing he may do is challenge the premise that our considered moral judgments provide a relevant criterion for assessing the adequacy of a normative theory. Rawls takes it that our considered moral judgments are the "class of facts against which conjectured principles can be checked" (51), because he is attempting to provide a "theory of the moral sentiments... setting out the principles governing... our sense of justice" (51). But the utilitarian (or any other moral theorist) may feel that *he* is trying to establish the true, or correct, principles of justice, not simply characterize or systematize the judgments we currently make. And he may feel that our actual moral sentiments are no test of truth, since they are likely to "derive from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances which now lie in the distant past."[11] Thus he may choose to ignore the pronouncements of common sense.

Other utilitarians have taken a slightly different attitude towards our reflective moral judgments. On their view, our judgments do not express discarded religious systems or outmoded empirical beliefs. Rather they are based on a moral code whose currency *under our social conditions* tends to promote human welfare. Thus even though our judgments may appear nonutilitarian in character, in fact they typically accord with what the principle of utility would recommend. And it is *this* fact which makes them a valid test of truth. Now, it is frequently argued, with some plausibility, that the principle of utility only allows grave infractions of personal liberty in social conditions which are significantly different from our own.[12] Our code deems such infractions to be unjust. However, there is no reason to suppose that the currency of our code in social conditions significantly different from our own would promote human welfare. Thus it is invalid to apply our code to social circumstances unlike ours. We feel that slavery is always unjust, but our feeling is conditiond by our own circumstances, which may not be replicated elsewhere. Where slavery would promote human welfare, it must be recognized as permissible, even though *we* feel, on the basis of our code, that it is wrong. Once again, inconsistency with reflective moral judgments is rejected as irrelevant.[13]

To resolve the issue of whether or not considered moral judgments form a test for the adequacy of any moral theory would require us to provide a complete account of how moral theories are to be justified, a matter beyond the scope of this paper. We must leave the problem here, therefore, with only the preceding sketch of the arguments which may be made on either side.[14]

Let us turn to the substance of the argument, the claim that the precepts of utilitarianism and those of reflective morality do differ in at least some cases. According to Rawls, we intuitively feel that it is unjust to impose sacrifices on a few in order to gain a larger sum of advantages for others (3-4). He states that utilitarianism violates this precept (178). It is not completely clear how we are to interpret this objection. Rawls may have in mind here the classic complaint against utilitarianism that it is *non-egalitarian*, i.e., that it justifies any distribution of goods as long as it maximizes average utility, however unequal that distribution may be. Thus imagine a society which must choose between two arrangements, in the first of which the worst-off persons receive an annual income of five hundred dollars while the best-off persons receive five million, and in the second of which the

worst-off persons receive five thousand while the best-off persons receive fifty thousand. If the first of these arrangements would maximize average utility, then utilitarianism prescribes it even though it involves far greater disparity between economic classes than the second. Many of us would find the first arrangement the less just and so would disagree with utilitarianism's recommendation.

Many utilitarians have felt they have an adequate response to this charge. They concede that it is *logically* possible for a society to confront the choice just described. However, they claim that the facts of human psychology make it *empirically* impossible for the first arrangement to be the one which maximizes average utility. They argue, for example, that human beings have similar utility functions for goods such as income, utility functions which satisfy the conditions of diminishing marginal utility (see Rawls' discussion, 159). Thus a hundred dollars taken away from someone who has five million will produce greater utility when given to someone who only has five hundred. In light of this, utilitarians have argued that under conditions as we know them, utilitarianism would not produce radically unequal distributions of goods or utility.

Rawls is quite prepared to reject arguments against utilitarianism which claim that it generates unacceptable prescriptions for *merely* logically possible conditions. He states that he is not interested in what he calls the "ethics of creation," but only in determining what ethical theories are appropriate for the natural and human world as we know it (159–160). Indeed he defends his own theory against certain objections by pointing out that those objections involve applying his theory to abstract, but empirically unrealistic, possibilities (157–159). Thus he must argue against utilitarianism that it generates unacceptably non-egalitarian results under realistic conditions. But his statements on this score are surprisingly ambivalent. Sometimes he says that utilitarianism may "demand that some should forego advantages for the sake of the greater good of the whole" (177). But at other points he seems to concede the utilitarian response: "...there is no reason in principle why the greater gains of some should not compensate for the lesser losses of others....It simply happens that under most conditions, at least in a reasonably advanced state of civilization, the greatest sum of advantages is not attained in this way" (26). Rawls' ambivalence is understandable here since no one really knows enough about the relevant empirical facts to be sure what utilitarianism implies for our world.

However, we can point out that Rawls himself makes empirical assumptions which would be sufficient to show that utilitarianism probably mandates strongly egalitarian distributions. In arguing that the parties in the original position would employ a maximin strategy in choosing their principle of justice, he assumes that there is a satisfaction threshold beyond which goods such as income have a sharply declining marginal utility, and also that there is a toleration point below which amounts of such goods would be intolerable (see "The Second Feature" and "The Third Feature" below). But if these conditions hold, it is extremely likely that average utility would be maximized by allocating to each person an amount of these goods which falls between these two points. Any amount of goods less than the toleration level would produce such severe disutility that it could only be counterbalanced by huge numbers of persons above that level; while amounts of goods greater than those at the satisfaction point would have to be possessed by huge numbers of people in order to counterbalance any amount below that level experienced by others. Neither of these possible patterns of distribution is likely to be one which realistically would be faced by a society. Since Rawls also assumes that the satiation point could be guaranteed to every member of society under Rawlsian justice, it seems unlikely that the difference between that point and the tolerance level can be too large. Thus utilitarian distribution, while it may not be precisely egalitarian, should not involve too glaring disparities between the rich and the poor—at least on Rawls' own empirical assumptions. Given Rawls' reluctance to criticize theories of justice for their applications in unrealistic conditions, and given the empirical assumptions he makes in arguing for his own theory, he himself is not in a strong position to press the objection that utilitarianism leads to radically non-egalitarian distributions in realistic cases. Neither are we in a position to accurately assess the matter until we have more empirical information and in particular some account of how utility is to be measured, a problem which has long hamstrung utilitarian theory. The egalitarian or non-egalitarian tendencies of utilitarian theory must remain a matter of conjecture only.

Rawls' language in his statements that utilitarianism violates common morality suggests that he may have something slightly different in mind than the objection we have examined so far. As we have seen, he repeatedly asserts that utilitarianism requires "sacrifices" from a few in order to gain a larger sum of advantages for others. Let us con-

sider what might be meant by this. Normally, when a person sacrifices something, he voluntarily gives it up. Clearly, however, Rawls does not envision voluntary sacrifices but rather something like the *imposition*, through institutions like the taxation and welfare systems, of lesser prospects for some than they might have enjoyed, so that others are assured of better prospects than they might have endured.[15] But what are we to use as the standard of reference, the situation these individuals might have enjoyed or endured, relative to which their sacrifice or benefit is measured? One possibility is that we should use as this standard of reference the situation that *would have obtained* if the present social system had not been in force. Thus to ascertain whether or not some members of a utilitarian society are making sacrifices, we must ask what their expectations would have been under the principle of justice which would have prevailed if utilitarianism had not. But this is a useless question: in any concrete situation, it would probably be impossible to know what system would have prevailed instead of utilitarianism. Insofar as we have any grasp on the issue, what seems to be true is that utilitarianism and Rawls' principles are on an equal footing here. That is, in most utilitarian societies, there will be some members who would have done better under the form of justice which would have prevailed if utilitarianism had not. But the parallel statement is true of most Rawlsian societies. By this criterion, utilitarianism fares no worse than Rawls' own principles —and probably no worse than any other plausible theory of justice, either.

However, this standard for measuring sacrifice is probably mistaken. When we believe someone has been sacrificed, we mean, not that he is worse off than he would have been in the situation that would have obtained if the present social system had not been in force, but rather that he is worse off than he would have been had the *correct* system of minimal justice or morality prevailed. Thus, when stolen goods are taken from a thief and restored to their rightful owner, we do not say that the thief's interests have been sacrificed for those of the owner—even though, had the present social system not been in force, the thief would have retained possession. A thief is required by minimal justice or morality to return whatever he has taken, so that his doing so, or being forced to do so, is not seen as an occasion of sacrifice.

By this standard, the question for Rawls is whether or not reflective morality would view the arrangements dictated by utilitarianism as requiring transfers or allocations of goods beyond those required

by minimal justice. Quite conceivably this is so. (It should be kept in mind here however that many of the "sacrifices" it is traditionally alleged utilitarianism would require involve only one or two "victims." What may have positive utility when done, perhaps in secret, to one individual is far less likely to have the same utility when institutionalized in the basic structure of society.) However, it appears that reflective morality would also view *Rawls'* principles as calling for sacrifices. In defending his principles on this score, Rawls emphasizes the fact that those in the worst-off position in society are nevertheless in the best position they could hope for, and hence cannot be viewed as making sacrifices. However, this defense is inadequate. First, while it may be true that the worst-off class is better off than the *analogous* worst-off class in any alternative society, it is *not* true that any given individual who is a member of this class is *himself* as well off as he might be under some different organization of society. Typically, for every individual, there will be some alternative society in which he would have done better. Moreover even the worst-off class itself, identified as Rawls suggests—e.g., as "unskilled workers" (98) —might well do better under a different organization. Since we are not yet at the point in Rawls' argument where the veil of ignorance can be introduced as a device to prevent individuals from being concerned with their personal fates, Rawls' focus here on the respective worst-off classes under different social arrangements is premature.

Second, whatever we conclude about the sacrifices suffered by those in the worst-off class, it is certainly not true of those in better-off positions that they are in the best possible position under Rawlsian justice as opposed to some alternative, for his conception of justice only allows them to better their positions when doing so is not to the detriment of those below them. If improved wages for them do not act to stimulate the economy and so indirectly to benefit those lower down, then transfers of income through welfare payments are called for which will directly benefit those on the bottom (§ 43). The question is whether or not reflective morality regards such transfers as *sacrifices* on the part of the better-off, perhaps sacrifices which generosity or benevolence recommends but which justice does not require. There is ample evidence that such transfers are precisely so regarded by many who are presently required to make them: the current American "taxpayers' revolt" is at least partly fueled by this feeling. Of course, it is unknown to what extent such a feeling is "reflective," in Rawls' sense, much less "philosophically reflective" (§ 9). Still, the force of the sentiment should make us suspect that

Rawls' principles, as well as utilitarianism, would be seen as requiring sacrifices. Thus in comparing utilitarianism and Rawls' principles on the issue of sacrifices, we must ask which theory would be found by common opinion to require the *more objectionable* sacrifices. Presumably Rawls believes his theory does better in this regard. But until we know more surely what current moral feeling holds as the minimal standard of morality, and what precise distributions utilitarianism requires, it appears that we had better leave the issue unresolved without attempting to declare either conception of justice the clear winner in this matter.

Let us look at the much narrower claim to which Rawls devotes the greater part of his attention, namely the contention that utilitarianism may require unacceptable sacrifices of *liberty*. In an earlier work, Rawls maintained that "ordinary conception of justice... [holds] that slavery is always wrong."[16] This is probably incorrect, but let us grant for the sake of argument that this is true. It is also undoubtedly true that utilitarianism condones the institution of slavery under some imaginable conditions, perhaps conditions in which slavery is more humane and less exploitative than the forms with which we are familiar.[17] (Note that our condemnation of slavery becomes less and less certain as its form becomes more humane, and as it becomes more likely that utilitarianism would allow it.) Thus utilitarianism and common morality (we shall say for the sake of argument) do conflict. But this fact in itself is not enough to show the superiority of Rawls' principles of justice, for they too permit slavery and serfdom under some circumstances. The illusion that they absolutely prohibit these forms of servitude is encouraged by Rawls' practice of speaking as though we need only compare utilitarianism with his *special* conception of justice, that is, the well-known two principles. These principles do indeed appear to bar slavery because they do not allow liberties to be traded for economic goods.[18] However, the proper comparison is between utilitarianism on the one hand and Rawls' special *and* general conceptions of justice on the other. And the *general* conception of justice, which permits liberties to be traded for other social values, would allow slavery in a case where the slaves were better off overall under slavery than they would be if all social values were distributed equally. Such a case might arise when a harsh natural environment and low degree of capital development make it impossible to sustain everyone unless some submit to a condition of servitude, or when the necessity of repelling powerful external enemies makes it necessary to organize the state on a militaristic basis. Thus Rawls' system of justice—taken in its entirety—violates the common moral stricture against slavery and serfdom just as utilitarianism does.

The question we must pose then is whether or not utilitarianism violates this stricture in some more objectionable fashion than Rawls' principles do, for example by allowing slavery in circumstances where the special conception of justice is in force and would prohibit slavery and serfdom. At this juncture of the debate, several critics have claimed that utilitarianism would bar slavery and serfdom in all the same circumstances that mandate the application of Rawls' special conception of justice, and hence that there is nothing to choose between utilitarianism and the special conception on these grounds.[19] Their argument proceeds as follows. Rawls claims that the special conception of justice would come into force when the advancing level of civilization makes the effective exercise of basic liberties possible, and so gives them such value to each individual citizen that he would be unwilling to accept a lesser liberty in trade for any increase, however great, of other goods such as income and wealth. Under these circumstances it is in the best interests of each to have in force a principle of justice, such as Rawls' special conception, which accords absolute priority to justice.

> Now the basis for the priority of liberty is roughly as follows: as the conditions of civilization improve, the marginal significance for our good of further economic and social advantages diminishes relative to the interests of liberty which become stronger as the conditions for the exercise of the equal freedoms are more fully realized. Beyond some point it becomes and then remains irrational from the standpoint of the original position to acknowledge a lesser liberty for the sake of greater material means and amenities of office.... To be sure, it is not the case that when the priority of liberty holds, all material wants are satisfied. Rather these desires are not so compelling as to make it rational for the persons in the original position to agree to satisfy them by accepting a less than equal freedom. (542-543).

Thus Rawls makes the empirical assumption that at a certain point in economic development, each individual places such a high relative value on liberty that he finds no increase in his material wealth to be worth the amount of liberty he would have to give up in order to secure that increase. But this is an assumption about individuals' *utility* functions for the various goods: the satisfaction they would derive from economic and social goods is infinitesimal compared to what they would derive from liberty. We may infer that Rawls also

assumes that the value of liberty for one individual is not infinitesimal compared to the value of liberty for another. If these assumptions are correct, then utilitarianism would require giving each individual the maximum amount of liberty compatible with no reduction in the amounts of liberty assigned to other citizens. Among the states which accomplish this goal, it would then select the one which maximizes average satisfaction over economic goods. But it would not trade liberties for economic goods.[20] Thus, in the circumstances envisioned, *neither* Rawls' special conception *nor* utilitarianism would allow some to be enslaved that others might enjoy greater economic advantages.

This leaves the possibility that under social circumstances where the special conception would come into play it might maximize average utility for society to reduce one person's (or group's) liberties in order to increase the *liberties* of others. However, in Rawls' more elaborate statement of his special conception, he also allows liberties to be unequal if the inequality is acceptable to the person who bears the lesser liberty (250, 302). Presumably this would occur when the least well-off person, in terms of liberties, sees that even so he is better off in terms of liberties than he would have been had liberties been arranged equally (see 247). Utilitarianism of course will permit this sort of case as well since average welfare would thereby be increased. The question is whether utilitarianism would go *beyond* this, allowing some to have less liberty than they would at the equal-liberty point so that others may have more. In principle of course this is possible. But whether or not it would actually occur depends on empirical facts about the level of economic and social development necessary to bring the special conception of justice into play, the character of people's utility functions at that point for liberty, and other matters which Rawls gives us little guidance on and which it is difficult to judge in the abstract. The moral we should draw here is that (at least if Rawls' own empirical assumptions are true) the contrast between utilitarianism and Rawls' principles of justice with respect to treatment of basic liberties is far less dramatic than much of Rawls' discussion would suggest. Rawls himself admits this in one passage, where he states that "It simply happens that under most conditions, at least in a reasonably favored stage of civilization, the greatest sum of advantages is not attained [by a system in which] violation of the liberty of a few... [is] made right by the greater good shared by many" (26). This "favored stage of civilization" is exactly

the same one in which Rawls' own special conception of justice prohibits violations of liberty, and outside of which his general conception does not. Despite the rhetoric which has arisen following publication of his book, Rawls' arguments in themselves do not establish a great disparity between the adequacy of his treatment of liberty and that given to us by utilitarianism, although the door remains open for future investigators to accomplish this task.[21]

The Second Extra-Contractarian Argument: Merging Persons

Rawls' second argument against utilitarianism, from the extra-contractarian standpoint, criticizes it on the grounds that reasoning in favor of the principle of utility fails to take seriously the distinctions among persons. The reasoning in favor of utilitarianism is depicted as taking two different forms.

This reasoning, as it appears in the first form, goes as follows. An individual, in attempting to advance his welfare, must take into account the fact that he has competing desires, not all of which can be satisfied. Most thinkers have held that the rational solution to this problem is for the individual to act so as to maximize his overall satisfaction: to sacrifice the satisfaction of less intense desires in order to satisfy more intense desires, and to sacrifice the satisfaction of a smaller number of desires in order to satisfy a larger number. But, the argument continues, this same principle may be applied to a society as well as to an individual because the society faces the same problem, that of advancing the welfare of the group when the desires of some members conflict with the desires of others. If it is rational for an individual to maximize overall satisfaction, it is rational for society to do so as well, sacrificing the satisfaction of less intense desires in order to satisfy more intense desires and sacrificing a smaller number of desires in order to satisfy a greater number. Thus a utilitarian principle is argued for by extending to society as a whole the decision principle which is rational for an individual (23–24).

The second version of this argument for utilitarianism takes the following form. It is pointed out that we believe a person is likely to make a correct moral judgment when he is impartial, considers all sides of the issue, knows all the relevant facts, and so forth. Our belief is then elevated into a definition of rightness: a social system is said to be right when an ideally rational and impartial spectator would

approve of it from a general point of view should he possess all relevant knowledge of the circumstances. From this definition one cannot derive any actual assessments of the rightness or wrongness of particular systems because it is still indeterminate what such an ideal observer would approve of. However, this defect can be remedied by stipulating in addition that the ideal observer is to be *sympathetic*—that is, he completely identifies with, and in fact acquires himself in the same degree of strength, the desires of all the parties involved in a given case. Thus if Jones desperately wants the last eclair and Smith desires it only mildly, the ideal observer wants it desperately for Jones and mildly for Smith. Such an ideal sympathetic observer will incorporate in himself all the interests and desires which are relevant to a particular assessment and so approve of a social system only if the existence of that system would maximize the satisfaction of his (expanded) set of desires. Thus from such a definition of "right" is deduced the principle of utility as the correct criterion for determining social justice (27, 184–188).

Rawls appears to object to these lines of reasoning on three separate grounds. The second form of the argument, according to him, has been put forward by those who believe that the sympathetic impartial observer provides us with the correct interpretation of the notion of "impartiality" which is so crucial in moral judgment. However, he claims, another interpretation of this notion is available, namely that provided by the idea of principles which would be chosen in the original position, for the parties in the original position are ones whose situation and character enable them to judge without bias or prejudice (189–190). This point is not decisive, for the fact that *two* interpretations of some notion are available does not show one of them to be mistaken.

Second, and more relevantly, Rawls seems to believe that the interpretation of impartiality embodied in the second form of reasoning depicted above is incorrect because it mistakes impersonality for impartiality. Evidently Rawls believes that the ideal sympathetic spectator must be understood as an impersonal being, not an impartial one. But this is wrong: the sympathetic spectator may be superpersonal insofar as he incorporates in one system of desires, desires identical to all the desires of those around him, but he possesses a full personality and all other attributes of human character. He is impartial because he gives equal consideration to the desires of everyone concerned. Any serious worry that impersonality may have illegitimately been substituted for impartiality should instead be directed at *Rawls'* account of impartiality, because the parties in the original position possess no recognizable human personality. Rawls denies them any knowledge of their place in society, their social status, their natural abilities and assets, their idiosyncratic psychological features, even their conception of the good (137). Some writers have claimed that even *ignorance* on their part of these matters is not sufficient since we can be moved by desires of whose existence we are unaware; to achieve his aims Rawls must deny that they possess these desires at all.[22] But surely such a creature is an impersonal agent if anyone qualifies for this title. Thus Rawls' second concern here seems to cut more severely against his interpretation of "impartiality" than it does against that proposed in the utilitarian reasoning.

Rawls' third objection is his charge that extending to society the principle of choice appropriate for one individual involves not taking seriously the distinction between persons (27, 187). It is true that utilitarianism does not take the distinction between persons seriously in the sense that it does not protect an individual in principle from having his interests neglected in order to promote the interests of others. Presumably this is what Rawls really objects to. But we need not read into this doctrine any confusion concerning the metaphysical difference between persons, as Rawls' words might suggest. There is no more grounds for discovering a confusion here than there is for discovering one in Rawls' own theory that the distribution of natural talents is to be viewed as a common asset whose benefits are to be shared by all (101–102, 179).

Kenneth Arrow argues in addition that the notion of conflating all desires into one system cannot be faulted. A theory of justice is presumably an ordering of alternative social states and therefore is formally analogous to the individual's ordering of alternative social states. Moreover, there is widespread agreement that justice should reflect individual's satisfactions, so social choice made in accordance with *any* of these theories of justice necessarily involves "a conflation of all desires"—albeit one which is purely formal, and not envisioned as embodied in some individual.[23] Rawls' objection cannot be so much to the conflation as to the principle of choice utilitarianism employs.

Before leaving a consideration of this objection to utilitarianism, we should note briefly that the lines of argument Rawls describes only serve as reasoning in support of the *classical* form of utilitarianism, not the *averaging* view which is his and our selected rival to his

theory. Moreover they are ones which only a few advocates of the classical view have actually proposed (see 188n.), and are not found in any contemporary defenders of utilitarianism.

The Third Extra-Contractarian Argument: Interpersonal Comparisons of Utility

Rawls' third objection to utilitarianism, from a standpoint independent of contractarian theory, arises from the fact that utilitarianism requires us to make theoretically difficult interpersonal comparisons of utility while Rawls' principles supposedly do not. If we are to maximize average utility, we must have a cardinal measure of each person's welfare, plus the ability to make sensible comparisons between the welfare level of one person and that of another (90). Some utilitarians have been content to leave these comparisons and measurements to unguided intuition, but as Rawls correctly points out, this is a poor basis for social policy. We do not want large-scale allocative decisions to rest on intuitive judgments which are likely to be distorted by self-interest or irrelevant moral notions. Some objective basis for these judgments is necessary so that widespread agreement to them can be elicited and social discord minimized. At the present time, no satisfactory method for making such objective judgments concerning welfare has been found (90-91, 321-325).

Rawls' theory, on the other hand, measures social expectations in terms of "primary goods," things which are necessary means to the success of one's rational life plan, so that it can be supposed a rational man wants them whatever else he wants. The primary goods identified by Rawls are rights and liberties, opportunities and powers, income and wealth, and, in the context of some issues, a sense of one's own self-worth (92-93). Rawls does not suggest this different index of human welfare reflects a different theory of the good for man; on the contrary, he states that the theory of the good is not in dispute between utilitarianism and the contract doctrine (92). According to both theories, a person's good is determined by his rational long-term plan of life. A person is happy when he is more or less successfully in the way of carrying out his plan. And a person's plan of life is determined by his desires, since the plan is designed to permit the harmonious satisfaction of these desires. Primary goods provide a suitable index of good because they are the necessary means for the accomplishment of this long-term plan, whatever its precise content (92-93).

Rawls' argument that his theory is preferable on this score to utilitarianism rests on the claim that while it is possible for society to arrive at an objective measure of primary goods, it is not possible to measure the satisfaction of desire. Partly because he does not assume this is a *theoretical* impossibility, Rawls does not wish to rest much weight on this objection. Nevertheless he does regard it as a significant difficulty (91, 321).

Of course, for Rawls' theory to compare favorably with utilitarianism, it must in fact be possible to measure the quantity of primary goods in some objective way. Little attention has been paid to this fact, but the availability of such a measure seems questionable. Rawls evidently understands "income and wealth" in terms of monetary sums which are relatively easy to measure. However, in a society such as ours, an important part of one's income (much contested in union contracts) is frequently the "fringe benefits" and "perquisites" which accompany one's salary, such as health insurance, paid vacations, or access to recreational facilities.[24] It may be far more difficult to measure the value of such benefits. Moreover the same income in different societies will have different value. Commodities which are available in exchange for financial considerations in one society will not be so available under different cultural and economic arrangements (for example, human organs for transplant purposes are not typically available through the market in our society, although one can imagine them being so under different circumstances.) And the same income has one value when faced with one price structure in one society and a different value when faced with a different price structure in another society (if you like pears and hate apples, ten dollars is worth less to you in a society where pears cost a dollar apiece while apples are fifty cents than it is in a society where apples cost a dollar and pears are fifth cents apiece).[25] How to measure rights, or liberties, or powers, or opportunities, seems an even more difficult problem.[26]

Rawls discusses whether or not his theory involves a serious "indexing problem," that is, a problem of measuring the value of a quantity of one primary good *as compared to* the value of a quantity of some other primary good. He argues the problem is not serious, for the following reasons. Under the special conception of justice, the first principle concerning liberty and the principle of equal opportunity call for all persons to have the *same* amounts of the *same* liberties and opportunities. Thus we need not consider whether or not, for example, the right to vote is equivalent to the right to pro-

pose candidates for office since everyone will have both rights, or neither one. Moreover, these principles are ranked lexicographically relative to the difference principle, so the primary goods covered by them need not be compared with those covered by it. Thus, states Rawls, the only possible problem arises in the application of the difference principle itself. But in effect its application only requires us to identify the least well-off representative person (or position) and maximize his situation by finding the social scheme which will leave him best off. Everyone else will perforce be better off than he is. In deciding which social system makes the worst-off representative person best off, we need only make ordinal judgments about his situation, not cardinal judgments as is required in utilitarianism. Rawls claims that defining the bottom position will not in practice be a problem, since although in principle primary goods could vary relative to each other, which would make it impossible to define the bottom position without some weighting scheme, in fact the primary goods tend to vary together, so that persons in the better positions tend to have more of *every* good. Thus the indexing problem reduces to the problem of deciding when the worst-off person is best-off for the various possible combinations of primary goods which may define his expectations. According to Rawls, this judgment can be made by taking up the standpoint of the representative individual from this group and asking which combinations of primary goods it would be rational for him to prefer. This involves an unavoidable, but limited, reliance on intuitive judgments, but one which is less egregious than that required by utilitarianism (93-95).

This argument has not been well received by Rawls' commentators. For one thing, the alleged insulation of liberties, rights and opportunities from comparisons with other primary goods only occurs under social conditions when the special conception of justice is in force. Under more primitive conditions, when the general conception obtains, rights, liberties and opportunities *may* be traded for the economic goods, and we need some measure of their value relative to each other. Critics have also found it extremely unlikely that the primary goods will vary with each other (so that a representative person who has more power also has more income), as Rawls suggests, under all possible social arrangements. Moreover, the likelihood that persons will have different degrees of desire for the various primary goods throws into question Rawls' solution to the problem of ascertaining which possible society's worst-off position is best. One group may find increased powers and responsibilities a more effective tool in

pursuing their life plans than increased income while another group may find the opposite. Any society which assumes a universal scale of relative values in determining what arrangements to make for the worst-off group will not be acting with sufficient sensitivity to human variation. Altogether, Rawls' use of the notion of primary goods to solve the problem of interpersonal comparisons of utility is subject to problems almost as severe as those which vitiate the utilitarian account of justice.[27]

There is a related problem which we should note. As remarked above, Rawls states that the theory of the good is not at issue between contract theory and utilitarianism—at bottom they are both concerned with the satisfaction of rational desire (92-93). However, according to Rawls' theory, "justice as fairness...does not look behind the use which persons make of the rights and opportunities available to them in order to measure, much less to maximize, the satisfactions they achieve...once the whole arrangement is set up and going no questions are asked about the total of satisfaction..." (94). One may ask why this failure to enquire into satisfactions is appropriate. An interpretation that has been proposed is that society should concern itself with the distribution of *opportunities*, not the distribution of happiness. What is unjust is not that some people are less happy than others but rather that some people have fewer chances than others. It is up to the individual to decide how to employ his life-chances.[28] This may indeed be Rawls' guiding idea. However, the language in Section 15 suggests another interpretation, namely that the state is genuinely concerned as a matter of justice with the *satisfactions* obtained by individuals; but because of the practical problems in measuring extent of satisfaction, it attempts to affect satisfactions indirectly through the allocation of primary goods which are means towards desire satisfaction. It assumes that "the members of society are rational persons able to adjust their conceptions of the good to their situation" (94), that is, able to maximize their satisfactions given their allotment of primary goods. However, there is no reason to suppose that quantity of primary goods correlates well with degree of desire satisfaction. Although primary goods are *necessary* means to satisfaction of most desires, there is no assurance that they are *sufficient* for the satisfaction of these desires. To show this, Arrow adduces an example of two individuals whose incomes (and presumably other primary goods) are equal, and who therefore qualify as "equally well-off" under Rawls' theory. Yet one of these individuals may be a hemophiliac who requires four thousand dollars a year o

coagulant therapy to achieve a state of security from bleeding at all comparable to that of the other. It is by no means clear that they are equally well off in terms of desire-satisfaction or any other intuitive measurement.[29] In this connection one might also point out that whether or not possession of primary goods assures men of greater success in advancing their ends depends on how society is organized. In our society, if I have a consuming desire to collect Persian rugs, a greater income will assist me in pursuing this interest. But here are other imaginable societies in which Persian rugs are not available on the open market at all, but only (say) passed down through family lines. In such societies, possession of primary goods will be of far less use to me. For these kinds of reasons, we cannot expect any important correlation between primary goods and satisfaction of desires. Of course, any argument that there is an important correlation would require us to measure satisfaction of desires, something which Rawls (and many others) have claimed we cannot do. Thus it appears the employment of primary goods to measure social expectation cannot be justified on the grounds that it gives us an indirect, but practical, way of measuring satisfaction of desires. (If it could be so justified, it would be open to utilitarians to use primary goods as well, so that Rawls' theory would not be superior to theirs in terms of practicality. Rawls himself admits the possibility of so reconstruing utilitarianism (175)). But without this kind of connection, it is unsufficiently clear why primary goods provide a relevant measure of social expectations, particularly in light of Rawls' theory of the good. And as we have seen, even the alleged ease of measuring them is in serious dispute.

The Fourth Extra-Contractarian Argument: The Source and Quality of Desires

The final extra-contractarian argument we shall consider goes as follows. Rawls points out that utilitarianism sets as its goal the maximal satisfaction of people's desires. It makes no judgment about the source of those desires or their content or nature. Thus utilitarianism takes existing desires as given, whatever their characteristics. However, it is well known that political and economic institutions influence the desires of those who live under them—thus it is frequently claimed, for example, that capitalism creates the desire for more material goods. Moreover, some desires may seem morally objectionable in themselves, such as the desire to see members of other

races occupy lower social positions, or the abhorrence of certain sexual or religious practices. But utilitarianism must take all these desires as grist for its mill, as legitimate as any others, so that the course for the state which it charts out will always be affected by the particular character of the time at which its reforms are initiated.

According to Rawls, however, *his* theory enables one to define an "Archimedean point" for assessing social systems without contamination by existing institutions, whatever their nature might be. His principles of justice are chosen by the parties in the original position on the grounds that they will best advance their interests as measured in primary goods. But the primary goods are taken to be things which are wanted as parts of rational plans of life which may include the most varied ends. Thus basing a conception of justice on these goods does not tie it to any particular pattern of interests as these might be generated by an historical arrangement of human institutions. Since the two principles of justice are not contingent on existing desires or social conditions, they define the long range aim of society, regardless of what it is like at the time they are implemented. Moreover, the parties in the original position implicitly agree not to press claims on each other which violate whatever principles of justice are chosen. Thus, since they give liberty lexicographical priority, they agree not to give any weight to any subsequent desires that the liberties of some be restricted. Such desires receive no value at all, unlike what happens under a utilitarian principle of justice (30-32, 258-263, 450).

It is not wholly clear what the nature of Rawls' complaint against utilitarianism in these passages is. There seem to be three separate concerns. One is that utilitarianism's placing equal weight on all desires, whatever their content, will lead to institutions which would be considered unjust by reflective judgment (see 450). For example, not enough assurance would be given of individual freedom. This concern is simply a restating of the reflective equilibrium argument considered in "The First Extra-Contractarian Argument" above rather than an independent line of reasoning. As we saw there, the real degree of difference on this subject between Rawls' theory and utilitarianism is open to question. Powerful utilitarian arguments for liberal institutions are well known (see 209ff.).

A second concern seems to be the thought that some desires in themselves are recognized by reflective moral judgment as immoral, and consequently that utilitarianism goes wrong in allowing these desires the same status as other more innocent desires. However, utilitarians have long argued, with a good deal of plausibility, that no

desire is evil in itself, but only evil insofar as it involves evil consequences. Thus we might normally think of a sadistic desire that others suffer as an evil desire. But on reflection, there seems nothing wrong with the satisfaction of such a desire simply taken *as satisfaction;* the desire is only objectionable because satisfying it necessarily involves the suffering of others, suffering which they desire not to undergo. One of the cases traditionally urged against utilitarianism is that where the sadistic desires of one group (say, the Roman masses) outweigh the desire not to suffer on the part of others (say, the Christian martyrs). However, this sort of infringement on individual freedoms would be objectionable even if the majority's desires were perfectly innocent in character, for example, the desire to see how human beings react under stress.

Although utilitarians have argued that no desire is evil in and of itself, they have also recognized that some desires tend to have worse effects than others because they are incompatible with the satisfaction of opposing desires. They have urged therefore that society encourage the replacement of these desires by other, more harmonious ones which will lead to a greater overall level of satisfaction within society (262). Thus utilitarianism provides us with consequentialist grounds for criticizing certain desires, and plausible grounds at that.

The third concern Rawls expresses in these passages is the concern that utilitarianism does not take into account the *source* of the desires whose satisfaction it seeks to maximize, and in particular disregards the fact that these desires may be the outgrowth of existing institutions. It is difficult to see precisely what is objectionable about this. Rawls may have in mind the fact that these institutions will in some cases be unjust and so give rise to desires which will lead to further unjust institutions, or which are "immoral" in themselves. If so, this point is not distinct from the first two. Or Rawls may be concerned by the fact that social systems tend to breed in their members artificial desires designed to perpetuate the system. Thus many corporations in capitalist economies spend substantial sums on advertising upholding the values of free enterprise. However, this is only objectionable if the resulting attitudes and desires are irrational, that is, based on false beliefs. Since we are concerned with a form of utilitarianism that maximizes the satisfaction of *rational* desires alone, any irrational desires stemming from this source would not be counted in the calculus. Alternatively, Rawls may believe that principles of justice should define a social ideal which is universal, that is, which *any* society should aspire to, whatever its present social and cultural forms.

In any very strong form, such a thesis seems implausible. In our culture, there is great interest in organized sports, and it is prima facie desirable that an ideally just version of this culture would maintain the economic arrangements which make this possible. It would appear that maintaining such arrangements would not necessarily be part of the idealized version of some culture which has no interest in such activities. But if we weaken the thesis to allow room for such diversity, it becomes unclear how the results would diverge from those obtained by applying utilitarianism.

We must conclude that insofar as this final extra-contractarian argument has weight, it is primarily the weight which accrues to it as a version of the argument that utilitarianism fails to accord with our reflective moral judgments. And as we saw before, no conclusive argument has been proposed which shows that utilitarianism fares worse in this arena than Rawls' own theory.

EVALUATION OF RAWLS' CONTRACTARIAN-DEPENDENT ARGUMENTS

In the last four sections, we considered the most salient arguments against utilitarianism that Rawls offers from a perspective which is independent of his contractarian approach. None of these arguments proved to be compelling, and certainly not as demonstrations that Rawls' theory is clearly superior to utilitarianism in the respects at issue. In the following sections we will consider the major arguments Rawls advances against utilitarianism from within the contractarian standpoint. All take the form of attempting to establish that the parties in the original position would choose Rawls' principles of justice, rather than utilitarianism, to govern the basic structure of their society.

To this sort of argument, the defender of utilitarianism has four different kinds of response. He can claim, first of all, that what principles the parties in the original position would select is an entirely irrelevant test for the correctness of a given principle of justice, and therefore that utilitarianism is in no way shown defective by any argument that it would not be so chosen. This is clearly an important strategy, but I shall not explore it here since evaluating it involves a deeper foray into matters of meta-ethics than is appropriate within the confines of this paper. The utilitarian can claim, secondly, that what principles the parties in the original position would choose

is indeed relevant, but that Rawls has failed to describe the correct original position, and that the parties in the correct original position would choose utilitarianism rather than Rawls' principles. Rawls recognizes this line of response when he states that for each traditional conception of justice there is an appropriate original position from which it would be selected, and that one of the objects of his argument is to establish that his original position is the correct one (121, 141). The third response available to the utilitarian is to concede that Rawls' original position is the correct one, but to argue that the parties in it would choose utilitarianism rather than Rawls' principles. The fourth and final possibility is to admit that Rawls' original position is the correct one, but to claim that utilitarianism and Rawls' principles are equivalent for the central range of cases, and therefore that the parties in that original position must be indifferent between them. We shall see each of these strategies adopted in response to one or another of Rawls' arguments.

Rawls' argument that his principles of justice, rather than utilitarianism, would be chosen by the parties in the original position rests on the claim that these parties would adopt a maximin strategy in choosing principles, and that his theory of justice constitutes the maximin solution to their decision problem (152, 175). The maximin strategy of decision-making requires choosing an alternative whose worst possible outcome is better than the worst possible outcome of any other alternative. It is the strategy one would select if one knew one's enemy were to determine which outcome would come about. Rawls argues that his principles of justice are the appropriate ones to select if the parties in the original position are trying to maximin, since these principles, in effect, only condone societies which maximize the position of the least well-off group. Thus if one assumed one would be assigned one's place in society by one's enemy, it would be rational to select Rawls' principles to govern that society, for then one would be assured that one's expectations would be as high as possible for that society (152-153).

Rawls is at pains to point out that the parties in the original position are not to make the false assumption that their enemies will assign them their places, and that in general the maximin strategy is not an adequate guide to decision-making (153). The question, then, is why it is appropriate for the parties in the original position to employ it. Rawls has two answers to this question. One is that there are three features of decision-situations which are generally recognized to call for employment of the maximin rule, and that the situation

of the parties in the original position manifests all three of these features to a high degree. The second is that the strains of commitment which accompany any public conception of justice and the necessity for psychological stability offer special reasons for the parties to select his principles of justice. We shall look at these in turn and finally shall examine Rawls' criticism of Harsanyi's attempt to derive utilitarianism from a variant on Rawls' own original position.

The First Feature: No Knowledge of Probabilities

Since the maximin rule takes no account of the likelihoods of the possible outcomes of choice, use of the rule is much more plausible if there is some reason for sharply discounting estimates of the probable consequences of one's choices (154).[30] Rawls argues that there is good reason in the original position to discount such estimates. The parties in the original position not only lack all knowledge about themselves as individuals or their place in society, but they also lack any knowledge of the course of history or how often society has taken one form or another (200). In such circumstances, Rawls claims, estimates of probabilities cannot be objective or based on knowledge of particular facts (172-173). The only ground on which the parties might make probability estimates is appeal to the Principle of Insufficient Reason, a principle which directs the decision-maker, when he cannot assign probabilities on actual evidence, to identify the possibilities in some natural way and then assume that each is equally likely. But Rawls argues that it is inappropriate to make use of this principle in the original position because the decision is of such fundamental importance, and because one would desire to have one's decision appear responsible to one's descendants who would be affected by it (169). Thus the parties have "no basis for determining the probable nature of their society, or their place in it" (155). Without probability estimates, they must make use of a decision-principle which does not take probabilities into account such as the maximin rule.

This argument has perhaps attracted more attention from Rawls' critics than any other. The grounds for their objections are diverse. First we might note that even if it is agreed that the parties must employ some decision-rule which takes no account of probabilities, it doesn't follow that they should employ the maximin rule. The maxi-

min rule is only one member of an entire family of decision-principles which take no account of probabilities, and the fact that *some* member of this family must be used hardly shows that maximin in particular must be.[31]

Second, as some critics have pointed out, there is something slightly bizarre about Rawls' arguing that the knowledge of the parties in the original position is insufficient to allow them to make probability estimates. The argument that they cannot make such estimates relies on the *stipulation* that they possess no knowledge of the course of history or the frequency with which society assumes various forms. But this stipulation seems to have no independent rationale. Unlike the stipulation that they possess no knowledge of themselves as individuals, it is not needed to prevent them from "tailoring principles to their own circumstances" (139). Unlike the stipulation that they have no knowledge of the particular details of their own society, it is not needed to secure proper justice between generations (137). It does not seem necessary to secure unanimous agreement on principles (140). Thus its *only* apparent role is to prevent them from making probability estimates of the various possible outcomes, despite Rawls' statement that all features of the original position are to be "natural and plausible" (18). The same comments apply to the parties' inability to estimate the probability of their turning out to be any given individual in a society. The parties' transformation into members of society is not a natural process. a matter of fact about which they could have ordinary evidence. Rather it is a fictitious event whose nature is wholly governed by Rawls' stipulations. His failure to stipulate what the probabilities are, and to allow the parties to know what these probabilities are, needs explanation, and the explanation seems simply to be the desire to secure their selection of his principles of justice. But if this is the case, then Rawls could have secured the same result more straightforwardly by simply stipulating that the parties in the original position are not to make probability estimates. It would then have been clearer that the only rationale for this feature of the original position is that it enables us to derive principles which Rawls finds to be in accord with our considered moral judgments, and in particular to avoid utilitarianism.[32]

Some critics have argued that the situation may be even worse than this discussion reveals, for they contend that independent considerations show that the parties in the original position *must* assume they have an equal probability of being any person in the society for which they are choosing principles of justice. Thus Harsanyi has suggested that the equiprobability assumption is not to be interpreted as the result of using the Principle of Insufficient Reason, but rather a feature designed into the original position which reflects our normal moral assumption that the interests of each person in society are to be given equal weight—as opposed to giving more weight to the interests of the poor.[33] Somewhat along the same lines, Narveson has argued that it is one of our pre-analytic precepts of fairness that rules are fair only when they do not load the dice in favor of anyone, whether rich or poor, and that this is what the equiprobability assumption amounts to.[34] We can certainly agree that it would be coherent for a theorist to design an original position to include an equiprobability assumption for this reason.

It might be noted, incidentally, that Rawls' specific arguments for rejecting the use of the Principle of Insufficient Reason in the original position are not [both] equally compelling. The first argument, that the decision is so important, has found many adherents.[35] But the second one, that abjuring use of the Principle allows the parties to defend their decision to their descendants who will be affected by it, seems off the mark. First, as Hare has noted, it is not at all clear that the descendants will thank their parents for being so conservative; they might well respond, "Nothing ventured, nothing gained."[36] But more telling is the fact that the descendants themselves are potential parties in the original position; by hypothesis, if the parents adopt a certain principle of justice in that situation, the children will also. And it would hardly seem rational for the children to reproach their parents for choosing a principle affecting them when they would choose the very same principle on their own behalf. Thus possible reproach from one's descendants does not seem to be a factor which the parties in the original position need to worry about.

We have seen in this section that Rawls' argument for the parties adopting the maximin strategy because they lack the knowledge to make probability estimates has no independent grounds, beyond the fact that adopting this strategy leads to a principle of justice which allegedly accords with our reflective moral judgments better than utilitarianism. And we have seen reason to question this latter claim.

The Second Feature: A Guaranteed Minimum

Use of the maximin strategy in the original position is plausible, Rawls argues, because of a second feature of that situation: the fact that the parties in it know their conception of the good is such that they

care very little, if anything, for what they might gain above the minimum stipend they can be sure of by following the maximin rule. Since this is so, Rawls claims, there is little reason for them to try to do better, for example by following the rule of maximizing expected utility (154–155).

This argument relies on the assumption that the principles of justice selected by following the maximin strategy, namely Rawls' principles, will ensure a socially acceptable minimum which no one will care greatly about going beyond. Rawls argues that his principles will guarantee this minimum because they provide a workable theory of justice and are compatible with reasonable demands of efficiency (156). However, as a number of critics have pointed out, this argument is dubious.[37] Whether or not such a minimum is achievable depends on the natural resources available to the society, the health of its members, its relations with other societies, and other matters which are largely unaffected by the principle of justice which governs the society. In this connection it should be remembered that Rawls' *general* conception of justice, in addition to his special conception, represents a maximin solution to the choice in the original position.[38] But the general conception is explicitly designed to apply to situations of extreme poverty and social underdevelopment, situations where establishment of a minimum beyond which no one is much interested in going seems ruled out by definition. Some critics have pointed out that indeed the satiation point for wealth and power is higher for most people than the minimum obtainable by Rawlsian principles in even the richest societies.[39] Barry pursues this point by arguing, persuasively, that if the minimum level achievable by maximin policies in a society falls either below or above the threshold of satiation, there is little reason to think, merely on the basis that such a threshold exists, that the maximin policy is obviously the right solution. For example, if apples are to be distributed among ten people whose Rawlsian threshold level is twelve apples, and we can choose between giving everyone ten apples, or giving nine of them twelve apples and the remaining person nine apples, it is not at all clear that the second division isn't better.[40]

Last, it should be pointed out that if there is such a threshold point, and if it is the same for everyone, as Rawls implicitly assumes, then this means that the marginal utility of primary goods effectively diminishes to zero after this point. If this is true, then in all probability utilitarianism would mandate the same institutions as Rawls' principles of justice, since it would increase average utility to distribute

goods so as to raise everyone up to the threshold point, rather than giving fewer to some and more to others. If so, utilitarianism satisfies the second demand as well as Rawls' principles of justice do.[41]

We can see, then, that the "second feature" of the original position involves an implausible assumption about human good whose truth would not support Rawls' conception of justice more strongly than it would utilitarianism.

The Third Feature: Avoidance of Intolerable Outcomes

Rawls cites a third feature of the original position which makes it rational for the parties to employ the maximin strategy: use of other choice rules (for example, the rule of maximizing expected utility) would lead to conceptions of justice (for example, utilitarianism) which would permit intolerable institutions (such as slavery and serfdom) which the parties could hardly accept (156).

There are three problems with this argument. First, the desire to avoid intolerable outcomes does not in itself require the maximin strategy. It requires what Hare calls an "insurance" strategy, one which guarantees avoidance of unacceptable outcomes. If the maximin strategy could ensure acceptable outcomes, then it would qualify as an insurance strategy. However, at best it is only one of several possible insurance strategies because it goes beyond what is required of a policy to qualify as an insurance strategy. In particular, it dictates what social arrangements must be selected even in a rich society where the minimum level obtainable is far above the point of toleration. Thus the "intolerable outcome" argument at most gives us reason to think the maximin strategy satisfies a necessary condition for being adopted, but not a sufficient condition.[42] Second, as the argument in the last section allows us to infer, there is no good reason to think that adopting the maximin strategy will actually avoid intolerable outcomes, especially at the low levels of social advancement which call for Rawls' general conception of justice. Finally, as we have already seen in "The First Extra-Contractarian Argument" above, it is far from clear that utilitarianism at any rate would lead to intolerable institutions under conditions when Rawls' principles would not. If slavery and serfdom are genuinely intolerable, then it appears their disutility is so great that no social system permitting them (except under very severe conditions) would maximize average utility. This means that utilitarianism may provide as good "insur-

ance" against these institutions as Rawls' principles of justice do, for the two conceptions of justice appear to rule out intolerable institutions in roughly the same range of conditions.

We can see, then, that the desire to avoid intolerable outcomes provides an insufficient argument in support of the maximin strategy, and that there is reason to suppose utilitarianism may satisfy this desire at least as well as Rawls' principles of justice do.

The Strains of Commitment: Psychological Stability

In the last three sections, we have seen Rawls' argument for claiming that the parties in the original position would be rational to employ the maximin strategy in selecting their principles of justice. The case appears far from conclusive. Inadequate support is offered for the inability of the parties to make probability estimates; and the putative existence of "satiation threshold" and "intolerable outcome" points does not show that maximin must be followed, since it may not achieve the former or avoid the latter, and other strategies might work as well. In addition we have seen evidence that utilitarianism might succeed as well as Rawls' principles in guaranteeing achievement of the satiation threshold and avoidance of intolerable institutions. However, Rawls adduces several other arguments to show that his principles, rather than utilitarianism, are the genuine maximin solution to the problem in the original position. We shall consider these in this section, the arguments from what Rawls calls the "strains of commitment," and "psychological stability."

The reasoning regarding the strains of commitment appears to go as follows. In selecting a principle of justice which will satisfy maximin, the parties assume their society will act in "strict compliance" with that principle. That is, they assume everyone will accept the principle, and know that the others accept the principle, and also assume that the basic social arrangements will satisfy and be known to satisfy the principle (8, 145, 454). But they are not to assume the impossible. If members of society would not be able to honor a given principle of justice under all circumstances, even the most onerous, then the principle is disqualified as one they may select (175–176). Rawls argues that his special conception of justice has an advantage over utilitarianism in this regard, because utilitarianism, unlike his principle, may require people to sacrifice their freedoms for the sake of greater good for others (176–177). We may grant that it would be

psychologically possible for members of society to honor Rawls' special conception.[43] However, as we pointed out above, the empirical assumptions Rawls employs to argue for the priority of liberty under the special conception show that in circumstances where it would apply, utilitarianism would *also* refuse to trade the liberties of anyone for mere economic goods accruing to someone else. Thus honoring utilitarianism would be no more difficult, in these circumstances and for this reason, than honoring Rawls' special conception. We have also seen that Rawls' general conception of justice allows violations of liberties in some cases, and we have no conclusive argument that utilitarianism would do worse. We cannot conclude, then, at least on Rawls' own empirical assumptions, that it would be psychologically impossible to honor utilitarianism.

Rawls' reasoning concerning the psychological stability of a conception of justice has much the same flavor. The parties in the original position assume that they are choosing a conception of justice with which their society will strictly comply, in the sense explained above. However, society is an enduring entity, and strict compliance of the sort contemplated will only be achieved at the cost of social practices designed to maintain the relevant sense of justice in the members of society. The principles of justice must be promulgated and enforced; people must be trained to believe in those principles and to feel guilty when they violate them. The level and type of motivation necessary to maintain strict compliance may vary, depending on the content of the different principles of justice. In assessing a particular conception of justice, the parties in the original position must therefore take into account not only the effects of institutions which comply with that conception, but also the burden that maintaining compliance imposes on society. Rawls calls a conception of justice "stable" when public recognition of its realization by the social system tends to bring about the corresponding sense of justice, i.e., tendency to judge in accordance with the principles of that conception. Obviously, the more stable a conception is, the less burdensome the social cost of maintaining it (46, 177–183, § § 69, 76).

At some points in his discussion, Rawls suggests that different conceptions of justice must be compared with each other with respect to these burdens, and that a less burdensome conception is to be preferred, other things being equal (455, 498). At other points, he seems to suggest merely that an acceptable conception of justice must be *stable enough* (504). However, the structure of his argument implies that his position ought to be the following. The parties in the original

position, according to him, must employ a maximin strategy, that is, select that principle of justice whose worst possible outcome is better than the worst possible outcome of any alternative principle. But the level of well-being of the worst-off individual in society depends not only, for example, on the economic arrangements mandated by the conception of justice but also on the social practices which are necessary to sustain compliance with that conception. Two different conceptions of justice might sanction precisely the same economic and political arrangements but differ from each other according to the ease with which allegiance to the conception is elicited, and therefore with respect to the amount of social resources which must be devoted to maintaining compliance. The naturally more attractive conception would then guarantee a *better* worst outcome. Therefore the parties must pay attention to the *relative* stability of the conceptions of justice they consider, for this affects what their expectations under these conceptions would be.

Let us look, then, at a simplified version of the complex argument to show that Rawls' conception of justice would be less burdensome, or more stable, than utilitarianism. Rawls points out that any conception of justice requires an individual to perform some acts which are not in accord with his self-interest, narrowly conceived. Therefore strict compliance can only be maintained if the pressure of self-interest is adequately offset, for example, by an opposing sense of justice, or a concern for the welfare of others (454, 497). He believes that such a sense of justice can be produced by two circumstances: (a) knowledge that the institutions satisfying that conception enhance one's own good—knowledge which enhances one's self-esteem, and creates the tendency to cherish and support those institutions and the governing conception of justice, and (b) thorough understanding both of the precepts of the governing conception of justice, and of the reasoning which supports it (177, 498–499).[44] He argues that his conception of justice would create both circumstances, and so produce a strong corresponding sense of justice. His principles prohibit forced sacrifice of one citizen's good for that of others, and they require that institutions be established from which everyone benefits (§§ 29, 76). Thus social arrangements satisfying these principles enhance the good of each member of society and would be known to do so. Moreover, he argues, his conception of justice is clear enough so that it is easy to understand and apply, largely because of its use of primary goods in measuring social expectations; in addition the reasons given for it are easily understood and accepted (§ 49, 501).

By contrast, Rawls argues, the institutions governed by utilitarianism may require us to make sacrifices, even of our liberties, in order to increase average welfare. Thus they need not enhance our good as individuals and so will not tend to produce the corresponding sense of justice throughout society. Compliance can then only be produced by inducing people to identify strongly with the interests of others, but this is not easy to bring about. Moreover, utilitarianism is difficult to understand and apply because of the problem presented by the necessity for making interpersonal comparisons of utility (§§ 15, 49, 76). Rawls concludes that the contract view offers greater stability (501).

It is difficult to assess an argument such as this which relies so heavily on empirical hypotheses which have received inadequate testing. I will confine myself to making three points. First, as we have already seen in "The First Extra-Contractarian Argument" above, there is room for disagreement with Rawls' claim that his principles "benefit everyone," whereas utilitarianism alone requires sacrifices of the interests of some for the good of others. According to Rawls, his principles benefit everyone primarily in the sense that each person (or representative group) would do better, if the principles govern society, than he or his group would have done if the primary goods had been divided equally (80).[45] We might grant that this is so. However, the question before us is whether or not persons living in a society governed by utilitarianism, or by Rawls' principles, would *feel* that anyone was being required to make sacrifices. As we noted before, someone who regards a given person as making sacrifices evaluates the position of that person relative to what minimal justice or morality requires of him. Thus to know if members of society would believe sacrifices were being made, we must know what *they* would believe justice requires of the members of society. One possibility of course is that they simply believe justice requires what the principle governing their society requires. If so, then obviously no one living under either utilitarianism or Rawls' principles would believe sacrifices were being made. The two conceptions of justice would be on an equal footing. Another possibility is that there is some independent, "natural," conception of justice which people tend to adhere to no matter what principles govern their society. If so, then it is possible that *either* utilitarianism *or* Rawls' principles, or both, require sacrifices relative to what that conception requires. If this "natural" conception calls for "equal shares for all," as perhaps Rawls assumes, then his principles would not call for sacrifices whereas utilitarianism

might well. But there is no reason in advance to suppose the "natural" conception does call for equal shares. Certainly many in our society do not believe this and would believe that Rawlsian justice requires sacrifices by those who are better off for the sake of those who are worse off. Which conception of justice calls for greater sacrifices, and so is less stable, must be left an open question whose answer depends on resolution of the standard of "sacrifice" to be used and on empirical facts concerning human psychology.[46]

Second, it is far from clear that Rawls' principles have an advantage with respect to clarity and ease of application since as we saw before, the problem of indexing primary goods is in principle as difficult as the problem of making interpersonal comparisons of utility for utilitarianism. Moreover, the reasoning in favor of Rawls' principles is quite complex whereas some forms of reasoning in favor of utilitarianism have been quite simple (for example, Smart simply argues that utilitarianism is the principle that would be adopted by a perfectly benevolent person).[47]

Third, we have seen that it may be argued with some plausibility that utilitarianism would actually sanction much the same social arrangements as Rawls' principles. If so, it might be urged that utilitarianism would have the same effect on each person's good that the contract view would, hence would cause people to cherish just institutions to an equivalent degree, and hence develop an equally strong sense of justice. However, Rawls proposes a response to this line of thought. He points out that the affection elicited by a conception of justice depends not only on the consequences of the social arrangements which satisfy that conception, but also on the values which are expressed in the way the conception is stated. In particular he claims that the formulation of his conception of justice expresses respect for each member of society. According to him, expression by others of respect for oneself promotes one's own self-respect, and self-respect on the part of members of society increases the efficacy of social cooperation. Thus he claims that even if his principles mandate the same institutional arrangements as utilitarianism, public acceptance of his principles would increase the amount of self-respect in society, so make it more efficient, and so raise the expectations of every member of society relative to what they would be under utilitarianism. He concludes that the parties in the original position have reason to choose his conception of justice rather than utilitarianism (178–183).

According to Rawls, his conception of justice expresses respect for each member of society because it includes everyone's good in a scheme of mutual benefit, so that public affirmation of this scheme affirms the worth of each person's life plan (178–179). However, we have already seen it is far from clear that persons living under Rawlsian justice would view institutions satisfying his conception as benefiting everyone. They might well feel that those higher up on the social scale were being asked to sacrifice their life plans for those lower down. Thus we cannot be sure that his conception of justice would really promote universal self-respect in the manner described.

Rawls offers a variant on this argument. He suggests in another context that his conception of justice has an advantage over utilitarianism because members of a utilitarian society would constantly be aware that the guarantee of their civil liberties rests on certain assumptions concerning empirical facts, assumptions which may at any time be found erroneous, and facts which may change with the alteration of social conditions. Thus civil liberties are not assured once and for all, and the members of society may feel insecure in this knowledge. Under Rawlsian justice, it is claimed on the contrary that the priority of liberty is built right into the principles and cannot be taken away even if new facts come to light (159–161).

We may concede that members of a utilitarian society would necessarily be aware that their liberties rest on empirical assumptions about the importance of liberty, and so are not immune to change. Thus there is a built-in uncertainty which might have psychological cost. (One might ask whether this will necessarily be a *cost*, however. Why should the members of society be so concerned about the possible loss of something which would only be removed on the finding that they care rather little for it?) However, in all probability, as we have seen before, a situation in which utilitarianism does not support equal liberties for all is also a situation in which Rawls' special conception of justice, which grants liberties a special place, would no longer be appropriate. The special conception rests on empirical assumptions just as the derivation of civil liberties from utilitarianism rests on empirical assumptions. Members of a society governed by the special conception must recognize that these assumptions may prove false, or that the facts may change. What are they to imagine their options in such a situation would be? One possibility is that nothing could be done, i.e., that they are stuck with the special conception of justice and the priority of liberty, even if it turns out for

example that they prefer wealth to liberty. It would appear that knowing one lived under a conception of justice which could not be overturned, even if the empirical assumptions on which it rests turn out to be false, would involve just as high a psychological burden as that envisioned for members of a utilitarian society. A second possibility Rawls could allow is that members of this society would know that if the facts which mandate application of the special conception fail to obtain, then their society would shift to the general conception of justice. Thus they need not fear being stuck with an unwanted priority of liberty. However, this scenario would mean that they could not feel, any more than members of a utilitarian society could, that their civil liberties are guaranteed. If this knowledge in itself is costly, members of Rawlsian society must bear it as well. Moreover they must bear an additional cost. Members of a utilitarian society will know that their principle of justice is sensitive to *any* change in personal values. Members in a Rawlsian society cannot rest assured of this, for their society has only two options, the general or the special conceptions of justice. Society must accept one or the other of these even though there are possible, indeed probable, preference structures on the part of society's members which would make neither one appropriate (for example, they might rank wealth lexicographically relative to liberty). The moral to be drawn here is that any conception of justice which is not directly committed (rather than committed only on the basis of possibly false empirical assumptions) to the fundamental values of members of society will necessarily entail some psychological cost for those who live under it. But Rawls has hardly shown conclusively that the cost to be borne in a utilitarian society would be greater than that borne in a Rawlsian society.

Harsanyi's Argument for Utilitarianism from an Original Position

We argued in "The First Feature" above that the desire to avoid utilitarianism was Rawls' main ground for denying that the parties in the original position have knowledge of the course of history or (more importantly) of the probability of their being a given individual in the society they enter. Those who do not find utilitarianism objectionable will not of course be moved by this consideration. By allowing the parties in the original position knowledge that they have an equal chance of becoming any member of society, Harsanyi and Vickrey have developed an argument to show that utilitarianism,

rather than Rawls' conception of justice, would be chosen.[48] Let us look at this argument (the first detailed *positive* reasoning for utilitarianism we have seen so far), and Rawls' response to it.

Stated simply, the argument proceeds as follows. Suppose one were in the original position and knew that one was going to become a member of a given society about which one knew almost all the relevant details: the social structure of the society, the expectations of its members, their number, and their preference orderings. One does not know, however, which member of society one will turn out to be, although one knows that one has an equal chance of being any given member. Most theorists, including Rawls, agree that when probabilities can be taken into account, it is rational to choose the alternative which maximizes one's expected utility. Thus persons in this version of the original position would follow a strategy of maximizing their expected utility, and this dictates choosing a principle of justice whose public recognition would maximize average utility in the society they will enter. (If the levels of welfare of the members of that society are u_1, u_2, u_3, ..., u_n, then the total utility of the society would be the sum of these or Σu_i, and the average utility would be the total utility divided by the number of members of society or $\Sigma u_i/n$. Assuming that one has an equal chance of being any member of that society, the probability of being a particular member is $1/n$. This allows us to arrive at one's expected prospect for being in that society by weighting the utility of being a particular member (e.g., u_1) by the probability of being that member ($1/n$) and summing the results $1/n(u_1) + 1/n(u_2) + 1/n(u_3) + \ldots + 1/n(u_n)$ to get $\Sigma u_i/n$. Since this figure just *is* the average utility of society, one's prospect is equal to that average (165)). This is true for *any* society one might enter, whatever the structure, prospects, number, or preferences of its members. Thus if one did not know anything about the society one was to enter, except that one had an equal chance of being any member of it, and one also knew that a given principle of justice would maximize average utility in every society, then one would select that principle of justice. It maximizes one's expectations, even when one does not know which society one will be a member of. It is then assumed that the principle of maximizing average utility, when strictly complied with, actually succeeds in maximizing utility in each society. Consequently it is the one, not Rawls' conception of justice, which the parties in the original position ought to select.

Most of Rawls' responses to Harsanyi's proposal have already been examined above under other topics. However, he suggests, and David

Gauthier elaborates on, an argument which is specific to the debate with Harsanyi. Rawls' and Gauthier's reasoning is as follows.[49] Let us take it for granted, as Harsanyi must, that it is possible to make interpersonal comparisons of utility. Imagine, now, an individual Jones whose society I may possibly enter. Jones derives a utility of ten from experiencing an hour's worth of pleasure. Of course, ten utiles for him is worth ten utiles for me. However, it does not follow that I also assign a utility of ten to being Jones and experiencing an hour's worth of that same pleasure. My tastes—that is, my utility function—may differ, and that hour may only be worth eight utiles to me. In light of this, let us ask what utility I should assign to the prospect of entering Jones's society as Jones and experiencing an hour's worth of this pleasure. One might suppose, and Harsanyi does suppose, that the utility for me of being Jones during this hour is ten utiles. But in fact (the argument continues) the utility *for me* of being Jones and experiencing his pleasure is only *eight* utiles—since I am concerned with the utility *for me now*, in the original position, not the utility for me after I have become Jones. It is claimed this fact invalidates Harsanyi's argument, for his argument depends on the identification of my utility for being various members of society with their own utility for being themselves. Since the identification is incorrect, Harsanyi's argument does not succeed. Rawls and Gauthier conclude we cannot assume, then, that a principle of justice which maximizes average utility in a society will necessarily maximize the expected utility of a person in the original position who will enter that society as one of its members.

It is dubious that this objection to Harsanyi succeeds. First, as stated, it assumes that the parties in the original position know their own utility functions. However, Harsanyi cannot allow this. If a party in the original position knows what his utility function is, then he will know that he cannot turn out to be any member of society whose utility function is different from his own. If this knowledge is allowed, the interests of all members of society will not be equally taken into account. Of course, even if a person in the original position is not allowed to know his own utility function, he may know that it differs from that of some members of the society, for he knows that his is identical with that of one member of society, and he may know that the members of society have different utility functions from each other. But he must think about the situation as follows. When he tries to assess the utility to him of being Jones, he knows that if he *is* Jones, then his present utility function is identical with that of

Jones. Thus his utility for being Jones is the same as Jones' utility for being Jones. And this is true for every other member of society as well, even though their utility functions may differ. Consequently, his expected utility for entering that society is equal to the average utility level in the society, just as Harsanyi's argument requires, even though the expected utility figure is computed on the basis of different utility functions. There seems to be nothing in principle faulty about such a procedure, despite Rawls' hesitation on this score (175).

Even if we had to assume for some reason that the person in the original position must view himself as being *transformed* into a member of society who may have a different utility function from his own (unknown) one, it is still not obvious that it is illegitimate for him to evaluate the life of a given member of society in terms of that member's own utility function. We might understand the case as parallel to the following one. Suppose you must choose whether to enroll in a graduate program in philosophy or one in business. If you do the former, you will be poor but famous. If you do the latter, you will be rich but unknown. Right now you prefer being famous to being rich. However, whichever course of action you choose, you will undergo a character transformation and come to prefer being rich to being famous. Which of these three utility functions should be taken into account in your decision whether to enroll in the philosophy or the business program? A simple egoism-of-the-moment would dictate that you take into account only your *present* preferences, and so go into philosophy. However, it seems wholly appropriate to take into account instead the preferences of the future philosopher or businessperson, since they, as future aspects of yourself, will be the ones to actually lead and endure the lives in question. We might now transfer this solution to Harsanyi's original position since the time interval itself seems to make no difference. Thus even if the person in the original position knows that his utility function may be different from that of the member of society he will become, it may be appropriate for him to evaluate the life of that member in accord with the preferences of the member himself since he is the one who will actually live the life. He should not be concerned about the fact that, as chooser in the original position, he might "now" have a different utility function. It is unclear whether or not this appropriateness is a matter of rational prudence or rather a matter of moral principle. If it is the latter, then the parties in the original position, who are not themselves moved by moral considerations, cannot themselves argue on moral grounds that they should use the utility functions of

the members of society. However, it is open to Harsanyi to stipulate it as one of the conditions of the original position that they so judge the lives of the members of society, in order to ensure the fairness of the decision, just as he stipulates that they judge themselves to have an equal chance of becoming any member of society.

The argument from different utility functions therefore does not appear to succeed as stated. However, it should be noted that Harsanyi's whole enterprise depends on the posssibility of making appropriate interpersonal comparisons of utility. Harsanyi's own suggestion for how this may be done appears to be undercut by the possibility that two individuals may have different utility functions for the very same experiences.[50] Until some more satisfactory method has been arrived at for making such comparisons, Harsanyi's argument remains at best problematic. Thus we cannot say that the original-position argument in favor of utilitarianism has been made completely compelling.

SUMMARY AND CONCLUSIONS

Let us summarize, although briefly, what has been discovered in the course of our examination of Rawls' arguments against utilitarianism. These arguments were divided into two categories, those which are independent of the contractarian standpoint, and those which depend on it. Among the former, the most important is the argument that utilitarianism accords less well with our considered moral judgments than does Rawls' own conception of justice, and in particular that utilitarianism violates our conviction that the liberties of some may not be limited so that others may benefit economically. We saw that the actual extent to which utilitarianism violates common egalitarian precepts is unclear because the empirical facts necessary to determine this remain unknown. But we also saw that Rawls' own assumptions about people's utility functions for such goods as income do not leave him in a strong position to press this point. We saw in addition that if Rawls' assumption about our relative preference for liberty is correct, then utilitarianism and his special conception may treat the protection of civil liberties in very much the same way, since it would not in realistic situations promote utility to trade the liberties of one person for an increase in economic goods for others. The contrast between Rawls' general conception and utilitarianism with regard to liberty must also remain unsettled, because of

our ignorance of relevant empirical matters; but Rawls himself offers no reason to suppose utilitarianism would violate common precepts in any worse fashion than his general conception. The other extra-contractarian arguments he offers were also questioned: the complaint that utilitarianism illegitimately merges persons, and the complaint that it improperly disregards the source of the desires whose satisfaction is to be maximized, seemed unwarranted in the main part. The objection that utilitarianism has a serious problem in making interpersonal comparisons of utility was granted as a debilitating difficulty. But it was also pointed out that Rawls' own reliance on primary social goods to measure social expectations is subject to difficulties which in principle are just as grave. Thus on the whole we must conclude that Rawls' extra-contractarian arguments fail to show that utilitarianism is unacceptable—at least in comparison to Rawls' own principles, and judged on the basis of the empirical assumptions Rawls makes in arguing for his own conception of justice.

The arguments from within the contractarian standpoint do not fare significantly better. The three features of the original position which Rawls adduces to show that the parties in it must employ a maximin strategy in choosing a principle of justice to govern their society, fail to show that maximin, rather than some related strategy, must be used. It seems dubious that all of these features would hold true in an empirically plausible original position, and one feature at least seems to be a restriction added solely to avoid the derivation of utilitarianism from the original position. Given Rawls' empirical assumptions, it cannot clearly be argued that his principles of justice, rather than utilitarianism, constitute the unique best choice for the parties, even if these three features are all present. Rawls argues in addition that the parties in the original position would know they would be unable as members of society to adhere to the requirements of utilitarian justice, but we saw that utilitarianism rules out slavery and other forms of servitude in the same circumstances that Rawls' own principles do, and hence that there is no more reason to suppose it could not be adhered to. The argument from stability has serious flaws as well: it was shown that Rawls' principles would probably be seen as calling for sacrifices, just as utilitarianism might be, and hence that both would have some trouble eliciting popular support. It was also argued that utilitarianism is not significantly more difficult to understand or argue for than Rawls' principles, and so not less likely to command allegiance on those grounds. In addition we questioned whether Rawls' principles would be psychologically more reassuring

than utilitarianism for members of society. Finally, we inspected Rawls' and Gauthier's argument against Harsanyi's attempt to derive utilitarianism from an original position, and saw that it seemed answerable, even though Harsanyi's own project is handicapped by lack of an acceptable method for making interpersonal comparisons of utility.

Rawls' book has revolutionized contemporary discussions of political theory; it seems likely to be the most profound work in the field to be published in this century. His extended argument against utilitarianism has raised important new issues and forced us to examine the old ones from a new perspective. But contrary to what many critics have supposed, the argument does not appear to ring the deathknell for utilitarianism; defenders of that theory, and those of contractarianism, must feel that the battle is far from over.[51]

NOTES

1. See Rawls (22-23n.); "Utilitarianism" *The Encyclopedia of Philosophy* (1967), pp. 7 and 8, 206-212, Dan W. Brock, "Recent Work in Utilitarianism," *American Philosophical Review* 10, no. 4 (October 1973): 241-276; and Samuel Gorovitz, ed., *Mill: Utilitarianism* (Indianapolis: The Bobbs-Merrill Company, Inc., 1971).

   Some utilitarians believe that the welfare of *all* sentient creatures, not just human beings, should be taken into account, but I shall not pursue that line here.
2. William K. Frankena, *Ethics*, 2nd ed. (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1973), p. 52.

   It should be pointed out that some "ideal" utilitarians *do* face a similar problem, for they believe that disparate kinds of things must all be recognized as good in themselves and balanced against each other.
3. A.M. MacLeod, "Critical Notice of Rawls' Theory of Justice," *Dialogue*, March 1974, p. 158, as quoted in Jan Narveson, "Rawls and Utilitarianism," unpublished paper presented at the Conference on the Limits of Utilitarianism held at Virginia Polytechnic Institute and State University, May 18-21, 1978. See also Joel Feinberg, "Rawls and Intuitionism," in Norman Daniels, ed., *Reading Rawls* (New York: Basic Books, Inc.), p. 116.
4. See Rawls (302-303) for the final statements of these principles.
5. See Feinberg, "Rawls and Intuitionism," pp. 108-116 for an illuminating discussion of the relations between justice, overall rightness, and utilitarianism.
6. Strictly speaking, his argument for the average view makes use of the original position perspective, but I shall assume it has general application.

7. See Richard Brandt, *A Theory of the Good and the Right*, Chapters 13 and 16 for useful discussion of the problems with the desire-satisfaction theory, and for an innovative account of rational desire.
8. On Rawls' version of utilitarianism, the welfare to be taken into account is restricted to that experienced by members of the society in question, rather than all of humanity, although the latter would be more in the spirit of classical utilitarianism (22). This makes no difference to the points I shall discuss, but merits further investigation.
9. Rawls also invokes the "Kantian conception" as a justification for the contractarian approach, but the gap between this conception and utilitarianism is too large to attempt dealing with in this paper.
10. See Marshall Cohen, "Review of a Theory of Justice," *New York Times Book Review* (16 July 1972), pp. 1, 16, and Hugo Bedau, "Founding Righteousness on Reason," *The Nation* (11 September 1972), pp. 180-181, as quoted in David Lyons, "Nature and Soundness of the Contract and Coherence Arguments" in Daniels, *Reading Rawls*, p. 143.
11. Peter Singer, "Sidgwick and Reflective Equilibrium," *Monist* 58 (1974): 490-517.
12. Lyons, "Nature and Soundness," p. 148.
13. See Rawls (28). Allan Gibbard discusses this theory of ordinary moral judgments, as Sidgwick states it, in "If the Morality of Common Sense is Unconsciously Utilitarian, Does that Give Us Any Reason to be Utilitarians?" unpublished paper presented at the utilitarianism conference at Virginia Polytechnic Institute and State University, May 18-21, 1978.
14. For further discussion on the role of moral intuitions, see Brandt, *A Theory of the Good and the Right*, Chapter 1, and Ronald Dworkin, "The Original Position," in Daniels, *Reading Rawls*, pp. 27-37.
15. This was pointed out to me by Allan Gibbard.
16. John Rawls, "Justice as Fairness," in Wilfrid Sellars and John Hospers, eds., 2nd ed., *Readings in Ethical Theory* (New York: Appleton-Century-Crofts, 1970), p. 592.
17. See Lyons, "Nature and Soundness," p. 148.
18. Rawls actually admits that the special conception itself would admit slavery when it constitutes an improvement over a current unjust practice, for example if enslavement of prisoners of war were substituted for their automatic execution (248). Unfortunately it is unclear from his discussion whether he intends this as an application of the special conception in nonstrict compliance theory, and so not strictly relevant to its adequacy in a well-ordered society, or intends it to fall under the strict-compliance priority rule which stipulates that a less than equal liberty must be acceptable to those citizens with the lesser liberty (250). If he intends the latter, then the possibility is relevant to the assessment of the special conception in a well-ordered society. Presumably utilitarianism would allow slavery in these circumstances as well, so no contrast can be drawn.

19. Narveson, "Rawls and Utilitarianism," pp. 16-17; Kenneth J. Arrow, "Some Ordinalist-Utilitarian Notes on Rawls' Theory of Justice," *The Journal of Philosophy* LXX, no. 9 (May 10, 1973), 250; Brian Barry, *The Liberal Theory of Justice* (Oxford: The Clarendon Press, 1973), p. 106; Lyons, "Nature and Soundness," pp. 142-145.
20. See Arrow, "Ordinalist Notes," p. 250.
21. Rawls also argues that his conception of justice handles justice between generations more effectively than utilitarianism does (286). I will not attempt here to resolve this difficult issue.

    A number of authors have argued that Rawls' principles actually give rise to *less* intuitive results than utilitarianism does. See, for example, Arrow, "Ordinalist Notes," and Barry, *Liberal Theory of Justice.*
22. Brandt, *A Theory of the Good and the Right*, Chapter 12. The "impersonality" of Rawls' parties was pointed out to me by Gary M. Busch.
23. Arrow, "Ordinalist Notes," p. 257.
24. See Jane Bryant Quinn, "Perquisites: A Status Report," *Newsweek*, July 24, 1978, p. 13c.
25. For a discussion of this issue and the ways in which Rawls might try to avoid it, see Allan Gibbard, "Disparate Goods and Rawls' Difference Principle," forthcoming in *Theory and Decision.*
26. Some work has been done on the measurement of power, but it is unclear what the relation is between *power* and Rawls' notion of *powers.* On the measurement of power, see Alvin I. Goldman, "Toward a Theory of Social Power," *Philosophical Studies* 23 (1972): 221-268, and "On the Measurement of Power," *The Journal of Philosophy* LXXI, no. 8 (May 2, 1974): 231-252. The first of these contains references to literature in the social sciences.
27. See Arrow, "Ordinalist Notes," p. 254.
28. Frederick Schick, "A Calculus of Liberalism," unpublished paper presented at the conference on utilitarianism at Virginia Polytechnic Institute and State University, May 18-21, 1978, p. 2.
29. Arrow, "Ordinalist Notes," p. 254.
30. Rawls' citation of the three "features" follows William Fellner, *Probability and Profit* (Homewood, Ill.: R.D. Irwin, Inc., 1965), pp. 140-142 (154n.).
31. Barry, *Liberal Theory of Justice*, p. 91; R. Duncan Luce and Howard Raiffa, *Games and Decisions* (New York: John Wiley and Sons, Inc., 1957), Chapter 13.
32. See Narveson, "Rawls and Utilitarianism," p. 21, and Thomas Nagel, "Rawls on Justice," in Daniels, *Reading Rawls*, pp. 11-12.
33. John C. Harsanyi, "Can the Maximin Principle Serve as A Basis for Morality? A Critique of John Rawls' Theory," *The American Political Science Review* LXIX (June 1975), 598.
34. Narveson, "Rawls and Utilitarianism," pp. 21-22.
35. Although Nagel has noted, significantly, that the decision is "important" only within the range of social arrangements where disaster is a possibility; with respect to more socially developed circumstances, the "importance" argument looks far less compelling (Nagel, "Rawls on Justice," p. 11).
36. R.M. Hare, "Rawls' Theory of Justice," in Daniels, *Reading Rawls*, p. 103.
37. Narveson, "Rawls and Utilitarianism," pp. 10-11; Barry, *Liberal Theory of Justice*, pp. 97-98; Nagel, "Rawls on Justice," p. 12.
38. Rawls himself does not state this explicitly.
39. Barry, *Liberal Theory of Justice*, p. 105.
40. Barry, *Liberal Theory of Justice*, p. 98.
41. Narveson, "Rawls and Utilitarianism," p. 23.
42. See Hare, "Rawls' Theory of Justice," pp. 104-105, and Nagel, "Rawls on Justice," p. 12.
43. There are interesting problems here, however, about the nature of "psychological possibility." Rawls' remarks suggest that it may be psychologically impossible for a member of society to accept a given principle of justice if he does rather poorly under it, *and* if he knows he personally would have done better under some other principle of justice (174-175). But of course people living under Rawlsian principles might know this.
44. Rawls also mentions as a third circumstance the recognition of those who follow the governing principle of justice as being admirable. However, this appears to play a less significant role in his argument that his conception of justice fares better than utilitarianism.
45. Rawls also identifies a second sense in which his principles benefit everyone, but it relies on empirical assumptions on which he wishes to rest no weight, which many commentators have found suspect, and which, if true, would render utilitarianism and the difference principle equivalent in their prescriptions (80, 82). I shall therefore ignore this second sense.
46. See Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, Inc., 1974), pp. 190-197, for an illuminating discussion of the problem of "sacrifice."
47. J.J.C. Smart, "An Outline of a System of Utilitarian Ethics," in J.J.C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: The University Press, 1973), p. 7.

    Rawls may have in mind here the sort of argument which it would be necessary to use when selecting a principle of justice in the original position, although it is not clear this restriction would be fair. In any event, the argument from the original position for utilitarianism is no more difficult than the argument for his own principles.
48. John C. Harsanyi, "Cardinal Utility in Welfare Economics and the Theory of Risk Taking," *Journal of Political Economy* 61 (1953), and "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy* 63 (1955); and W.S. Vickrey, "Utility, Strategy, and Social Decision Rules," *Quarterly Journal of Economics* 74 (1960). See also Rawls' presentation of this argument in Section 27.
49. Rawls (173-175); David Gauthier, "On the Refutation of Utilitarianism," unpublished paper presented at the conference on utilitarianism at Virginia

Polytechnic Institute and State University, May 18-21, 1978, pp. 18-22.

50. Harsanyi appears to suggest that we can derive interpersonal comparisons of utility in the following manner. Suppose the problem is whether to provide Jones or Smith with an hour's pleasure of a certain sort. We arrange for Jones to understand exactly what being Smith and experiencing that pleasure would be like for Smith, and we also arrange for Smith to understand exactly what being Jones and experiencing that pleasure would be like for Jones. We then ask each to choose between the following alternatives: (a) taking a fifty-fifty chance of being Jones or Smith, and Jones' experiencing the pleasure, or (b) taking a fifty-fifty chance of being Jones or Smith, and Smith's experiencing the pleasure. It is assumed they would make the same choice, and that the choice of, say, alternative (a) shows that Jones experiencing the pleasure has greater value than Smith's. However, as Allan Gibbard points out, this procedure does not work if Jones places a different value on the sort of experience that Smith would have than Smith himself does. And it seems possible that Jones and Smith may indeed have different utility functions for the same subjective experiences.

51. I wish to thank John G. Bennett, Richard Brandt, Alvin Goldman, Jan Narveson, and particularly Allan Gibbard for their assistance with this paper.

Rawls and Marx

Joseph P. DeMarco
CLEVELAND STATE UNIVERSITY

INTRODUCTION

John Rawls' almost immediate success with *A Theory of Justice* suggests that he is able to articulate some beliefs about justice important in contemporary Western social and political thinking. He does this by taking a position within the social contract tradition which has had great importance as a foundation of political thoughts, in a way that seems to support current demands for equal opportunity and for greater economic equality. His work is meant as a challenge to basic patterns of institutional life, but the standards he sets are not out of line with some fundamental predispositions in Western societies. His position is basically individualistic; he is not a radical egalitarian—instead he is theoretically willing to tolerate even wide economic and social inequalities. His call for equal liberty is restricted to political liberties, and his demand for full equal opportunity is somehow seen by him as consistent with predictable disabilities in initial life prospects accruing from basic class differentiations.[1] Of course, these concessions are not to be viewed as a commitment to basic patterns of social and economic inequality; his clear intent is to force such inequalities into a position in which they carry the burden of proof. A defense of inequality must include everyone—it is only in mutual benefit through greater efficiency that inequalities become tolerable. Rawls taps dominant sentiments: productive efficiency and individual self-concern are essential ingredients of his theory.