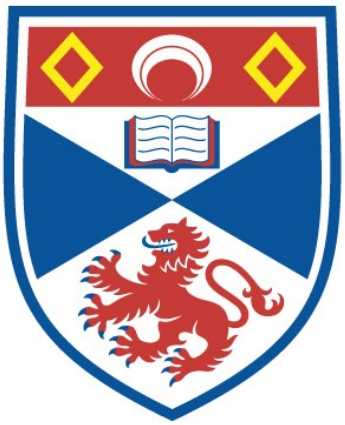
THE THEORY OF RATIONAL DECISION AND THE FOUNDATIONS OF ETHICS

Lanning Sowden

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



1983

THE THEORY OF RATIONAL DECISION

AND

THE FOUNDATIONS OF ETHICS

by

LANNING SOWDEN

B.A. (Hons), M.A.

A thesis submitted to the Faculty of Arts of the University of St. Andrews in fulfilment of the requirements for the degree of Doctor of Philosophy.

October 1982

FOR S.S.

... morals are the work of woman.

Tocqueville, *Democracy in America*, Vol.II, III, 9.

CONTENTS

DECLARATIONS

I, Lanning Patrick Sowden, hereby certify that this thesis which is approximately 67,000 words in length has been written by me, that it is a record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

date 28/10/82 signed .....

I was admitted as a research student under Ordinance No.12 on 1 October 1979 and as a candidate for the degree of Doctor of Philosophy on 1 April 1980; the higher study for which this is a record was carried out in the University of St. Andrews between October 1979 and October 1982.

date 28/10/82 signed .....

I hereby certify that the candidate has fulfilled the conditions of the Resolutions and Regulations appropriate to the degree of Doctor of Philosophy of the University of St. Andrews and that he is qualified to submit this thesis in application for that degree.

date 28/10/82 signed .....

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to my supervisor Professor Bernard Mayo who patiently read and carefully commented on the material that has gone to make up this thesis. Of those not officially involved in the supervision of the thesis I would especially like to mention Professor Crispin Wright and Dr Peter Clark whose help and encouragement has been invaluable. From these three gentlemen I have learned a lot: they have been the best of teachers. The opportunities they gave me to inflict my inchoate ideas on innocent students in various undergraduate courses was clear evidence (unfortunately for a central idea in my thesis) that here were individuals who did not value certainty as such: I can't thank them enough for their willingness to take a chance and let an untried post-graduate have a go. I would also like to thank the University of St. Andrews whose provision of a full St. Andrews University Scholarship over the last three years has enabled me to undertake this work. Finally, I am exceptionally grateful to Mrs Eileen McRobbie for her expert and very efficient typing.

ABSTRACT

The primary concern of this thesis is to investigate what light (if any) the theory of rational decision can throw on certain problems in first-order ethics. In particular, it examines whether given a *correct* theory of decision we can determine which of the two major rivals in the field of contemporary ethics, utilitarianism and contractarianism, is the more adequate moral theory. I begin by outlining what I call *orthodox* decision theory and note from this theory together with a minimal characterization of what it is to make a moral judgement we can deduce utilitarianism. The apparent conflict between utilitarianism and our moral intuitions is then examined. I criticize a common response made by utilitarians to this conflict, namely, their recourse to the distinction between rule and act utilitarianism. But I then ask the question of whether this conflict really matters? I conclude that in a sense it does not. I then turn from a consideration of the *implications* of utilitarianism to its *foundations*, particularly, its foundations in orthodox decision theory. I attempt to establish that orthodox theory has empirical content and that it has been falsified. I also consider the theory from the normative standpoint and construct a *prima facie* case against it. I now consider the dispute between the contractarian and the utilitarian and note that it is essentially decision theoretic in character. From a consideration of what was found to be mistaken about orthodox theory I now argue for a defence of the selection rule for rational choice presupposed by contractarianism and thereby offer a (partial) defence of a contractarian theory of justice.

CHAPTER I

RATIONALITY AND MORALS : INTRODUCTORY REMARKS

In this thesis we investigate an old philosophical problem, *viz.*, the relationship between rationality and morals, but from a somewhat different perspective. We tackle the problem from the point of view of contemporary decision theory: we will attempt to see what light, if any, this theory can throw on first-order ethics, in particular, the two major rivals in the field of contemporary first order ethics, utilitarianism and contractarianism.

There are two roles that will be played by the theory of decision in our argument. Its major role will be to see which of utilitarianism or contractarianism is the more adequate theory; its subsidiary role will be to see whether its modelling techniques and other analytical tools can throw further light on a famous distinction *within* utilitarianism.

It is important that we should have a motivation for employing decision theory in these two roles, especially in its role as arbiter between utilitarianism and contractarianism. For a very widespread view amongst

moral philosophers is that it is a rather simple matter to demonstrate the inadequacy of utilitarianism: we simply show that it can be brought into conflict with some of our most firmly held moral intuitions. This is a view also shared by many philosophers of utilitarian persuasion and hence they involve themselves in all manner of contortions to prevent their theory from coming into conflict with these intuitions. This raises an additional problem within the area subsumed under the foundations of ethics part of our title: we are not simply concerned to look at the foundation of ethical theories in the theory of rational behaviour, but we are also concerned with the foundational problem of whether we can decide between ethical theories on the basis of intuition, in particular, whether we can decide on the adequacy of *utilitarianism* on the basis of an appeal to our moral intuitions. I will argue that we cannot decide on the adequacy of utilitarianism by these means. And this provides us with the (additional) motivation to engage in an examination of the somewhat technical literature that surrounds the theory of decision and to look at the attempts to ground ethical theories in such a theory.

We begin in the next chapter by outlining what I will call *orthodox* decision theory. This is sometimes also known as Bayesian decision theory. This is a theory

of rational decision which deduces from four apparently innocuous assumptions or axioms the theorem that the utility of a "lottery" - a risky alternative - is simply equal to its mathematically expected utility. Orthodox theory then claims that a rational individual will choose (act) so as to maximize expected utility. If we then introduce an account of what it is, in some minimal sense to make a moral judgement, *viz.*, that it be impartial and impersonal, we can with a more formal definition of what this amounts to deduce that a rational individual, if he is to make a moral judgement, must do so as a utilitarian.

In the next chapter, Chapter III, we go on to consider whether utilitarianism has any unfavourable implications, i.e., whether it can be brought into conflict with our moral intuitions. We note that the normal response here is to distinguish between rule utilitarianism and act utilitarianism. The former is advanced because it at least appears to *not* be able to be brought into conflict with our intuition. However, there is a well-known argument to the effect that in all important respects rule and act utilitarianism are equivalent. That is, that *both* theories, when properly understood, prescribe the acts which we find morally objectionable. I call this the equivalence thesis. But the equivalence thesis has not gone unchallanged. Most notably, John

Harsanyi, a well-known decision theorist, has claimed not only that the thesis has already been shown to be mistaken by people such as Gertrude Ezorsky, but that by employing the techniques of decision theory we can show that most definitely the equivalence thesis is false. I agree with Harsanyi that this approach certainly makes more perspicuous in what the distinction between rule and act utilitarianism consists, but I disagree that he has shown that rule utilitarianism is a significant improvement on act utilitarianism. In other words, he has not shown that it is false that both theories prescribe the very acts which we find morally repugnant, i.e., the equivalence thesis still stands. Harsanyi, however, is not the only one to attempt to show that the equivalence thesis is mistaken. There are, for example, the arguments advanced by Ezorsky and John Mackie. But I show that Ezorsky's argument is invalid and that Mackie's argument reduces the status of the rules in rule utilitarianism to the status of the sorts of rules already countenanced by the act utilitarian. But has all this effort at an attempt to show that the theories are or are not equivalent really been worth it? We have *presupposed* that if utilitarianism can be brought into conflict with our moral intuitions then this demonstrates the inadequacy of utilitarianism. By considering the sorts of replies made by the utilitarian to the apparent

fact that his theory can be brought into conflict with our moral intuitions I attempt to show that the question of whether utilitarianism is shown to be inadequate via an appeal to our intuitions is irresolvable : we can provide no conclusive answer one way or the other. This sort of approach, the most common form of approach to the question of the adequacy of utilitarianism, is not a fruitful approach. I suggest, therefore, that we return to look at the attempts to ground utilitarianism in the theory of rational decision making.

Thus in Chapter IV we consider whether or not orthodox decision theory is an adequate theory of rational decision. By looking at an experiment performed by Amos Tversky and Daniel Kahneman I argue that it would appear that orthodox theory has been *falsified*. This experiment shows that individuals do not choose so as to maximize expected utility. However, some orthodox theorists, namely, those who present their theory as a *normative* theory need not apparently be concerned by the results of the Tversky-Kahneman experiment. Afterall, that individuals *do* not choose as their theory says they *ought* to choose is of no consequence. But I argue that we can at least construct a *prima facie* case against the orthodox theorist if he presents his theory as a normative theory. And this for the reason that what the Tversky-Kahneman experiment makes clear is that individuals

whom we would normally regard as quite rational do value certainty as such (and *this* is why they do not choose so as to maximize expected utility) and that no argument has been offered (and no argument seems in the offing) which could show that they *ought not* to value certainty as such. Nonetheless we have not *shown* and I don't think we *could* show that the orthodox theorist is mistaken in claiming that an individual ought not to value certainty. I suggest that if we want a more definite conclusion as regards the adequacy of orthodox theory - which surely we do want as we are now attempting to decide between ethical theories on the basis of an appeal to what constitutes the *correct* theory of decision - then we would do better to adopt an empirical stance towards decision theory. However, it has been argued that decision theory lacks empirical content. I consider two arguments here : one asserts that we must revise our ascription of belief and preference in the face of any apparently falsifying instance to the central hypothesis of decision theory; and the second asserts that if we are to ascribe beliefs and preferences to a person at all then we must presuppose the truth of the central hypothesis of decision theory. I argue that neither argument successfully demonstrates the empirical vacuity of decision theory. But in arguing that decision theory does have empirical content two theses in the philosophy

of action have to be presupposed. In the Interpolation that follows Chapter IV I present an argument for these two theses.

In the final chapter we consider the differing conceptions of justice proposed by utilitarianism and contractarianism and the differing conceptions of rationality that underlie them. According to utilitarianism rational individuals when choosing in a situation which will ensure that the choice is a moral choice will choose so as to maximize expected utility; according to contractarianism they will choose (assuming that choice situation has certain additional features) so as to maximin. The question arises, then, of whether it is ever rational to choose so as to maximin (i.e., to choose that alternative with the maximum minimum utility)? Looking back to what we found to be the error of orthodox decision theory, *viz.*, that it precludes individuals from valuing certainty as such, we see that we can provide a rationale for maximim and its restriction to certain types of situations. We then note that according to John Rawls (a leading proponent of contractarianism) the Original Position (a hypothetical choice situation which ensures that the society chosen by rational individuals is a just society) is indeed a situation of the type where application of the maximin principle is appropriate. By way of clinching our argument we then go on to consider

the objections that have been raised by the orthodox theorist to the maximin principle: these are the objections, presented by means of counter-example, that maximin gives rise to decisions that are clearly irrational and to decisions that are clearly immoral. I give arguments which, I believe successfully, defuse these counter-examples. I then go on to note in what respects the theory I have defended is similar to the theory advanced by the orthodox decision theorist and in what respects it is dissimilar and in accord with Rawl's theory. An appropriate name for the theory is, I suggest, *preference contractarianism*. Finally, I briefly consider the meta-level status of this theory, in particular, the status of the theory given its reliance on certain contingent facts. This is a topic more properly dealt with in a thesis on meta-ethics, nonetheless, the problem is sufficiently pressing for it to be worthwhile to pass some comment. I note that I side with Rawls in believing that it is quite appropriate that our fundamental ethical principles should depend on certain contingent facts about men and society. To a very significant extent this belief is justified by challenging those who think otherwise to come up with a theory that satisfies certain conditions of coherency and understandability. Thus, it is no weakness in our argument for preference contractarianism that both or either of the following is the

case: (a) that viewing decision theory from the normative standpoint, when we establish the *prima facie* case against orthodox theory we do not establish that rational individuals *ought* to value certainty, but only that if they do (a contingent fact) then they ought to choose in an unorthodox manner; and (b) that viewing decision theory from the empirical viewpoint we only establish that the correct theory of *actual* human decision is unorthodox theory.

As a quick perusal of this thesis will make apparent I have had to touch on many problems which have assumed a considerable importance in this and other areas of philosophical enquiry, for example: the problem of when a theory has empirical content, the relationship between reasons and actions, the nature of rules of conduct, holism, the debate between rule and act utilitarianism, and so on, each of which could have had a thesis written on it in its own right. With respect to some of these problems I have adopted a position and offered considerable supporting argument. But with respect to some other of these problems I have adopted a position with the intention that the position I adopt is at least plausible and not overly controversial. To the extent that as regards these problems other positions certainly are possible and that I have offered little reason to suppose otherwise then there will be those who may fail to be totally convinced of my main argument: this, unfortunately, is unavoidable.

CHAPTER II

THE DEDUCTION OF UTILITARIANISM FROM ORTHODOX DECISION THEORY

1. The Types and Attractions of Utilitarianism.
2. An Outline of Orthodox Decision Theory.
3. The Equiprobability Model and the Deduction of Utilitarianism.
4. Interpersonal Comparisons of Utility.

1. *The Types and Attractions of Utilitarianism*

It is possible to distinguish a number of different forms of utilitarianism. A traditional form of utilitarianism - *hedonistic* utilitarianism - claims that what one morally ought to do is bring about the greatest amount of *happiness*. According to *ideal* utilitarianism what one ought to do is bring about the greatest *intrinsic good*. And in its most modern guise - *preference* utilitarianism - what one ought to do is maximize *individual utilities*. Now each of these statements of utilitarianism is ambiguous in at least two respects. The first respect is this: it is not clear from the

above statements whether what we ought to do is bring about the greatest *sum* of happiness/intrinsic good/ individual utilities or whether we ought to maximize the *mean*.[1] Depending on which approach we take we can distinguish two types of utilitatianism for each of the major forms. For example, consider hedonistic utilitarianism: we could have, considering each individual level of happiness, that we ought to *either* maximize the sum total of such levels *or* maximize the arithmetic mean of such levels. The proper resolution of this ambiguity will not be of any great concern in this thesis[2] : the problems that we will be concerned with in our study of utilitarianism are not problems that arise because of this ambiguity nor are they effected by it. Indeed, the form of utilitarianism that will be the focus of attention in this thesis - preference utilitarianism - is generally put forward as a theory of the "average" type. Thus, according to John Harsanyi[3], a utilitarian is required to maximize *social utility* and this is defined as the arithmetic mean of all individual utilities. As we shall see, that this form of utilitarianism should be of the average type follows from what (supposedly) it is to make a rational choice conjoined with a minimal characterization of what it is to make a moral judgement.

I mentioned that the three statements of the

general forms of utilitarianism were ambiguous in *two* respects. The second respect arises because however it is that we specify the *maximand*, i.e., what it is that the utilitarian is supposed to maximize, we have not thereby fully characterized the utilitarian's theory for we have not specified the *constraints* (if any) of maximization. Here it is a commonplace to distinguish two types of utilitarianism known as *act* utilitarianism and *rule* utilitarianism. However, we put this distinction to one side for the moment: the motivation for and the specification and examination of the distinction will be the subject of Chapter III. It suffices to note here that whether we suppose that according to utilitarianism we ought to maximize the total or average of happiness or intrinsic good or individual utilities, we still have not ascertained what constraints the utilitarian is to place on that maximization, and what constraints *are* imposed may have important repercussions for any attempt to determine the adequacy or otherwise of utilitarianism.

Utilitarianism certainly has some attractions as a moral theory, particularly, I think, in its hedonistic and preference forms. It appears to offer a coherent theory of morality in the sense that all our moral views can be seen to flow from a single principle. In contrast, the deontological theories - theories that assert that

morality essentially concerns *justice* and *rights*- which are put up as rivals to utilitarianism often seem to be no more than an arbitary collection of principles that even on occasions conflict, and indeed often conflict with the precepts of commonsense morality unless some apparently *ad hoc* adjustments are made to the principles. Moreover, these latter theories generally do not seem to offer a meaningful account of how we come by these principles, whereas utilitarianism does offer an account of how we come by our moral principles. Utilitarianism seems to offer the prospect of more or less definitive conclusions as to what we ought or ought not to do, and the prospect of a means of deciding between the often conflicting precepts of commonsense morality and, indeed, of the deontological principles themselves. And finally, as regards hedonistic and preference utilitarianism, there certainly seems *something* right in the idea that morality should have something to do with happiness promotion or preference satisfaction: at least such an idea accords well with the humanitarian spirit of our age. To be sure, the notions of "happiness" and "preference" have to be understood in a somewhat technical sense, and we will return to this problem in Chapter V when we touch on the question of what preferences are to be properly included in a moral calculus.

The above considerations give us some reason for

taking utilitarianism seriously. But the considerations are not conclusive. First, it may turn out that utilitarianism has implications which are in some way totally unsatisfactory. I consider this possibility in Chapter III. In which case we might conclude that there is just no possibility of formulating a successful moral theory with utilitarianism's apparent advantages of coherency, etc.. Second, it may turn out that there is another moral theory, completely at variance with utilitarianism, but which nonetheless shares all or nearly all of the advantages of coherency, etc.. I argue for such a theory, which I call *preference contractarianism*, in Chapter V.

In this chapter I want to consider what more can be said in favour of utilitarianism by looking at its foundations. In particular, I want to consider the idea that utilitarianism is the only *rational* moral code. That is, that if an individual is rational, and if he is to make a moral judgement, then he must do so as a utilitarian. I will look at this idea as it has been expounded by John Harsanyi[4].

2. *An Outline of Orthodox Decision Theory.*

Orthodox (i.e., Bayesian) decision theory is advanced as a theory of *rational* behaviour or choice. Just in what sense

it is such a theory and whether it is an adequate theory are questions that I postpone until Chapter IV. The purpose of the present section is to simply present the theory in outline to facilitate an understanding of Harsanyi's deduction of utilitarianism from orthodox theory.

In decision theory decision situations are divided into three types: decisions under certainty, under risk, and under uncertainty. A decision maker is said to make a decision under *certainty* when he can predict the actual outcome of the action he chooses to perform. He makes a decision under *risk* when he knows the *objective* probability associated with each possible outcome. And finally, an individual makes a decision under *uncertainty* when he does not know some or even all of these probabilities.

It is assumed that in a situation of certainty an individual will choose that action with an outcome that has the *highest* utility. To say that one outcome has a higher utility to another is just to say that the individual *prefers* that outcome to the other. But for us to be able to claim that an individual will choose so as to maximize utility, we must make certain assumptions about the preference relation, in particular, the relation must be such that it orders the outcomes so that there will be a most preferred outcome (i.e., an outcome with the highest utility). This is ensured by supposing that the preference relation is complete and transitive.

That is, for any three outcomes O_r, O_s, and O_t, where $O_r \geqslant_i O_s$ means that individual i prefers O_r to O_s or is indifferent between them, we assume:

(1) Completeness

$$(\forall i)(\forall O_r)(\forall O_s)\ (O_r \geqslant_i O_s) \vee (O_s \geqslant_i O_r)$$

(2) Transitivity

$$(\forall i)(\forall O_r)(\forall O_s)(\forall O_t)\ (O_r \geqslant_i O_s) \wedge (O_s \geqslant_i O_t) \rightarrow (O_r \geqslant_i O_t)$$

This is sufficient to induce a weak ordering of the outcomes, i.e., an *ordinal* utility scale. These two conditions have some plausibility and are clearly *required* if we are to say that an individual chooses that action with an outcome that has the highest utility. Consider completeness: suppose this condition does not hold then we allow that there may be two outcomes between which the individual is not indifferent nor does he prefer one to the other. Obviously then an individual cannot choose between these two outcomes by choosing that with the highest utility (i.e., that which he most prefers). Consider transitivity: suppose this condition does not hold then we allow that there may be some set of outcomes such that there is no most preferred outcome because for any outcome there is another outcome preferred to it (i.e., without transitivity our preferences may be "cyclic"). Preferences that obey these two conditions

are said to be *consistent*.

But an *ordinal* utility scale while it may suffice for decisions under certainty will clearly be insufficient for decisions under risk or uncertainty. Suppose, for example, that I may choose between one action with an outcome that I value highly but which has a *low* probability and another action with an outcome that I value lowly but which has a *high* probability (and between all the other possible outcomes of those two actions I am indifferent), then we cannot say which action is to be chosen *unless* we have a quantitive measure of preference, i.e., unless we can say *how much* one outcome is preferred to another. In other words, at least some decisions under risk and uncertainty seem to require *cardinal* utility scales. To induce a cardinal utility scale we must impose further conditions on the preference relation. These conditions depend on the notion of a *"lottery"*. A lottery L is simply a probability distribution over the members of a set of possible outcomes. In making choices between actions under risk or uncertainty an individual can be regarded as choosing between lotteries where the "prizes" are the possible outcomes. For consider: in choosing between alternative actions under risk or uncertainty an individual chooses an action which has a set of possible outcomes associated with it, such that if he chooses a certain action then

certain outcomes will obtain given the occurence of certain events whose occurrence will be at certain probabilities. A lottery L can be described thus:

$$L = (O_1/e_1; \ldots; O_r/e_r; \ldots; O_R/e_R)$$

indicating that the lottery L (the action) will yield outcome O_r as "prize" if event e_r occurs ($r = 1,\ldots,R$). The events $\{e_1,\ldots,e_R\}$ are regarded as mutually exclusive and exhaustive. L is a *risky* or *uncertain* lottery according to whether the decision maker does or does not know the probabilities associated with all the events e_r. Note that a prize may itself be a lottery.

On Harsanyi's approach[5], the following two further conditions are imposed on the preference relation. They are:

(3) *Probabilistic Equivalence*

Let

$$L = (O_1/e_1; \ldots; O_R/e_R) \text{ and } L^* = (O_1/e_1^*; \ldots; O_R/e_R^*)$$

and suppose that the decision maker knows the objective probabilities associated with each of the events $\{e_1,\ldots,e_R\}$ and $\{e_1^*,\ldots,e_R^*\}$, and that he knows

$$\text{Prob}\ (e_r) = \text{Prob}(e_r^*) \text{ for } r = 1,\ldots,R$$

Then

$$L =_i L^* \quad \text{(where } =_i \text{ means "indifferent for } i\text{")}$$

In other words, a decision maker is assumed to be *indifferent* between two risky lotteries provided the lotteries have the same prizes (outcomes) with the same probabilities - even though the events which bring about those outcomes are quite different. In particular, as Harsanyi notes, he will be indifferent between a one stage and a two stage lottery if they are probabilistically equivalent. For example: suppose an individual is offered a lottery which is such that if a certain event occurs (say, that a pointer lands on a certain segment of fairly spun disc divided into ten equal segments) which has a probability of $\frac{1}{10}$ of occurring then he will receive a $1000 prize (and nothing should the event not occur). And suppose that that individual is offered another lottery which is such that if a certain event occurs (say, that a pointer lands on a certain segment of a fairly spun disc divided into five equal parts) which has a probability of $\frac{1}{5}$ of occurring then he will receive a *lottery* as a prize (and nothing otherwise) which is such that if an event occurs (say, that a fair coin lands up "heads") which has a probability of ½ of occurring then he will receive $1000 (and nothing otherwise). According to (3) above the individual should be indifferent between the former simple lottery and the

latter compound lottery; after all they offer the same prizes at the same probabilities - the events upon which the outcomes are conditional are different, but they are *probabilistically equivalent*.

(4) *Sure-thing Principle.*

Suppose

$$O_r^* \succcurlyeq_i O_r \quad \text{for } r = 1,\ldots,R$$

then

$$(O_1^*/e_1;\ldots;\ O_R^*/e_R) \succcurlyeq_i (O_1/e_1;\ldots;\ O_R/e_R)$$

That is, in Harsanyi's words "other things being equal, an individual will not prefer a lottery yielding less desirable prizes to a lottery yielding more desirable prizes". In particular, and for example, this condition requires the following (recall that a prize may itself be a lottery). Suppose an individual prefers the prize of receiving (for certain) \$400 to a prize of receiving \$1000 at probability ½ (and \$0 at probability ½). And further suppose that he is offered the following two lotteries. The first consists of his receiving \$400 should a certain event occur (and nothing otherwise). The second consists of his receiving a ½ chance of \$1000 and a ½ of \$0 should the *same* event occur (and nothing otherwise). Then according to (4) he will not prefer the second lottery to the first. Clearly, we must presuppose here that two events are the same (type) *only if*

they have the same probability of occurring, otherwise (4) might not hold[6]. For consider: it would be absurd to suppose that the first lottery is still preferred to the second when the probability of the event in the first was approaching zero, and the probability of the "same" event in the second was approaching one. Thus the Sure-thing Principle requires that if an individual is offered a lottery consisting of a prize of $400 should a certain event occur with probability $\frac{1}{5}$ (and nothing otherwise), and another lottery consisting of a prize of a ½ chance of $1000 and ½ chance of $0 should the *same* (type of) event occur - in particular, an event with probability $\frac{1}{5}$ of occurring - (and nothing otherwise), then, given the original preference, he will not prefer the second lottery to the first.

The general thrust of these two conditions is clear. The *Probabilistic Equivalence Postulate* roughly says that if two lotteries are equivalent in terms of prizes and the probability of prizes (computed according to the ordinary probability calculus), then an individual will be indifferent between those two lotteries. The *Sure-thing Principle* roughly says that if an individual prefers or is indifferent between two prizes, then he will prefer or be indifferent between any two lotteries, one involving one of the prizes and the other involving the other prize, provided the lotteries are otherwise

equivalent, in particular, that the probability of receiving each prize in each lottery is equivalent. As such we can see that these two conditions are essentially von Neumann and Morgenstern's "Reduction of Compound Lotteries Assumption" and the "Substitution Assumption" respectively.[7]

If an individual's preferences satisfy conditions (1) - (4) we can deduce the *expected utility theorem* which says, that the utility U_i of any lottery L is equal to its expected utility, i.e.:

$$U_i\ (L) = \sum_{r=1}^{R} p_r\ U_i\ (O_r)$$

Here p_r is the probability associated with the event e_r, r = 1,...,R. This probability is deemed to be *objective* if L is a risky lottery, and *subjective* if L is an uncertain lottery. An individual who chooses so as to *maximize expected utility* chooses, according to orthodox theory, *rationally*. Given this theorem we are able to construct a cardinal utility function for any individual i over any set of possible outcomes.

Most importantly we should note here that according to orthodox theory the utility of a lottery is equal to its *expected utility*, i.e., it is equal to the sum of the products of the utility of each of the (mutually exclusive and exhaustive) outcomes times its probability,

and that an individual who chooses - acts - so as to maximize expected utility, chooses rationally. Let us now go on to consider what is required of such a rational individual if he is to choose *morally*. Harsanyi takes two approaches here[8]:- an axiomatic approach and an approach that involves what he calls "the Equiprobability Model for Moral Judgements". For heuristic purposes we will follow the latter approach.

3. *The Equiprobability Model and the Deduction of Utilitarianism*..

Harsanyi takes it that the hallmark of a moral judgement is that it is based on *impartial* and *impersonal* criteria. About this I think Harsanyi is right: *at the very least* what we require of a moral judgement is that it be impartial and impersonal, i.e., that in making a moral judgement we should not attempt to further our own interests, even if these interests are understood to be not simply selfish, at the expense of others.[9] In any case, that is the view that will be adopted in this thesis and as we shall see it is a view that is common to the moral theory being advanced here and the one that will be advanced in opposition to it in Chapter V. But even if we grant that it is a minimal requirement on moral judgements that they be

impartial and impersonal we still have not got very far unless we have a more precise understanding of what that requirement amounts to. To this end Harsanyi advances the *Equiprobability Model*.

Suppose that an individual *j* is to make a moral choice between bringing about one of two social situations or arrangements, e.g., a choice between two alternative arrangements of income distribution. Of course, not all our moral choices are choices between such social situations, but supposing that we do have such a choice here will enable us to more easily, in Chapter V, draw the parallel between Harsanyi's Equiprobability Model and Rawl's Original Position. Moreover, as we shall see, the Equiprobability Model is readily extendable to other moral choices, e.g., the choice between whether I take Mr X's property or not take it. Call the situations A and B. One way of ensuring that *j*'s choice between A and B is not unduly influenced by his own self-interest is to require that he choose between A and B *without knowing* what his own position will be under either arrangement, i.e., without his knowing which individual he will be under either arrangement. Now as *j* does not know what position he will occupy or which individual he will be, then he seems required by the Laplacean principle of insufficient reason to assign an *equiprobability* to his ending up in any one of the *n* positions in society or being any one

of the n individuals in society. That is, he must assign a $\frac{1}{n}$ probability to his ending up in any one of the n positions or of his ending up as any one of the n individuals. But if a *rational* individual is to make such a choice under these constraints, i.e., if the individual is to maximize expected utility under these constraints, then he will choose that particular situation which *maximizes the average utility level in society*. That is, an individual who chooses under the Equiprobability Model so as to maximize expected utility will choose so as to maximize the social welfare function:

$$W_j\ (A) = \frac{1}{n} \sum_{i=1}^{n} U_i\ (A)$$

for some alternative A. That is, he will choose as an *"average" utilitarian.*

There are a number of points that are especially worthy of note here. First, as may be clear from the above equation, and as Harsanyi notes[10], when we say that in making a moral judgement an individual must do so on the assumption that he has an equiprobability of occupying any of the n positions in society we mean he must assume that

> he had an equal chance of being "put in the place of" any individual member of the society, with regard not only to his objective

> social (and economic) conditions, but also to his subjective attitudes and tastes. In other words, he ought to judge the utility of another individual's position not in terms of his own attitudes and tastes but rather in terms of the attitudes and tastes of the individual actually holding this position.

This explains my vacillation in the previous paragraph between speaking of an individual assuming that he has an equiprobability of occupying any of the *n* social positions or of being any one of the *n* individuals in society. For in saying that the individual must assume when making a moral judgement that he has an equiprobability of occupying any one of the *n* social positions *with the preferences of the individual in that position*, then to all intents and purposes he must assume that he has an equiprobability of being that individual. At least that is a handy short-hand expression for what we have in mind. The idea is a fairly common one in moral philosophy, particularly among those of utilitarian persuasion.[11] The second point I wish to make is that having said this it is easier to see how the Equiprobability Model is extendable to moral judgements more generally, e.g., in choosing between whether I take Mr X's property or not take it. If I am to make a *moral* judgement here then I must assume that I have the same

chance (50/50) of being in *my* position with *my* preferences, and being in *Mr X's* position with *Mr X's* preferences. Finally, it is not necessary that we take the Equiprobability Model *literally*. For an individual to make a moral judgement it is not necessary that he literally not know which individual he will be: it will be enough that he should choose *as if* he did not know.

Of course in this attempt to arrive at a utilitarian theory of morality we have had to presuppose *interpersonal* comparisons of utility. And although this presupposition will not figure in our criticism of utilitarianism and, indeed, such comparisons will also be presupposed in the alternative moral theory advanced in Chapter V, nonetheless, interpersonal comparisons of utility have been regarded with some suspicion for some time. Hence it would be worthwhile to say *something* about interpersonal comparisons of utility no matter how briefly, even if in so doing we only succeed in indicating in what direction the argument would go for a full defence of such interpersonal comparisons.

4. *Interpersonal Comparisons of Utility*

Once again my argument draws heavily on the work of Harsanyi[12]. He distinguished two problems that concern interpersonally comparable utility functions: the

"psychological problem" and the *"metaphysical problem"*, or, as we might call them, the *practical* and the *theoretical* problem. It is the latter in which, as philosophers, we are primarily interested and about which we are competent to judge.

The above problems arise when we come to consider *how* we are to estimate another person's present utility level. The way we attempt to do this in everyday life (e.g., when I attempt to estimate whether my wife would prefer to go to the movies or stay at home, and whether her preference to go to the movies sufficiently outweighs my parents desire for the letter that I could write if we stayed home) is based on what Harsanyi calls *"imaginative empathy"*. This is the ability to imagine ourselves to be in the shoes of other people. Now imaginative empathy does *not* amount to my attempting to estimate another person's utility level by supposing that I am in *his* position with *my* preferences: rather, it requires that I suppose that I am in *his* position with *his* preferences. (Note that this is just what we said was required of a moral agent under the equiprobability model). But how do I manage to do this? Well, what we normally do is to ask, "How much utility would *I* derive from that situation supposing that *my* personality had been formed by those biological, psychological and sociological factors that have formed *his* personality?

(Hence, in the rough and ready reckonings of everyday life I might say, "If *I* were a parent, with an elder son, etc. I would greatly appreciate a letter...".) To be sure, it is very difficult, in most cases, to work out what our psychological reactions would be to these determining factors and, indeed, to isolate those determining factors themselves. But the difficulties here seem to be of an empirical or factual nature and they seem settleable, at least in principle, by the normal methods of empirical science. These difficulties constitute the *practical* or *psychological* problem in interpersonal utility comparisons.

The *theoretical* or *metaphysical* problem arises because we have *assumed* in the previous paragraph that *different* people will have *similar* psychological reactions to the same situation given the same psychological background. This assumption Harsanyi thinks is justified by what he calls, the *"similarity postulate"*. Of this postulate Harsanyi says:

> By this I mean the principle that, given the basic similarity in human nature (i.e., in the fundamental psychological laws governing human behaviour and human attitudes), it is reasonable to assume that different people will show very *similar* psychological reactions to any given objective situation,

> and derive much the *same* utility or disutility from it – *once proper allowances have been made for any empirically observed differences* in their biological make-ups, in their social positions, in their educational and cultural backgrounds and, more generally, in their past life histories. In other words, in the absence of clear evidence to the contrary, the presumption must always be that people's behaviour and psychological reactions will be similar in similar situations.[13]

Notice that just such a postulate or principle is at work in other areas. It is used when we attempt to gauge the degree of another person's pain or, indeed, whether they are in pain at all. For example, I see you struck on the head with a mallet, and I say, "He is in great pain". And, if I'm asked, "How do you know that he is in great pain or even in pain?", the typical response is, "Well, wouldn't *you* be in considerable pain if you had just been struck on the head with a mallet?" But in so saying we presuppose that in the absence of evidence to the contrary people's psychological reactions will be similar in similar situations. Of course, my claim that you are in great pain may be wrong; maybe that wasn't a mallet but only a stage prop, maybe you have no nerves

in your head, and so on. But the force of the similarity postulate is that *in the absence* of such evidence we are entitled to presume that you are in great pain. Incidently, it is of no consequence that in the case of other person pain ascriptions I have more to go on than the mere observation that they were, for example, struck on the head with a mallet - there are their verbal utterances (e.g., you say, "I am in great pain"). We also have just such "evidence" in the case of other person utility level ascriptions. The problem is how can I know that you would utter these words just when I would utter those words (i.e., if I were in your situation) - is what you would count as being in great pain what I would count as being in great pain? And this remains a problem even when I determine that you are not play-acting, etc.. The purpose of the similarity postulate is to bridge the gap between the *evidence*, e.g., that you were struck on the head with a mallet, that you have a similar physiology to other people (in particular, to me), that you uttered the words "I am in pain", and the *claim* that your psychological reaction is the same as other people (in particular, that it is the same as mine). In other words, the similarity postulate is advanced as a means of dissolving the age-old philosophical problem of other minds. And hence the theoretical or metaphysical problem that arises with inter-

personally comparable utility functions is just a special case of the general problem of other minds.

Thus Harsanyi remarks that those who reject the idea of interpersonal comparison of utility on the aforementioned theoretical or metaphysical grounds must, if they are consistent, also be skeptics about other minds. But, says Harsanyi,"as common sense tells us, *all* normal humans are fully self-conscious human beings."[14] Hence we cannot reject the theoretical possibility of interpersonally comparable utility functions. However, this is not, I believe, a very felicitous mode of argument on Harsanyi's part. No matter how strong our common sense belief that other people are conscious beings, this in itself does not constitute a refutation of the skeptic. The skeptic accepts that there is such a commonly held belief, but rejects the idea that we have any justification for it. In other words, he rejects the idea that the similarity postulate is *justified*. As a matter of strategy in philosophical argument the better approach for Harsanyi to take would be as follows. It is too much to expect that a philosopher in arguing for a *moral* theory should *provide* solutions to *all* the philosophical problems whose solution is presupposed in the advancement of his moral theory. In particular, he need not be expected to provide a solution to a very general problem like that of other minds which is not simply endemic to

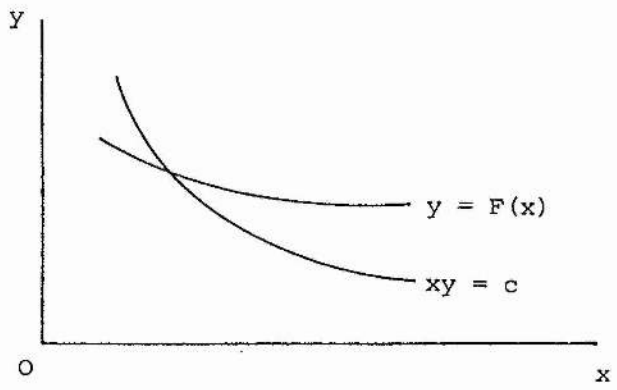
moral philosophy. These considerations are all the more pressing when we notice the apparently intractable nature of the problem of other minds. Rather the moral philosopher should proceed on the assumption that there is a solution, and after all,in the case of other minds, most of us hope there *is* a solution, and get on with the job of formulating his moral theory. If it turns out that the *only* objection to his moral theory centres on the problem of other minds then that is the time to return to a more thorough examination of that problem. In the interim there is important and fruitful work to be done. My point in general is that I take it that we do not want our rejection of utilitarianism to depend only on an appeal to a general and apparently intractable problem like the problem of other minds.

Footnotes:

1. Indeed, our statements of the various forms of utilitarianism are so vague as to leave open other possibilities, e.g., that considering each distribution of levels of happiness/intrinsic good/individual utilities we ought (implausibly) to maximize the *mode*. The two possibilities I mention are the two most commonly cited possibilities.
2. And maybe it generally ought to be of no concern. J J C Smart when discussing what he calls "total" and "average" utilitarianism claims: "In most cases the differences between the two types of utilitarianism will not lead to disagreement in practice. For in most cases the most effective way to increase the total happiness is to increase the average happiness, and vice versa." (Smart (1973) p. 28). However, what is important here is that we stress the phrase *"in most cases"*: there *can* be cases, admittedly we have to suppose that quite a number of assumptions obtain, where there will be an important difference between the two versions of utilitarianism. The total view requires that provided the average utility per person falls slowly enough when the number of individuals increases in some population, then the population ought to be increased in size no matter how low the average utility falls. In such a case

the total utility will be increased by a sufficient amount to make up for the decline in average utility per person. Hence on the total view a very low average of utility may be required. As John Rawls more formally puts it:

Where y is the average utility per person, x is the population size, then if the curve y = F(x) is flatter than the rectangular hyperbola xy=c then total utilitarianism requires that x be increased indefinitely.



For xy equals the total utility, and the area of the rectangle representing the increase in total utility increases as x increases just when y = F(x) is flatter than xy = c. (Rawls (1972) pp. 162-163)

Now provided we are in a world where it is *not* the case that the average utility per person falls slowly enough when there is an increase in popul-

ation size, then here is an instance where there will be a practical disagreement between the total and average utilitarianism. And such people as Rawls take it that we *are* on such a world and that it is preferable to bring about a greater average utility rather than a greater total utility. That is, the more plausible utilitarian theory is *average* utilitarianism.

3. See, for example, John Harsanyi (1978) especially p. 228.

4. Harsanyi has written on this topic on many occasions., for example, see his (1955) and)(1978).

5. See Harsanyi (1978) pp. 224 - 225 and (1979) pp. 290-291.

6. Of course we might not require this as a necessary condition of sameness of events for events *generally*, but it does seem right for these events upon which outcomes are conditional and where we're trying to make a rational choice on the basis of those outcomes. Whether two events which are of the same type must have the same probability of occurring will very much depend on what we take to be our criterion of same type.

7. See R D Luce and H Raiffa (1957) pp. 26-27.

8. See Harsanyi (1978) pp. 226-228 and (1979) pp. 292 - 295.

9. As I also note in Chapter III this has been disputed, see, for example, N Rescher (1975) especially pp. 70-72, and for a reply see R M Hare (1981) pp. 135-140. I do not find this particular line of argument very convincing and this for the reason, examined in Chapter III, that attempts to determine the adequacy or inadequacy of utilitarianism via an appeal to our moral intuitions are irresolvable.

10. Harsanyi (1955) p. 316 footnote 16.

11. For example, see Hare (1963) pp. 112-113 and (1981) pp. 110-111.

12. For example, see Harsanyi (1955) pp. 316-321 and (1979) pp. 300-302.

13. Harsanyi (1979) p. 301.

14. Harsanyi (1979) p. 302.

CHAPTER III

RULE AND ACT UTILITARIANISM, DECISION THEORY AND INTUITION

1. The Conflict between our Moral Intuitions and Utilitarianism.
2. The Attempt to resolve this Conflict by introducing the Distinction between Rule and Act Utilitarianism.
3. The Equivalence Thesis for Rule and Act Utilitarianism.
4. A Decision Theoretic Approach which attempts to falsify the Equivalence Thesis.
5. Other attempts to falsify the Equivalence Thesis.
6. Do conflicts between our Moral Intuitions and Utilitarianism really matter?

1. *The Conflict between our Moral Intuitions and Utilitarianism*

We have already mentioned in the previous chapter the apparent attractions of utilitarianism, namely, that it seems to offer the prospect of a coherent theory of morality, it seems to offer a meaningful account of how we come by our moral principles, and it seems to offer the prospect of definitive conclusions as to what we ought and ought not to do. As we said these considerations give us some reason for taking

utilitarianism seriously. However, as we also noted, these considerations are not conclusive. The first possibility we mentioned was that utilitarianism might have implications that are totally unacceptable. The idea here is that utilitarianism is inadequate as a *moral* theory for it gives rise, if it is consistently followed, to the prescription of actions that are clearly *immoral*. Now if utilitarianism were the *only* theory that offered the aforementioned advantages then we would have to conclude (reluctantly) that there was no possibility of an adequate theory of morality that had these apparent advantages. But the idea that concerns us in this chapter is simply the thought that utilitarianism is inadequate because it has unsatisfactory implications. And it is this form of attack which, I think it will be agreed, has constituted the main and most common form of attack on utilitarianism.

It has for a long time been commonly observed that utilitarianism appears to be in conflict with some of our most firmly embedded moral intuitions. W D Ross, for example, criticised ideal utilitarianism on the grounds that it is mistaken to suppose that it is always right to break a promise whenever one can thereby bring about the greatest intrinsic good.[1] More recently H J McCloskey has pointed out that utilitarianism conflicts with our fundamental beliefs about justice. McCloskey's argument proceeds by way of a well-know example. We are asked to

consider a sheriff in some small town who can prevent the occurrence of riots which will kill and harm, let us say, ten people only if he frames an innocent man and sentences him to death. It seems that according to utilitarianism what the sheriff ought to do is sentence the innocent man, and yet this is counter-intuitive for surely such an act would be *unjust* and therefore ought *not* to be done. It is tempting to think that the utilitarian is *not* committed to saying that what the sheriff ought to do is sentence the innocent man, for if the sheriff were found out this would weaken people's respect for and confidence in the law. Such a consequence may well be worse than the death and harm resulting from the riots. McCloskey grants that this *may* be the case, but as the example is set up it is *not* the case: we can suppose that, *ex hypothesi*, only the sheriff and the innocent party know or could know that the innocent man is in fact innocent. Similarly, we can suppose that *ex hypothesi* the sheriff knows (or, at least, can estimate to a sufficiently large degree of probability) that if he does not sentence the innocent man ten people will die in the riots. As McCloskey emphasises, it is at least *logically possible* that there should be such a situation even if we would never expect such a situation to arise in practice. And that, McCloskey thinks, is sufficient to undermine utilitarianism: "to expose the inadequacy of utilitarinism in dealing with the problem of justice, *only* the logical possibility of such an

'unjust' utilitarian system of punishment needs to be indicated."[2]

Now, by and large, examples such as these where utilitarianism is thrown into conflict with some of our most firmly held moral beliefs have been regarded as demonstrating that the utilitarian has to make some move to restore a semblance of cogency to his theory. And it was in response to these putative counter-examples that the distinction was made between act and rule utilitarianism.

2. *The Attempt to resolve this Conflict by introducing the Distinction between Rule and Act Utilitarianism.*

As I mentioned in Chapter II there remains the possibility that even when we have specified the *maximand* of utilitarian theory, i.e., what it is that the utilitarian is supposed to maximise, we have not thereby fully characterised the nature of his theory, for we have not specified the *constraints* (or lack thereof) of maximization. In this respect it is a commonplace to distinguish two types of utilitarianism known as *act* utilitarianism (AU) and *rule* utilitarianism (RU).

J J C Smart, a leading proponent of AU, offers the following statement of AU (for ease of exposition Smart puts it forward in a broadly hedonistic form):

> Let us say, then, that the only reason for performing an action A rather than an alternative action B is that doing A will make mankind (or, perhaps, all sentient beings) happier that will doing B.[4]

Presumably, if two or more actions result in the *same* amount of happiness each of them is *a* right action. So, whether one defines the maximand of utilitarianism in terms of happiness, intrinsic good, or individual preferences, and whether one supposes that it is the sum or mean of these that is to be maximized, one can be classified as an AUian if one subscribes to the following idea: the utilitarian criterion is to be applied directly to the individual course of action available.

A classic statement of RU is given by J Austin:

> according to that theory, our conduct would conform to *rules* inferred from the tendencies of actions, but would not be determined by a direct resort to the principle of general utility. Utility would be the test of our conduct, ultimately, but not immediately: the immediate test of the rules to which our conduct would conform, but not the immediate test of specific or individual actions. Our

> rules would be fashioned on utility; our conduct, on our rules.[5]

By the *tendency* of an act Austin emphasizes that he does not mean the consequences of that specific act. For the RUian

> The probable *specific* consequences of doing that single act, of forebearing from that single act, or of omitting that single act, are not the objects of the inquiry. The question to be solved is this:- If acts of the *class* were *generally* done, or *generally* forborne or omitted, what would be the probable effect on the general happiness or good?[6]

So, the central idea of RU seems to be this. In order to determine whether some particular act ought or ought not to be done we must consider whether it is in accordance with some rule which enjoins or forbids acts of that type. Such a rule is justified on the grounds that if everyone acted in accordance with that rule then the consequences would be better than if everyone acted in accordance with some contrary rule.[7] Suppose that the rule *forbids* acts of a certain type, then that rule is justified if it is the case that should everyone act in

accordance with that rule the consequences would be better than if everyone acted in accordance with a rule that did *not* forbid acts of that type. This central idea of RU has some initial plausibility. For consider: the performance of some individual act, say, the act of this sheriff sentencing this innocent man, may have consequences that are quite beneficial; but should acts of that type or class be generally done -- if it became the normal practice in cases of punishment to sentence the innocent -- then the consequences would be disastrous. All confidence in and respect for the law would evaporate. So it would appear that the RUian can say in response to McCloskey's example that the sheriff *ought not* to sentence the innocent man for such an action is forbidden by a rule which is justified on utilitarian grounds: the consequences of everyone acting in accordance with a rule that forbade the sentencing of the innocent would be better than the consequences of everyone acting in accordance with a rule that did not forbid the sentencing of the innocent. Thus RU seems impervious to McCloskey's counter-example: RU seems to be able to accommodate an intuition we have with respect to justice.

Before proceeding further with our major argument there are two important points that need to be made about my comments above. I have said, suppose that a rule *forbids* acts of a certain type, then that rule is

justified if it is the case that should everyone act in accordance with that rule the consequences would be better than if everyone acted in accordance with a rule that did *not* forbid acts of that type. There are two objections that can be made against that statement. First, to have a rule which did not forbid acts of that type, i.e., to have a rule which did not forbid acts of sentencing the innocent, is not to have a rule which enjoined or *required* acts of sentencing the innocent. Such a rule would merely *permit* the sentencing of the innocent. And to have a rule which permitted individuals to sentence the innocent is not to say that they must or will sentence the innocent. This point must be granted, but it is of no real consequence. For, in general, we are clearly presupposing that there is some point to having a rule which forbids acts of a certain type, namely, that without that rule (and without that rule appropriately enforced) we would expect at least some individuals to perform acts of that type, i.e., without a rule forbiding the sentencing of the innocent we would expect some individuals to sentence the innocent. Hence if we did not have a rule which forbade the sentencing of the innocent, i.e., if we had a rule that *permitted* the sentencing of the innocent, we would expect at least some people to sentence the innocent. This brings me to the second objection, and it is this: that to *not* have a rule

which forbade the sentencing of the innocent is not to have a rule which *permitted* the sentencing of the innocent, for there are *no* permissive rules. Bernard Mayo has argued that there are no permissive rules on the grounds that "it is analytic that rules can be conformed to or infringed" but we cannot conform to or infringe a permission.[8] This point too may be granted and it indicates that we must sharpen up our expression. I take it that it would not be denied that if there is no rule that forbids the sentencing of the innocent, then an individual may infer that the sentencing of the innocent is permitted. That is, it is *implicit* in the body of rules in which there is no rule that forbids the sentencing of the innocent that the sentencing of the innocent is permitted. Thus our statement of the RUian justification of some rule should be put as follows: a rule that *forbids* acts of a certain type is justified if the consequences of having such a rule were better than were the consequences of not having such a rule, i.e., if the consequences of a rule that forbids acts of a certain type were better than the consequences where it could be inferred - where it was implicit in the body of rules - that acts of that type were permitted. And, as we have already said, we presuppose that if there was not a rule that forbade acts of that type then at least some individuals would perform acts of that type, i.e., if individuals were permitted to perform acts of that type

they would perform them. Conversely, a rule that *enjoins* acts of a certain type is justified under RU if the consequences of such a rule were better than the consequences where it could be inferred that *not* performing acts of that type was permitted.

But what we have said should not be taken to imply that AU has *no* place for rules of (moral) conduct. As Smart notes, we may well decided to choose according to certain rules, even if we subscribe to AU, but these rules will be merely "rules of thumb" or, as we shall call them, *rules of convenience*. For at least two reasons the AUian will maintain that we employ such rules. First, in many cases it will be extremely impractical to work out the consequences of an action. Indeed, the job of working out the consequences might be such that while the job was logically possible it might well outstrip human capacities.[9] All a utilitarian can require of an agent if he is to perform the act he ought to perform is that, as regards the consequences of the act, he has considered the consequences that it is *possible* for him to predict. Thus, to use a common example, the doctor who saves a certain baby does the right thing, even when the baby grows into the mature Adolf Hitler. The second reason that the AUian may give for the employment of rules of convenience is that in many cases the effort expended in working out the consequences of an act may give rise to a disutility

that outweighs the utility of those consequences. In short, the rules proposed by the AUian are necessary because of our human weaknesses in knowledge acquisition and computational skills. Of course, not just *any* rules will be permitted to function as rules of convenience by the AUian; but only those rules which, in the light of our knowledge about the world and ourselves, are such that in performing actions in accordance with them it is most probable that we will perform an action that has better consequences than any other action available. That is, according to AU we will adopt a rule as a rule of convenience if it is such that by acting in accordance with that rule we perform an act which is the act we most probably would have performed if, lacking our human weaknesses, we had been in a position to compute the consequences of that act.[10] Rules of convenience are not, however, inviolate: they are there to be broken. For should there be a situation where some act which was contrary to a rule of convenience had the best consequences, and it was practicable to work out that that was the case, then an individual *ought* to act contrary to the rule. As a consequence, if RU is to be a theory quite distinct from AU then the rules of RU cannot be merely rules of convenience. That is to say, the rules of RU *cannot* be rules such that if there is some individual act for which it is humanly feasible to work out that it has the best consequences and that this

computation does not constitute a waste of effort, then that act *ought* to be performed even though it is contrary to one of those rules.

Now the reader might think that this last statement has an odd ring about it; why, *as a utilitarian*, would one *not* perform an act which maximized utility even though it was contrary to some rule? Indeed, here lie the seeds of destruction for RU. If RU is to be a genuine advance on AU then RU must be significantly different from AU. However, it has been claimed that if RU is put forward as a *bona fide* utilitarian theory then RU is *not* significantly different from AU.

3. *The Equivalence Thesis for Rule and Act Utilitarianism.*

It has been claimed that RU and AU are extensionally equivalent, i,e., that whatever action is prescribed by the former is also prescribed by the latter, and *vice versa*. The idea here is that if proper attention is paid to the two-stage procedure proposed by RU for the justification of actions, then it will be seen that this procedure prescribes just those acts prescribed by the one-stage procedure of AU. I call this "The Equivalence Thesis". Now as we shall see (in section 4) this claim must be false. However, this is not to say those who claim that RU is not a significant improvement over AU are bereft of an argument to that effect. All that has to be shown

is that there are some acts prescribed by AU which also must be prescribed by RU and these acts are such that according to our moral intuitions no *moral* theory which prescribes these acts can be an *adequate* moral theory. My eventual aim will be to demonstrate just such a limited equivalence between AU and RU: i.e., I will argue for a *partial* extensional between AU and RU over a class of actions such that if AU is shown to be inadequate because it prescribes these acts, then RU is also inadequate because RU likewise prescribes these acts. Briefly, I will try to show that if AU is inadequate because on occasion it prescribes the sentencing of the innocent, then RU is also inadequate because it too must on occasions prescribe the sentencing of the innocent.

The argument I shall give which appears to establish the full equivalence thesis is essentially due to J L Mackie[11] and T L S Sprigge[12]. It proceeds by way of *reductio ad absurdum*. Henceforth, we talk of the action that has the best consequences as that which maximizes utility and for simplicity we ignore - as we have done already in this chapter - that in general there will not be an action that maximizes utility but only one that maximizes expected utility. We suppose, contrary to what we want to prove, that there is some action *A* which maximizes utility (and hence is prescribed by AU) but which is forbidden by RU. RU forbids *A* because *A* is

contrary to some rule *R* which is supposedly justified on utilitarian grounds. In other words, RU forbids *A* because if everyone acts in accordance with *R* which forbids *A* then this will give rise to greater utility than if there were no such rule and everyone acted accordingly, i.e., if everyone were permitted to do *A* and did *A*.

The act *A* we will say is of type *S*. Recall that the single performance of *A* maximizes utility whereas if everyone performs acts of that type, i.e., of type *S*, then this, supposedly, would not maximise utility. Therefore, there must be some feature of *A* or its circumstances which brings about the difference in utility and which distinguishes it from other acts of type *S*. Call this causally relevant feature *D*. Note that the feature might be no more than that the act is not performed by more than *n* individuals (where *n* may equal 1): i.e., the feature may be that acts of type *S* are not performed by *everyone*. We can now pick out a new class of acts which we will call *SD* acts, i.e., acts of type *S* which also have the feature *D*. Clearly, *A* is an element of that class. Let us formulate a new rule *R'* which enjoins acts of the class *SD* but *not* acts of the class *S non-D* (i.e., acts of type *S without* the feature *D*). In other words, we can think of our original rule *R* as saying, "Acts of type *S* ought not to be performed" and our new rule *R'* as saying, "Acts of type *S* ought not to

be performed, *except* when they are also *D"*. Now *R'* must maximize utility relative to *R*, i.e., if everyone acted according to *R'* rather than *R* this would give rise to greater utility. For consider: *R'* forbids all the acts that *R* forbids *except* those of the *SD* type which have a feature, *viz.*, *D*, which brings it about that they maximize utility. Hence, as a utilitarian the RUian should select *R'* over *R*. But, as we have seen *R'* enjoins *A*, and hence RU does *not* forbid *A*. Therefore, it cannot be the case that *A* maximizes utility and *A* is forbidden by RU.

Before going on to consider some objections to the equivalence thesis and thereby fleshing out the above argument, there is one point to which I wish to draw particular attention. The argument is addressed against those who regard RU as an *improvement* on AU. That is, it is addressed against those who regard RU as a theory of the same type as AU -- i.e., as a genuinely utilitarian theory -- and as a theory which is able to accommodate the putative counter-examples to AU. Our argument has the following structure: if in response to the putative counter-examples to AU the utilitarian advances RU then we can show, provided that the rules of RU are supposed to be justified on utilitarian grounds, that RU must also prescribe these acts which offend against our moral intuitions but which nonetheless

maximize utility. Now the supporter of RU could make recourse to either of the following two responses. First, he could claim that his theory is a utilitarian theory distinct from AU but that it was not designed nor was it capable of accommodating the counter-examples to AU. This response is not very interesting in the present context, for the motivation for introducing RU was precisely that it was an attempt to formulate a utilitarian theory which did not have the unfortunate implications of AU. Second, the RUian could claim that the rules of RU were not, or were not simply, justified on utilitarian grounds. But then such a theory hardly seems to deserve to be called "utilitarian" at all: rather, it is, to use a common phrase, merely a form of "rule worship".

So, if the RUian is to provide a defence of his theory he has to keep two things in mind *even supposing* he can show that his theory is distinct from AU: (a) his theory must be able to accommodate the counter-examples to AU, and (b) it must remain a recognizably utilitarian theory (it must not degenerate into mere rule worship). As will become apparent from my examination of attacks by RUians on the equivalence thesis I do not think that the above two requirements can be met simultaneously.

4. *A Decision Theoretic Approach which attempts to falsify the Equivalence Thesis.*

The argument against the equivalence thesis that I wish to consider here has been advanced by Harsanyi.[13] Harsanyi believes that the modelling techniques and other analytical tools of decision theory (and also game theory) can profitably be used to demonstrate the non-equivalence of AU and RU. About this I think Harsanyi is right. However, as I shall argue, I do not believe that Harsanyi has shown that RU is a significant improvement on AU: he has not shown that a limited version of the equivalence thesis is false. All the same Harsanyi's argument is a very important argument and it helps clarify our argument to the effect that RU is not a significant improvement on AU.

Harsanyi proposes a model of a moral decision problem which he then uses to determine the strategies of utilitarian agents. The notion of a *strategy* is a familiar one in decision theory and game theory: the actions of agents are determined by strategies which are conceived of as mathematical functions assigning one specific action to each possible decision situation, subject to the proviso that if two situations are of the *same* type they must have the *same* action assigned to them. In Harsanyi's model it is supposed that society consists of (n + m) individuals of whom 1,...,n are utilitarian

agents, and of whom $n + 1,\ldots,n + m$ are non-utilitarian agents. The choices of the non-utilitarian agents are determined by, say, self-interest or a non-utilitarian moral code (e.g., conventional morality). The strategies of these non-utilitarian agents are said to be *given*, i.e., constant. In other words, as regards these strategies any utilitarian agent must allow that they may not be the strategy he employs in some particular situation -- obviously, because what determines their strategy choice is quite distinct from what determines his strategy choice: all any utilitarian agent can do in any particular situation is to choose his strategy in the light of what he *expects* their strategy choice to be. Given this model we now ask, what do AU and RU require of the utilitarian agents in their choice of strategy?

Supposing that the maximand of utilitarianisim is social utility (i.e., the arithmetic mean of individual utilities) AU requires that each utilitarian agent i $(i = 1,\ldots,n)$ choose his strategy s_i in such a way as to maximize social utility *on the assumption that* all non-utilitarian strategies are given *and* that all utilitarian strategies are given. In contrast, RU requires that each utilitarian agent i choose his strategy s_i in such a way as to maximize social utility, *but* on the assumption that all other utilitarian agents will employ the same strategy, i.e., on the assumption that their strategies are *not* given, while all non-utilitarian strategies are given.

Letting S be the set of all strategies available to each agent (assumed to be the same set for each agent) and W be social utility we have that a utilitarian agent under AU or under RU must solve one of the following two quite distinct mathematical problems.

(A) Mathematical problem to be solved by i under AU:

Maximize $W = W(s_1, \ldots, s_i, \ldots, s_n; s_{n+1}, \ldots, s_{n+m})$

subject to the constraints

(A_1) $s_i \varepsilon S$

(A_2) $s_j = r_j = const.$ for $j = 1, \ldots, i-1, i+1, \ldots, n$

(A_3) $s_k = t_k = const.$ for $k = n+1, \ldots, n+m$

(B) Mathematical problem to be solved by i under RU:

Maximize $W = W(s_1, \ldots, s_i, \ldots, s_n; s_{n+1}, \ldots, s_{n+m})$

subject to the constraints

(B_1) $s_i \varepsilon S$

(B_2) $s_1 = \ldots = s_i = \ldots = s_n$

(B_3) $s_k = t_k = const.$ for $k = n+1, \ldots, n+m$

From this characterisation it is clear that AU and RU are certainly not logically equivalent and that they may lead to quite different moral decisions. But before proceeding further let me expand on what has already been said. Consider the mathematical problem (A): AU requires that a utilitarian agent i select his strategy so as to

maximize social utility which is equal to the social utility (i.e., the arithmetic mean) of each of the strategies employed by all individuals in society. However, he selects that strategy on the following assumptions. First, (A_1), it must be a strategy in the available set. Second, (A_2), it must be a strategy chosen on the basis of what strategies he expects the other utilitarian agents to employ: we suppose that i expects the other utilitarian agents, *viz.*, agents $1,\ldots,i-1,i+1,\ldots,n$, to use strategies $r_1,\ldots,r_{i-1},r_{i+1},\ldots,r_n$ respectively. Furthermore, this strategy will not necessarily be the strategy he employs; the strategies are constant. Third, (A_3), it must be a strategy chosen on the basis of what strategies he expects the non-utilitarian agents to employ; these likewise are constant. In contrast, under RU a utilitarian agent i must select his strategy so as to maximize social utility subject to a similar set of assumptions except that for (A_2) we have (B_2) which says that he must assume that the strategy he employs is the strategy to be employed by all moral (or, at least, utilitarian) agents. The utilitarian agent i is then selecting his strategy as if he were selecting a strategy for *all* utilitarian agents providing it is the *same* strategy for each. This substitution of (B_2) for (A_2) simply reflects the idea that a utilitarian under RU must, when deciding what he ought to do, do so by considering what would be the consequences for social

utility supposing that everyone (or, at least, everyone who is doing what they ought to do) were to perform that act. There is, of course, no such requirement on a utilitarian under AU.

We can demonstrate that AU and RU will give rise to different decisions by considering the following simple decision situation presented by Harsanyi.

Example: consider a society of 1000 voters who are asked to vote on some socially important measure *M*, but voting involves some minor inconvenience. All voters are in favour of *M* and *all* are assumed to be utilitarians. Suppose *M* will pass only if *all* voters actually vote. The voting situation has the nature of a game with 1000 players in which each player *i* tries to maximize *W*. In the first instance, suppose each player *i* chooses under AU. Clearly, he will vote only if he is reasonably sure that all the other 999 voters will vote — otherwise he will not vote. That is, which strategy *i* employs will depend on what he expects each of the other players to employ: if *i* expects all the other 999 voters to vote, *i* will vote as this, given his expectation, will maximize *W*; but if *i* expects at least one of the other 999 voters not to vote, *i* will *not* vote as this, given his expectation, will maximize *W*. In the terminology of game theory, this game has *two*

Nash equilibrium points: one where everyone votes and one where no-one votes. (A Nash equilibrium point can be understood informally as the case where no player finds it to his advantage to change to a different strategy given that he expects the other players not to change.[14]) On the other hand, if each player *i* chooses under RU, there is only *one* Nash equilibrium point -- only one possible outcome -- namely, where everyone votes. For consider: *ex hypothesi* everyone favours *M* and there is only a minor inconvenience associated with voting, hence it would be better if everyone voted than if at least one person did not vote. That is, it would be better if there was a rule which enjoined voting than if there was no such rule.

Now given a plausible assumption about the subjective probability that a player *i* under AU will assign to all the other utilitarians voting it will be the case that in the above situation no-one votes. Thus, in this situation, RU is shown to be a *superior* theory to AU in the sense that utilitarians choosing under RU can be sure of a *global* maximization of social utility. That is, they can be sure that they will do better under RU than under AU. And it is not difficult to see why this should be so. Because RU requires that any player *i* select his strategy so as to maximize social utility on the assumption that *all* utilitarian agents will employ

the *same* strategy, RU will be far more effective than AU in securing socially desirable *co-operation*, i.e., it will be far more effective in securing *co-ordination* of strategy choice when it is necessary that everyone (or, at least, every moral person) employ the same strategy in order to maximize social utility. RU is superior to AU in respect of what Harsanyi calls "the coordination effect".

But now it is necessary that we recall my remarks at the end of the previous section. There I said that even supposing the RUian has shown that his theory is distinct from AU, he must also show that his theory is able to accommodate the putative counter-examples to AU without his theory degenerating into mere rule worship. Now I think Harsanyi has certainly shown that RU is a distinct theory from AU -- he has shown that on occassions it determines a different choice of strategy -- and, indeed, that in certain situations RU is a superior theory to AU. But has he shown that RU is able to accommodate the counter-examples to AU?

To answer this question consider the fact that the sort of situation examined by Harsanyi is a situation where it is necessary if there is to be a global maximization of social utility that everyone employ the same strategy: it was a situation where it was socially desirable to co-ordinate strategy choice. But a moments

thought will make apparent that not all situations where we are attempting to decide what we ought to do will require a co-ordination of strategy choice in order to maximize social utility. For example, take the case of McCloskey's sherriff; his problem is not to co-ordinate strategy choice with other agents; for him to maximize social utility it is not necessary that he co-ordinate his activities with others. Therefore, the obvious question to ask is whether an RUian will choose differently to an AUian in situations which do *not* require the co-ordination of strategy choice to achieve a global maximization of social utility? In other words, we are asking whether the RUian will choose differently to the AUian in situations like that of McCloskey's sherrif? Now unless we have a positive answer to that question we have not shown that RU can accommodate the counter-examples to AU. Harsanyi himself has already anticipated this objection (he puts the point in relation to promise keeping):

> Will the two versions of utilitarianism reach different conclusions about the conditions under which promises ought to be kept? Surely, if they are to reach different conclusions at all, this will have nothing to do with the co-ordination effect. Admittedly, we do sometimes make

> promises which can be fulfilled only by our undertaking co-ordinated efforts with other people, But the moral problem posed by promise making would not essentially change if we never made promises that could be fulfilled only by such co-ordinated activities, so that the possibility of a *co-ordination effect* (i.e., of co-ordination with other people who have promises to fulfill) would not even arise.[15]

For us to see that RU will prescribe different courses of action to those prescribed by AU as regards the sentencing of the innocent or the breaking of promises, Harsanyi believes that we must take cognizance of what he calls "the expectation and incentive effects". By the *expectation effect* he means the effect the adoption of any given strategy by the utilitarian decision maker will have on the expectations, the ability to form definite expectations, and the feelings of confidence and security of other agents in society. By the *incentive effect* he means the effect the adoption of any given strategy by the utilitarian decision maker will have on the other agents' incentive to engage in various types of socially beneficial behaviour. To illustrate the expectation and incentive effects Harsanyi

concentrates on the example of promise keeping. I too will keep to this example for the moment, merely remarking here that the argument presented below is equally applicable to the question of punishment raised in McCloskey's sheriff example.

The crucial assumption from the point of view of the *co-ordination effect* was that the decision maker would choose his strategy on the assumption that this was the strategy to be employed by all utilitarian agents: this assumption followed from the very definition of RU. The crucial assumption from the point of view of the *expectation* and *incentive effects* is that all agents, whether utilitarian or not, will know and will act on the knowledge, that the decision maker will use the strategy that is optimal under RU. This assumption follows from the fact that any agent can compute the optimal strategy under RU by solving the appropriate maximization problem.[16] Note then that the assumption that is crucial for the expectation and incentive effects does *not* follow from the very definition of RU. Harsanyi emphasises that this assumption

> is not a *casual postulate* about physical transmission of information from some agents to some other agents, but rather is a quasi-logical postulate about the nature of

> optimal strategies, and about free access of information to all agents concerning the nature of these optimal strategies.[17]

I will make further comment on this assumption in a moment. But for the present we simply point out that it is this assumption which gives rise to the expectation and incentive effects. Suppose that a utilitarian decision maker under RU has adopted a strategy which, say, permitted "many easy exceptions to promise keeping"[18]; then if the other agents knew that that strategy had been adopted

> then they would have much less ability to form definite *expectations* about (the decision maker's) future behaviour, and in general would feel less confident and less secure about the future.
>
> They would also have much less *incentive* to plan their future activities on the expectation that promises made to them would be kept (e.g., that their friends would actually turn up at the places and times they had promised to). Similarly, they would have much less incentive to perform useful services for other people on the mere basis of promised future

> rewards, without any immediate compensation, etc.[19]

There are two respects in which I think Harsanyi's argument is mistaken. The first objection concerns the crucial assumption which was necessary to induce the expectation and incentive effects. Consider, as Harsanyi recognizes, that even under AU a decision maker must consider the unfavourable consequences of some individual act, say, some act of promise breaking. So, we might ask, how can there be any difference between RU and AU with respect to the expectation and incentive effects? Harsanyi's answer is that

> barring some very special situations, the causal consequences of one isolated act of promise breaking will be very, very small, because people will not infer -- and cannot rationally infer -- from *one* such act that promise breaking has suddenly become a *general* practice in their society.[20]

In contrast, according to RU,

> if this strategy were in fact the *optimal* rule utilitarian strategy, then all interested

> parties would *know* that this strategy would represent the *general* practice in matters of promise keeping.[21]

These comments make clear the essential role played in Harsanyi's defence of RU by the assumption that *all agents, whether utilitarian or not, will know and will act on the knowledge, that the decision maker will use the strategy that is optimal under RU*. If this assumption did not hold there would obviously be no difference between AU and RU as regards the expectation and incentive effects. Now I do not dispute that this assumption is not a causal postulate about the physical transmission of information from some agents to some other agents, it is, rather, a quasi-logical postulate in that it will hold if it is the case that any agent can compute the optimal strategy under RU by solving the appropriate maximization problem. But we should now consider the following: suppose the decision maker -- a utilitarian agent acting according to RU -- knows or has good reason to believe that it is *not* the case that any other agent (or that a sufficiently large number thereof) can compute the optimal strategy under RU by solving the appropriate maximization problem. The decision maker could know, for example, that no one else in society had the requisite computational skills to solve the appropriate maximization problem. And further suppose

that the decision maker knows that the information as to what strategy he employs will not be physically transmitted to the other agents. Thus the other agents will not know what strategy the decision maker has decided to employ either by the quasi-logical means of solving the appropriate maximization problem nor by the normal causal/physical means. But if so, it will not be the case that all agents, whether utilitarian or not, will know and be able to act on the knowledge, that the decision maker will use the strategy that is optimal under RU. That is, the essential assumption for the inducement of the expectation and incentive effects would not hold. And in that case a decision maker could adopt a strategy which permitted exceptions to promise keeping knowing that the adoption of such a strategy would *not* have disastrous effects on the expectations and incentives of other people in society. Now I am not saying that it *must* or even *generally* will be the case that no-one else in society knows either by quasi-logical means or physical means what strategy the decision maker has adopted but only that it *might* be the case, and that in such a case a decision maker *even choosing under RU* should choose a strategy which permits of exceptions to promise keeping. But I take it that our intuition would be that whether we ought or ought not to keep a promise is not thus dependent on other people's ability to solve certain

computational problems or on "gaps" in the causal chain of information flow to those agents. In other words, I take it that we do not believe that I (say) morally ought to adopt a strategy which permits me to break promises just because I know that no one in society knows that that is the strategy I have adopted because everyone else in society lacks the intellectual capacity to solve the appropriate maximization problem and cannot find out what that strategy is by other means. To fix this idea, suppose that McCloskey's sheriff is attempting to decide on what strategy to adopt in this situation which involves the sentencing of this innocent man. Suppose that by solving the maximization problem appropriate for AU he determines that as an AUian in this situation he ought to sentence this innocent man. Now if as a RUian he is to adopt a *different* strategy in this situation then the assumption must hold that everyone else will know and will act on the knowledge that he will use the RUian optimal strategy. For it is from the fact that this assumption holds that the expectation and incentive effects are supposed to arise. But suppose this assumption does *not* hold -- which is, as I've said, surely logically possible -- then he will adopt exactly the same strategy as he would if he were choosing as an AUian. But such a strategy offends against our sense of justice: the RUian is committed to the view that on

occasions it would be just to punish an innocent man, namely, on just those occasions where for various reasons no-one else in society knows what strategy it is that he will employ and where to sentence this innocent man maximizes utility.

But even supposing that the crucial assumption for the inducement of the expectation and incentive effects holds it is still possible to show that a utilitarian choosing under RU is committed to views which offend against our moral intuitions. Note that Harsanyi asks us to consider what would be the effects on people's expectations and incentives supposing that they knew that the decision maker had adopted a strategy which permitted *"many easy exceptions to promise keeping"*; that is, we are asked to consider what would be the effects supposing that a strategy which represented a *general* practice of promise breaking was in fact the optimal RUian strategy. Now if we grant that everyone in society will know and will act on the knowledge that that strategy has been adopted then the adoption of such a strategy will have disastrous expectation and incentive effects, and hence could *not* be the strategy required of a utilitarian decision maker under RU. But why suppose that the optimal RUian strategy permits of *many easy exceptions* to promise keeping or that the strategy adopted would represent a *general* practice of promise breaking? To

revert to McCloskey's sheriff example, why suppose that the strategy he adopts as a RUian permits of many easy exceptions to not sentencing the innocent or that he adopts a strategy which would represent a general practice of sentencing the innocent in matters of punishment? Consider the following two points.

First, it may be that those situations where to break a promise or to sentence an innocent man will maximize utility are situations which as a matter of contingent fact, are very rare because the feature or features of those situations which bring about the maximization of utility are as a matter of fact very rare. Now if this were so then surely the strategy for the RUian to adopt would be one that permitted exceptions to promise keeping or not sentencing the innocent in just those very rare circumstances. (We may note that on McCloskey's account we would hardly expect to actually come across a case where on utilitarian grounds it would be just to sentence an innocent man -- such a case is thought only to be logically possible.) More fully the idea is as follows. Recall that a strategy is a mathematical function that assigns a specific action to a specific situation, with the proviso that if two situations are of the same type then they must have the same action assigned to them. (Harsanyi says two situations are of the same type when the decision maker cannot, with the information available

to him, distinguish between the situations.) We will distinguish two *different* types of situation that fall under a general type. The *general* type of situation will involve sentencing or punishing people. The two distinguishable *sub-types* will be situations that (1) involve sentencing the innocent where to do so maximizes utility and (2) involve sentencing the innocent where to do so does *not* maximize utility. (Of course, these two sub-types are not *exhaustive* of the general type, for example, there are the sub-types that involve the sentencing of the guilty. However, as will become apparent, this fact has no bearing on the rest of my argument.) That is, we suppose that there is some feature, the feature *D* to which we have referred previously, that situations of type (1) have and which situations of type (2) do not have which brings it about that to sentence an innocent man maximizes utility. As situations which fall under type (1) are not of the same type as type (2) it is not required that the same (type of) action be assigned to them. Thus if a strategy assigns a certain action to the first type of situation it need not assign that (type of) action to the second type. Now we suppose that situations of type (1) are very rare because, as a matter of fact, situations with the feature *D* are very rare. And we further suppose that a decision maker under RU adopts a strategy which

requires that in situations of type (1) he sentence the innocent. He is not, of course, required by the adoption of that strategy to sentence the innocent in situations of type (2) and thus, more particularly, he is not thereby required to sentence the innocent in all situations of the *general* type which involve the sentencing or punishing of people. That is, if he adopts a strategy which requires the sentencing of the innocent in situations of type (1) he is not required to sentence the innocent in all situations that involve sentencing or punishing people. We suppose, then, that a decision maker adopts a strategy which requires that he sentence the innocent in situations of type (1) *only*. Now even if we allow that all other individuals will know and will act on the knowledge that the decision maker has adopted just such a strategy, why should this knowledge have disastrous expectation and incentive effects? Afterall, they cannot infer that in matters of punishment that it is at all probable that as innocent persons they will be sentenced. *Ex hypothesi*, the strategy adopted requires the sentencing of the innocent in only very rare circumstances, indeed, in circumstances which may never *actually* occur but which are merely *logically* possibly. That is, in knowing that such a strategy has been adopted the other individuals cannot infer that sentencing the innocent has become a *general* practice in

matters of punishment. But is the RUian required to adopt such a strategy? Surely the answer is, "Yes". For such a strategy will require that in situations of type (1) the innocent are to be sentenced where *ex hypothesi* to do so maximizes utility; such a strategy does not require that the innocent be sentenced in situations of type (2) for the situations are not of the same type; and as in situations of type (2) to sentence the innocent does *not, ex hypothesi,* maximize utility a utilitarian ought not to adopt a strategy which required the sentencing of the innocent in situations of that type. Notice that even though the RUian by the very definition of RU, must make his strategy choice on the assumption that all individuals *ought* to adopt that strategy (i.e., on the assumption that all utilitarian agents *will* adopt that strategy) the knowledge that the above strategy is the optimal RUian strategy will still not have disastrous expectation and incentive effects. We can illustrate this point by way of the following example. Suppose that those situations where to sentence the innocent will maximize utility are so rare that as a matter of fact any sheriff can only be expected to come across such a situation in 1 out of 100,000 cases of sentencing people. I have suggested that the optimal RUian strategy will be that which requires the sentencing of the innocent in just these

rare circumstances. Consequently, even supposing that all other individuals in society know that *all* (utilitarian) sheriffs will adopt that strategy this would not have seriously detrimental effects on those individuals' expectations and incentives. And this because it is highly improbable that they will be sentenced when innocent.

If this argument is correct then it follows that RU can be brought into conflict with our moral intuitions even allowing that the crucial assumption for the inducement of the expectation and incentive effect holds. For the RUian is committed to the view that we ought to adopt a strategy (rule) which requires the sentencing of the innocent just when situations where to sentence the innocent would maximize utility are very rare (and maybe only logically possible).[22]

The second and, I think, more important point to be considered against Harsanyi arises from the fact that, as I mentioned in the argument for the equivalence thesis in section 3, the causally relevant feature D which brings about the maximization of utility might be no more that that the act A is not performed by more than n individuals (where n is some small number). So there we hypothesised a rule which said that an individual ought to perform the act A *provided* that no more than n individuals have or will perform that act. The idea here is that at least on occasions to break a promise or to sentence an

innocent man will maximize utility although if such acts became a *general* practice, i.e., if in matters of promising or punishment, promise breaking or sentencing the innocent became common practice, this would not maximize utility. Hence, it would be desirable on utilitarian grounds to have a rule which permitted the breaking of promises or the sentencing of the innocent *provided* such a rule did not allow such acts to become the norm or to become common in matters of promising or punishing. Even if we now suppose that everyone else in society knows that the RUian decision maker has adopted such a rule we could not conclude that the adoption of such a rule would have detrimental expectation and incentive effects or, at least, the effects might be so minimal as to be outweighed by the utility gained by, e.g., the sentencing of the odd innocent person. The reason that the knowledge that such a rule had been adopted would not have detrimental expectation and incentive effects is much as before, *viz.*, that the individuals could not infer that in, e.g., matters of punishment it was at all probable that they would be punished if innocent. But the adoption of such a rule is impossible on Harsanyi's account of RU, for it follows from the very definition of RU that a decision maker choosing under RU must adopt a strategy on the assumption that that is the strategy to be adopted by *every* person

choosing as they ought to choose, i.e., by every person choosing as a utilitarian, in that sort of decision situation. But this just goes to show that RU is inadequate *as a utilitarian* theory -- the theory advanced is not genuinely a utilitarian theory at all. For if a theory by definition will not permit *some* but not *all* individuals to adopt a certain strategy when *ex hypothesi* to do so would give rise to a greater utility than if *no* individuals were to adopt that strategy then that theory should be rejected *on utilitarian grounds*. The supporter of RU is caught in the horns of a dilemma by the sort of criticism I have addressed against his theory in this paragraph: either in response to the criticism he can stick with his theory and thus claim that his theory does not commit him to the view that at least on occasions what morally ought to be done is to sentence the innocent -- but then his theory is not a genuine *utilitarian* theory, or he can insist that his theory is a genuine utilitarian theory -- but then his theory is seen to commit him to views which offend our moral intuitions, e.g., that at least on occasions what ought to be done is to sentence the innocent. And as I remarked at the conclusion of section 3 even if the supporter of RU has shown that his theory is in some way distinct from AU he must also show (a) that it is able to accommodate the counter-examples to

AU, i.e., it must not commit him to the view, e.g., that it would sometimes be right to sentence an innocent man, and at the same time he must show (b) that it remains a genuinely utilitarian theory. In response to the sort of criticism I have made in this paragraph it would seem that Harsanyi cannot meet both of these requirements.

That completes my criticism of Harsanyi's argument that RU is a significant improvement on AU. But before turning to other attempts to show that a supporter of RU is not committed to the morally objectionable views to which the supporter of AU is committed we must turn to the argument for the equivalence thesis as presented in section 3. At first blush that argument seemed to establish a *complete* extensional equivalence between RU and AU; i.e., whatever action was prescribed by AU would also be prescribed by RU and vice versa. But in this section when discussing the *co-ordination effect* we have seen that that claim must be false and more importantly that RU is a superior theory to AU in certain circumstances in that it ensures a global maximization of utility in those circumstances. It is important therefore for us to understand where our initial assessment of that argument went wrong. Such an understanding is not hard to come by. For clearly in that argument we presupposed that the individuals were not in a situation where in order to achieve a global maximization of utility all individuals must employ the

same strategy. That is, we presupposed that the situation was not one where the *co-ordination effect* would be at all significant. In other words we presupposed that while RU forbids *A* it was not the case that a global maximization of utility could only be achieved if *everyone* did not do *A*. Hence we asserted that a maximization of utility would ensue if acts of type *S* (to which *A* belongs) were performed, but not performed by everyone. But this presupposition, while important, does not seriously affect the argument which concludes that RU is not a significant improvement on AU. For all that argument attempts to establish is that RU prescribes acts in certain logically possible situations which offend against our firmly held moral intuitions. The moral intuitions appealed to here to relate to matters that concern our common beliefs as regards, e.g., justice (e.g., our belief that the innocent ought not to be sentenced). Now as Harsanyi admits, and as we have seen, whether RU will result in different prescriptions to AU in matters of justice cannot have anything to do with the co-ordination effect. Hence even if it has been shown that RU is a distinct theory to AU and, indeed, is a *superior* theory to AU in situations where the co-ordination effect will be evident, this does *not* show what is of the main interest, namely, that RU is a significant improvement on AU in matters relating to

justice. That is, it will *not* have been shown that RU does not, unlike AU which *does*, prescribe acts in certain logically possible situations which offend against our intuitions as regards justice. To establish this we need some other argument. And as we have seen Harsanyi employs an argument which involves an appeal to what he calls the expectation and incentive effects. But I have argued that this argument is mistaken on at least two counts and it is worthwhile here to briefly restate my arguments to make clear just precisely what is and is not established by each of them. First, I claimed that Harsanyi's argument rests on an assumption, *viz.*, that all individuals in society will know and will act on the knowledge that the decision maker adopts the optimal RUian strategy, which may not hold. (We can grant that this assumption -- sometimes called "the assumption of mutually expected rationality" -- is commonly assumed to hold in decision theory and game theory. But it seems to be merely a matter of commonsense to realise that this assumption does not or, at least, need not hold in everyday life.) Of course, with this argument we have *not* established that *whenever* AU prescribes a particular act, RU prescribes that act: We have only established that RU prescribes the intuitively objectionable acts that AU prescribes *given*, as is certainly possible, that a certain assumption does not hold. But this is surely enough for those who claim

that Ru is not a significant improvement on AU for, as we remarked, whether this assumption does or does not hold seems to be irrelevant for the moral assessment of those actions. Second, even if we allow that this assumption *does* hold -- i.e., even if we take a purely decision theoretic approach to the attempt to demonstrate the putatively important difference between AU and RU -- we can show that Harsanyi's argument about the expectation and incentive effects will not suffice to show that RU is significantly different from AU. There are two points here: first, the situations where, e.g., to sentence an innocent man will maximize utility may, as a matter of fact, be extremely rare and hence a strategy which prescribed the sentencing of the innocent in just those rare circumstances would not have seriously detrimental expectation and incentive effects. Once again we have not established with this argument that *whenever* AU prescribes a particular act, RU prescribes that act: we have only established RU prescribes the intuitively objectionable acts that AU prescribes *given*, as is certainly possible, that where to perform these acts will maximize utility is extremely rare. And once again this seems sufficient for those who wish to claim that RU is not a significant improvement on AU: afterall RU is committed to the view that on occasions we ought to perform an injustice given that an apparently

morally irrelevant assumption holds. The second point is that a policy as regards punishment which permitted some (possibly *very* few) individuals to sentence the innocent would conceivably maximize utility and would not have seriously detrimental expectation and incentive effects, and if a supposedly utilitarian theory rules out such policies by definition, then so much the worse, from the utilitarian point of view, for that theory. Now with *this* argument we are able to claim that RU will prescribe (if it is a genuine utilitarian theory) the intuitively objectionable acts that AU prescribes *without* supposing that certain contingent facts hold, i.e., without supposing that the other individuals in society do not know what is the optimal RUian strategy or that the situations where to perform one of these acts will maximize utility are very rare. Our argument for the extensional equivalence of AU and RU over these class of acts only requires that the RUian present his theory as a genuinely utilitarian theory.

We can therefore conclude that Harsanyi's attempt to show that RU is a significant improvement on AU by using the modelling techniques and other analytical tools of decision theory is unsuccessful. But as I mentioned there have been *other* attempts to show that RU is a significantly different theory from AU and it is to these that I now turn.

5. *Other attempts to falsify the Equivalence Thesis.*

There are two arguments that I propose to focus upon in this section. The first has been advanced by Gertrude Ezorsky[23]. In her defence of RU Ezorsky asks us to consider what it is about actions that interests the RUian. They are concerned with the consequences of *social practices*, i.e., they are interested in "what would happen if certain kinds of actions were performed by everyone in a social group"[24]. And as we have noted there are certain types of actions such that if they became a *general* practice the consequences would be disastrous. Now with respect to these kinds of actions we suppose that there is a tendency for people to perform these type of actions unless there were an (enforcable) rule which forbade such actions. This, as we have already noted, was the central idea that underlay the introduction of the rules of RU. But, then, Ezorsky claims, if some action is to be the subject of a rule under RU, i.e., if it is to be an action that is forbidden by a rule of RU, then it must be possible for the action to be "contagious, or universalizable".[25] Ezorsky puts the point in terms of the properties of actions, and, in particular, the property of an action (or its circumstances) which brings about the maximization of utility -- these she calls "consequential properties". Thus she says, the RUian

> has a right to demand that a consequential property pass a test for being possibly contagious, or universalizable. Otherwise it isn't the sort of property he is interested in.[26]

Properties which do not admit of being possible contagious or universalizable Ezorsky calls "discriminatory". A property of an act is discriminatory if and only if one or both of the following suppositions is self-contradictory:

> (1) All members of the group perform acts exemplifying that property.
> (2) All acts of the general kind which is further specified by the property, exemplify the property.

Let us now consider the act of sentencing an innocent man which has the property which brings about the maximization of utility -- the property or feature which we earlier referred to as the property or feature *D*. Now it is not, as Ezorsky notes[27], self-contradictory to suppose that all members of a social group should perform such acts. For we could suppose that everyone in a social group performed just one such act out of (say) 100,000 acts of sentencing each, and such a practice

would not have detrimental effects -- in Harsanyi's terminology, such a practice would not have disastrous expectation and incentive effects. (This is just the idea we introduced in the previous section, *viz.*, that from the utilitarian point of view a practice or policy as regards punishment should permit of exceptions to not sentencing the innocent providing such exceptions were relatively few and far between. This will insure that individuals cannot infer from such a practice or policy that it is at all likely that as innocent persons they will be sentenced. For if they cannot infer this then such a practice or policy cannot have (seriously) detrimental expectation and incentive effects.) However, it *is* self-contradictory to suppose that all acts of sentencing should be acts of sentencing the innocent with the feature *D*. For if all acts of sentencing were acts of sentencing the innocent this would have disastrous effects. Therefore the feature *D* is discriminatory, and hence it is impossible that acts of the type "sentencing the innocent with the feature *D*" should ever become a social practice. This, according to Ezorsky, "enables the RUian to snip off the irrelevant maximizing circumstance" and specify what the sheriff did as simply a case of sentencing the innocent and "the generalized consequences of doing *that* are disastrous". Hence, the RUian will adopt a rule which forbids what the sheriff is doing, namely, sentencing the innocent. Ezorsky

concludes by saying that her argument

> frees RU from coextension with AU. The RUian's moral disquisitions can be different from (and better than) those of his AUian predecessor.[28]

Now the reader may have, as I have, some difficulty in following this argument for the argument seems to establish precisely the opposite of what Ezorsky wants it to establish. Let us grant that Ezorsky has correctly specified the necessary and sufficient conditions for a property of an act being discriminatory, i.e., for it being possible that an act should become a social practice. (It may be disputed that Ezorsky *has* specified the necessary and sufficient conditions for a property of an act being discriminatory, but to investigate this further would take us too far afield into an investigation of what constitutes a social practice.) Now we noted, at the beginning of this section, that if it is the case that an action is forbidden by a rule of RU, then it must be possible for the action to be contagious or universalizable, i.e., it must be possible for the action to become a social practice, i.e., it must not have a property or properties that are discriminatory. Ezorsky has shown that the property *D is* discriminatory.

Hence the plain philosopher might have thought that by a simple application of *modus tollens* we can conclude that it is *not* the case that an action with that property is forbidden by a rule of RU. Which seems to be precisely the conclusion that those who argue that RU is *not* a significant improvement on AU would want to arrive at. That is, from Ezorsky's argument we seem to be able to conclude that an action such as sentencing an innocent man where such an action has the property which brings about a maximization of utility is *not* forbidden by a rule of RU. In other words, we seem able to conclude that such an action is permitted by the rules of RU.

However, it might be thought that Ezorsky could reply as follows. We note that if an action maximizes utility, e.g., if this act of sentencing an innocent man has the feature *D*, then that act *ought* to be performed according to AU, i.e., that act is *prescribed* by AU. But the conclusion of our argument above is merely that such an act is *permitted* by RU -- there is no positive prescription to the effect that the act *ought* to be performed. So here, it might be thought, is a difference between AU and RU: according to AU the act (which has the feature *D*) of sentencing this innocent man is prescribed, whereas that act is merely permitted by RU. There are two points to be made in reply to this argument.

First, it is surely cold comfort to the supporter of RU that he should find that his theory does not prescribe acts like the sentencing of the innocent and does not forbid them, but actually permits such acts. I take it that our intuition is that we ought not to prescribe such acts -- and hence any theory which *does* prescribe them is inadequate; *and* that we ought not to permit such acts -- and hence any theory which *does* permit them is likewise inadequate. Second, if our argument above is correct then the supporter of RU must admit that there is no rule in his theory which will forbid such acts, and as such an act *ex hypothesi* maximizes utility then as a utilitarian he ought to prescribe such acts. This brings me to the final reply that I think that Ezorsky might make by way of defence of her argument.

Consider the following statement of Austin's which we have quoted previously: according to RU "The question to be solved is this:- If acts of the *class* were *generally* done ... what would be the probable effect on the general happiness or good?" This suggests the following idea: if, according to RU, an act ought to be done, i.e., if there is to be a rule that prescribes such acts, then the consequences of that act's becoming a social practice must be maximally beneficial. Thus, if RU is to prescribe a certain act then it must be

possible for that act to become a social practice. Now as it is not possible that an act of sentencing an innocent man which has the feature *D* should become a social practice, it follows that a RUian is *not* committed to prescribing such acts. Hence it is not correct to assert, as we seemed to assert in the previous paragraph, that an RUian is committed to prescribing such acts. But this reply is problematic in a number of respects.

Note that this reply rests on the idea that if RU is to *prescribe* certain acts then it must be possible for such acts to become a social practice. This is in contrast to the idea that informed our discussion of Ezorsky, namely, that if RU is to *forbid* certain acts then it must be possible for such acts to become a social practice. And it is worth remarking that as Ezorsky actually states her argument it would appear that she is actually interested in what acts RU can *forbid*, not what acts RU can *prescribe*: she is concerned to argue that a RUian can claim than an innocent man *ought not* to be punished (even when to do so maximizes utility). Now this might prove to be no problem for Ezorsky if it could be argued that given that a theory does *not prescribe* an act it follows that it *forbids* that act. For then it could be argued that as RU does not prescribe the sentencing of the innocent, it must forbid the

sentencing of the innocent. Such an argument has a conclusion which is reminiscent of the Aristotelian dictum "what the law does not prescribe, it forbids"[29], and it has long been recognized that such a view is, to say the least, problematic. We would have to suppose that RU is capable of generating a set of rules which would exhaustively divide actions into those that were prescribed and those that were forbidden: that is, the rules of RU would have to be supposed to not admit of a class of actions that were merely permitted. This is not the place to enter into this debate except to notice that Bernard Mayo has argued, to my mind convincingly, that there can be no body of rules which can prescribe or prohibit every action[30]. It is enough for our purposes to simply note the following: if we allow that the rules of RU do not admit of a class of actions that are merely permitted, then given our previous argument that RU does not forbid certain actions and given the *symmetrical* argument presented in this paragraph that RU does not prescribe these actions, we have a contradiction, namely, that RU forbids and prescribes these actions. Hence Ezorsky must conclude that such actions are merely permitted by RU. And this still leaves it open for me to make the point that I made previously. If certain acts are permitted by RU and these acts maximize utility then *as utilitarians*

we should prescribe such acts. That is, even granting the insights of RU, namely, that we should always bear in mind the consequences of an act should it become a social practice, if there is nothing in the theory of RU to forbid certain acts (because it is impossible that they should become a social practice) and these acts maximize utility, then on simple utilitarian considerations we should prescribe such acts. Maybe as an *RUian simpliciter* the supporter of RU is not committed to prescribing such acts, but as a *utilitarian* he seems to be so committed.

I said at the beginning of this section that there were two arguments addressed against the equivalence thesis that I was going to consider: the second argument is due to Mackie[31]. Mackie argues that the argument we gave in section 3 for the equivalence thesis is only decisive if the rules in RU are treated as "purely abstract entities" rather than as "social realities". For a rule to be a social reality -- a rule that is taught and passed on from one generation to another, a rule that is more or less consciously accepted and followed, appealed to in criticisms of violations, etc. -- "there are limits to the complexities and qualifications it can incorporate". Now the rule R may be such a rule, but there is no guarantee that the rule R' required to instantiate the equivalence thesis will

meet these requirements. Hence, if the RUian insists that the rules of his theory be social realities then, as Mackie concludes, "Rule utilitarianism, thus understood, can therefore resist the threatened collapse into act utilitarianism."

However, this initially attractive argument fails because Mackie's argument has the consequence that the logical status of the rules of RU is reduced to that of the rules of AU; i.e., they become merely rules of convenience (see section 2). To see this note that the reason Mackie cites for the rule R rather than the rule R' being adopted is essentially that the second may be too complex and include too many qualifications for it to be a rule that could function as a social reality. That is, its complexity may well outstrip human capacities for it to be the sort of thing that can be taught and passed on from one generation to another, etc. Nonetheless, it is granted that if the rule R' rather than R were adopted this would give rise to greater utility. The reason R rather than R' is adopted is just that, due to human weaknesses, the latter would outstrip our capacities for assimilation and manipulation: there is no logical necessity that R should be adopted rather R'. But recall that we noted that it was not the case that AU had no place for rules of moral conduct -- we called these "rules of convenience". We further noted that if RU was to be a theory quite distinct from AU

then the rules of RU could not be merely rules of convenience. Now the reason that an AUian would give for the adoption of a certain rule, say, R is just that, due to human weaknesses, it would be beyond our capacities to work out the consequences of individual acts: it may not be humanly feasible to work out those consequences or to work out the consequences may constitute a waste of effort (given *our* computational skills). However, there is no logical necessity that R should be adopted rather than that individuals should work out the consequences of each individual act -- indeed, to do so, if it were possible, may well give rise to greater utility. But now it seems that the logical status of the rules of RU, *according to Mackie's account of those rules*, is identical to their status under AU: they are both merely rules of convenience. For under AU the rule R, for example, is adopted because of human weaknesses -- but for limited human capacitites we could bring about greater utility by working out the consequences of each individual act. Similarly, according to Mackie, the rule R is adopted because of human weaknesses -- but for limited human capacities we could bring about greater utility by adopting the rule R'.

We seem entitled to conclude, therefore, that the argument to the effect that RU is not a significant improvement over AU still stands. Hence the utilitarian

cannot avoid the putative counter-examples to AU by introducing the distinction between RU and AU. But note that I have called these "*putative* counter-examples". It remains an open question as to whether we have demonstrated the inadequacy of utilitarianism by showing that it can be brought into conflict with some of our most firmly held moral intuitions. It is to this question that I now turn.

6. *Do Conflicts between our Moral Intuitions and Utilitarianism really matter?*

There are, it seems to me, two responses that the utilitarian can make to the above question. Each admits that utilitarianism can be brought into conflict with some of our moral intuitions but there are important differences in how they deal with this conflict. The first can be summarised by saying, "If utilitarianism can be brought into conflict with our moral intuitions, then so much the worse for those intuitions". Such a response is suggested by some of the remarks made by Smart[32]. The second can be summarised by saying, "The conflict between our intuitions and utilitarianism is more apparent than real - the intuitions appealed to by the anti-utilitarian have a limited legitimate application and their application in the the attempt to demonstrate

the inadequacy of utilitarianism is illegitimate". This response has been made by Sprigge[33] and R M Hare[34]. I will examine each of these responses in turn.

The first response has two related ideas underlying it. First, that it is clear that many of our intuitions, not just in the field of morality, have been found to be mistaken. Beliefs which were commonly held and even strongly held have later been abandoned. The second idea is that the intuitions in the moral sphere are even more shaky for we can observe many things that were thought right at one time or place are regarded as wrong at another time or place, and vice versa. This diversity of moral opinion should give us pause in thinking that our own commonly held opinions are the touchstone of what *is* right or wrong. Now I have some sympathy with this view: it is always dangerous to base some thesis on our intuitions. Of course, this is commonly done in philosophy, and no more so in moral philosophy than in other areas of philosophical inquiry. And maybe such a practice is unavoidable in the advancement and defence of philosophical theses (in the final analysis, when we are questioning the foundations of our views about ourselves and the world, what else is there to appeal to?). But it is singularly unedifying for a theorist to respond to a criticism of his theory to the effect that it has consequences which are counter-intuitive by

saying that as the intuitions are counter to the theory then this just goes to show that the intuitions are mistaken. This is not an *argument* to the effect that the intuitions are mistaken: we want some *independent* grounds for thinking that they are mistaken. That is, the utilitarian *would* have an argument for the claim that the intuitions are mistaken *if* he had an argument that did not simply presuppose the adequacy of the theory under test, namely, utilitarianism. However, while the utilitarian may not have *shown* that the intuitions are mistaken, nonetheless he may just be right in that claim. This is not an outrageous suggestion -- afterall, our intuitions have been mistaken before. In which case, while we can say that the utilitarian has in this response no argument to the effect that the intuitions that are being used to test the adequacy of his theory are mistaken, equally we must admit that we have not shown that they are correct. In short, if the utilitarian makes this sort of response he forces the present debate as to the adequacy of utilitarianism into a stand-off situation. Any attempt to argue for the adequacy or inadequacy of utilitarianism via an appeal to our moral intuitions is reduced to a non-starter.

However, it might be thought that the situation is not quite as bleak as all that and I develop the

following argument from a paper by J W N Watkins' concerning the rational appraisability of moral theories and principles[35]. Watkins' central idea is the supposition, which he takes to be uncontroversial, that "it would be unreasonable to demand *more* rational appraisability in morals than in science." And yet a commonly held view of contemporary moral philosophers has just this consequence and makes the problem of rational argument over conflicting moral principles particularly acute. This view Watkins calls "justificationism". It is the view that to rationally accept a moral principle or opinion consists in showing that the principle deductively follows from certain higher level (moral) principles (with or without the aid of factual minor premises). Now justificationism in itself does not present a problem for rational argument in morals: the problem arises when this view is conjoined with two other commonly accepted philosophical theses. These are, "non a priorism" and "autonomism". *Non a priorism* is the view that there are no self-evident and necessary moral principles and *autonomism* is the view that there are no "external" factors which require certain moral principles rather than others to be adopted (the autonomist would reject, e.g., an objectivist naturalistic theory of ethics). Thus if one accepts autonomism (as Watkins thinks we should) then moral principles

cannot be appraised by external evidence; and if one accepts non a priorism (as Watkins thinks we should) then moral principles are not self-evident; and if one accepts justificationism then the only way that some principle *could* be justified is by showing that it follows from certain higher level principles, but by autonomism and non a priorism *these* principles too are neither self-evident nor supported by external evidence:- they are just as much in need of justification. Justificationism renders justification in morals illusory.

Watkins now notes[36] that an influential view in the philosophy of science -- the Popperian view -- maintains that scientific theories or hypotheses are *not positively justified:* they are not verified or "confirmed" in any verificationist sense. Rather they may be *falsified* and their "justification" consists in their having successfully weathered attempts to falsify them. In the appraising of scientific hypotheses a crucial role is played by so-called "basic statements". These describe events or situations (like the position of a pointer on a measuring instrument) and are not themselves verifiable by (perceptual) experience. An hypothesis is rejected if it clashes with, i.e., if it implies the negation of, accepted basic statements. However, if a hypothesis is found to clash with a basic statement or statements it is not thereby automatically rejected. For the unverified

basic statemtents are themselves only accepted provisionally and the defender of the hypothesis may attempt to test *them*. These tests may overthrow the basic statements and thus the hypothesis survives. But, as Watkins notes, scientific rationality does *not* allow that accepted basic statements are rejected *merely* because they conflict with a favourite theory.

Now bearing in mind that we should not make greater demands on rational appraisability in morals than in science, we should not expect our moral principles to be *justified* (in the sense persupposed by justificationism). That is, we should not be looking for a justification of our moral principles from *above*, as it were, i.e., by showing that they are deducible from higher level principles; rather we should be looking for a "justification" from *below*, i.e., by showing that they do not have unsatisfactory implications. A scientific hypothesis has unsatisfactory implications when it can be shown that it clashes with accepted basic statements. Clearly, if there is to be an analogy between rational appraisability in science and rational appraisability in morals we will have to show that there is in the moral sphere a class of statements analogous to the basic statements in science. For the moment we proceed on the assumption that there is a class of such statements and we temporarily leave unanalysed the notion

of "unsatisfactory implication" as this applies in morals; we will return to the notion later.

An accepted scientific hypothesis should not be rejected without reason. By analogy an accepted moral principle should not be rejected without reason. The onus is on the critic and reformer to give us reasons for rejecting some commonly accepted principle . The reasons will consist in the critic showing that the principle has unsatisfactory implications.

Let us return now to my claim that if the utilitarian responds to the charge that his theory can be brought into conflict with our moral intuitions by saying, "Then so much the worse for those intuitions", then the debate as to the adequacy or inadequacy of utilitarianism via an appeal to moral intuitions is rendered irresolvable. We might now argue against my claim, using Watkins' account of rational appraisability in morals, in the following way. The principle (intuition) that we ought never to sentence the innocent is a commonly accepted principle. As such it ought not to be rejected without reason. Such a reason would be that it has unsatisfactory implications. But this has not been demonstrated by the utilitarian, and certainly he has not demonstrated this by saying that the principle must have unsatisfactory implications because it is in conflict with his theory. The principle, then, remains "justified" in the sense in which we could

reasonably expect it to be justified. My argument, it will be claimed, presupposed that a principle was not justified unless there were a deductive proof of that principle. But this is to demand too much for the rational appraisability of moral principles: I have presupposed justificationism. Hence we can conclude that a theory, viz., utilitarianism which clashes with the principle that we ought never to sentence the innocent is an inadequate theory.

Now I take this objection seriously and I certainly do not want to take on all the presuppositions that underlie it in the philosophy of science and meta-ethics: that would be to take on too much. But it is not necessary to do this in order to resurrect the essential part of my original claim, namely, that attempts to determine the adequacy or inadequacy of utilitarianism via an appeal to moral intuition are irresolvable. To see this it is necessary to return to the notion of an *unsatisfactory implication* in morals.

As Watkins remarks, "There is sometimes a lingering feeling that there can be rational appraisal only where there is truth or falsity."[37] However, as he notes, the account of rational appraisability in science given by Popper before he was acquainted with Tarski's theory of truth did not presuppose anything of the sort, Popper thought it was possible to avoid using the concepts

"true" and "false" and instead talk of logical considerations of derivability. There is no need to talk of a theory being "false", rather we can say that it is contradicted by a certain set of accepted basic statements. And these basic statements need not be said to be "true" or "false" because "we may interpret their acceptance as the result of a conventional decision"[38]. Now whether we can build up a correct account of rational appraisability *in science* by eschewing talk of truth and falsity is not the main question at issue here: rather the point is that it does not seem that we need truth and falsity for there to be the possibility of *rational appraisal*. That is, for there to be a rational method for us to arrive at a conclusion on some matter (note, not necessarily a conclusion that can be said to be true) does not require that there be truth and falsity. To be sure, what does seem clear is that if there is no possibility of truth and falsity (which is commonly thought to be the case in morals) then for there to be the possibility of rational appraisal we need some fair measure of *inter-subjective agreement*. More particularly, we will need in morals some fair measure of inter-subjective agreement as regards the analogues to the basic statements in science.

Now it is here, as Watkins realises, that the

parallel he has drawn between morals and science is likely to break down. He says:

> It will be objected that I have done nothing towards reinstating *inter-personal* rationality ,.. According to justificationism, differences between different people's top-level principles are unarguable; and it will be said that unarguable differences will re-appear, in our inverted scheme, unchanged except that they are now at the bottom: at the basic statement level the factual/moral parallelism completely breaks down, surely; for whereas people usually reach agreement about easily observed situations, what one person finds morally unsatisfactory another may find indifferent or even good; and there may be no hope of reconciliation.[39]

However, Watkins points out that what is crucial in his account of the rational appraisability of moral principles is that there be agreement as to what counts as *bad* or *wrong* -- *not* about what is good or right. For on Watkins' account a principle is not rejected unless it is shown to have *unsatisfactory* consequences.

> Now I believe it is a plain fact that people's judgements about what is wrong or bad are far more confident, and display *considerably less personal variation*, than their judgements about what is right and good.[40]

Watkins demonstrates this point by way of example: he cites a particular situation where almost all people will surely say that what is being done is wrong. Of course, we are logically free not to appraise the situation in that way, but equally we are logically free not to appraise some factual situation in the way that it is normally appraised. There is no deductive proof available which can prove that a person is mistaken who refuses to appraise as wrong the actions performed in the situation described by Watkins. But similarly there is no deductive proof available which can prove that a person who insists that there is an elephant in my living room at the moment is mistaken. In either case that there is a *large* measure of agreement is a *contingent* fact.[41]

Now the point I wish to make against Watkins is simply this: I think that he is far too sanguine in his belief that his example will generalize to others, and that, coming to our own case there will be a large measure of agreement as to whether the implications of the

principle "Never sentence the innocent" are or are not unsatisfactory. Note that the application of Watkins' idea in this case would be as follows: the principle "Never sentence the innocent" is commonly accepted and as such should not be rejected unless it has unsatisfactory implications, i.e., that its application in some particular case gives rise to a situation whereby there is agreement that the actions performed *ought not* to be performed. The particular case we have in mind is that involving McCloskey's sheriff. Here application of the principle would give rise to harm being done to ten rather than one person. Are our judgements about this particular case going to be in agreement? I want to suggest that this may not be so. Indeed, even a utilitarian may be in two minds; consider the following remarks by Smart:

> Surely, if it is shown that, in certain circumstances ... a utilitarian ought, on his own principles, to commit a serious injustice, such as punishing an innocent man, then it seems that this *does* and *should* weaken the appeal of utilitarianism. And yet one can be made to vacillate back again. We also reflect that the serious injustice would *ex hypothesi* be the only

> possible alternative to an even greater total misery than would be caused by the injustice.[42]

Or again:

> I am not happy to draw the conclusion that McCloskey says the utilitarian must draw. But neither am I happy with the anti-utilitarian conclusion. For if a case really *did* arise in which injustice was the lesser of two evils (in terms of human happiness and misery, that is) the anti-utilitarian conclusion is a very unpalatable one too, namely that in some circumstances one should choose the greater total misery.[43]

This, I think, is a vacillation that most people would be subject to if they seriously considered the implications of the principle "Never sentence the innocent". I assert this as a matter of *fact* and put forward as some evidence for this assertion the quotations from Smart above. I also think that the vacillation is perfectly understandable. For there is a duality in our moral thinking which informs our judgements in particular cases about what is bad or wrong. On the one hand we often take a *teleological*

view and this inclines us to say that if to perform an injustice will lead to less misery then we ought to perform an injustice. On the other hand we often take a *deontological* view and this inclines us to say that it cannot be right to perform an injustice no matter what the consequences. To be sure it is then clear that our particular judgements about what is wrong is informed by our theory. But this in itself does not break down the analogy between science and morals at the basic statement level. For we can point to what is now a common place in the philosophy of science, namely, that basic statements are theory laden. As Watkins says:

> All judgements about badness or wrongness involve *some* ideology, imply that some standard, however commonplace, has been departed from, just as all statements in science are theory impregnated.[44]

But whereas in science we *may* be able to ensure that the theory does not inform our interpretation of the situation currently being used to test that theory, and *may be* we can appeal to basic statements "thinner" in theory content than others, neither of these options seems to be available in the present case. Consider Watkins' idea that moral rationality requires an analogous descending

criticism to that found in science so that as the "critical investigation proceeds, objections will become more earthy and less ideological" and eventually objections to a principle "come down to this, that it would cause so much pain, unhappiness, or death".[45] Now if our disputants were teleologists we could expect some agreement as the criticism of the principle descended, but a deontologist, or even a teleologist in his deontological moments, will not be impressed by the fact that the consequences of failing to act according to a certain principle would cause less pain or unhappiness *if* to so act is to commit an *injustice*. That is, for a deontologist appeal to pain or unhappiness caused by the adoption of the principle will be irrelevant: to perform an act will be wrong if it is to commit an injustice no matter that to choose otherwise is to choose a lesser misery.

So we have reached a similar conclusion as previously albeit by a different route and certainly without presupposing justificationism. An attempt to determine the adequacy or inadequacy of utilitarianism by an appeal to our commonly held intuitions (principles) will not succeed. For the attempt to"justify" *those* principles by making judgements about their application in particular circumstances will require that we make a judgement about the wrongness of particular acts which is informed by our

deontological or teleological views, the very views which we were trying to choose between by making appeal to the original intuitions (principles) currently under test. Now any attempt to judge the wrongness of an action in some particular case *without* that judgement being informed by our general deontological or teleological views (note: *not* the *principle* currently under test) seems doomed to failure, *unless* we think that at some level in the descending criticism either of two things will happen. Either, that at some level we will reach a judgement which, while it presupposes *some* standard, this standard is *common* to the deontologist and teleologist. This *may* happen with respect to some particular principle, but where it does happen is of no interest in the present context. For the problem we are considering concerns those cases where we have a dispute as to the wrongness of an act *precisely because* one disputant focuses on the justness of the act and the other on its bad consequences. The second alternative is this: we think that at some level in the descending criticism we can appeal to *naive* intuition, i.e., that we think that at some level we will just see or in some fashion straightway apprehend the wrongness of an act. But this is as vain a hope as those who think that in science there is a level at the corresponding point where judgements will be based on pure sense data.

I must emphasise that nothing I have said should be taken to imply that I think that rational appraisability in morals is *generally* impossible. Quite the contrary: I think that Watkins has indicated how we can have rational argument about many, maybe even most, of our moral principles and policies; as a matter of *fact* I think it would work in many instances. But in respect of the debate about utilitarianism Watkins' method offers no way out.

Let us turn now to the second response I mentioned: I will concentrate on the argument as it is presented by Hare. Hare notes that the normal ploy of those who attempt to argue that utilitarianism is mistaken as a moral theory is to show that the utilitarian is committed to views that everyone, or nearly everyone, finds counter-intuitive. We have already mentioned McCloskey's argument *ad nauseam*. Hare's answer to this sort of argument involves his major distinction between what he calls "two levels of moral thinking": the "intuitive level" and the "critical level". The *intuitive* level of moral thinking is the product of one's moral education and upbringing. This level of thinking contains "a set of dispositions, motivations, intuitions, prima facie principles (call them what we will)"[46] which we try to inculcate in ourselves and others. But why do we attempt to inculcate just those intuitions or principles rather

than some others? No doubt we do so because we think they are *correct*, that they are the *best* set of such intuitions. But what constitutes the *best* set? According to Hare, "The best set is that whose acceptance yields actions, dispositions, etc. most nearly approximating to those that would be chosen if we were able to use critical thinking all the time."[47] The obvious question now then is, what is *critical* thinking? According to Hare it "consists in making a choice under the constraints imposed by the logical properties of the moral concepts and by the non-moral facts and by nothing else."[48] Hare's analysis of the moral concepts, in particular their alleged universalizability and prescriptivity, has the consequence that to choose at the critical level is to choose as a utilitarian.[49] Now it is not possible for we ordinary human beings to choose on every occasion between alternative courses of action at the critical level, this is only possible for an individual whom Hare calls "the archangel". This is a hypothetical individual with superhuman powers of thought, superhuman knowledge, and who is generally free from all human weaknesses. The archangel has no need for the intuitions or prima facie principles that guide the rest of us.[50] Just because we do not have superhuman powers, etc. *we* cannot judge every action at the critical level. Instead we judge between actions at the intuitive level,

i.e., we employ the intuitions or prima facie principles that we have absorbed in our moral education. Now as we said, in judging between various intuitions -- in attempting to select the best set -- we choose those whose acceptance would yield actions that most nearly approximate those that would be chosen on the basis of critical thinking, i.e., those that would be chosen by an archangel. This means that in choosing amongst intuitions "we have to look at the consequences of inculcating them in ourselves and others; and, in examining these consequences, we have to balance the size of the good and bad effects in cases which we consider against the probability or improbability of such cases occurring in our actual experience."[51] In other words, some intuition will be included in our set of *best* intuitions if it prescribes acts which have *good* consequences in situations we are *likely* to encounter, and *bad* consequences only in situations we are *unlikely* to encounter. In this way the acceptance of the intuition -- its inculcation in ourselves and others -- will yield actions most nearly approximating those that would be chosen at the critical level by an archangel.

We return now to the argument which uses examples such as McCloskey's to demonstrate the inadequacy of utilitarianism. When presented with such an argument Hare suggests that we ask, at what level is the argument taking place? If it is at the *critical* level then the

anti-utilitarian can advance *any* example he wishes, no matter how unlikely it may be or how outside our actual experience. But then our intuitions *cannot* be appealed to. On the other hand, if the argument is at the *intuitive* level then the anti-utilitarian *cannot* advance just *any* example no matter how unlikely. We must remember that our moral intuitions were inculcated in us to enable us to deal with situations that it was *likely* we would come across -- they were not designed to enable us to deal with highly improbable situations.

McCloskey claimed that in order to demonstrate the inadequacy of utilitarianism all that was necessary was that we show that in some *logically possible* situation a utilitarian is committed to the view that what we ought to do is sentence an innocent man. It is easy enough to see why McCloskey makes this move: that all the conditions should obtain that are necessary for it to be the case that that the utilitarian is committed to such a view is extremely unlikely. (Just how likely is it that a sheriff should know or be able to estimate to a sufficiently high degree of probability that if he does not sentence this innocent man then ten people will die; just how likely is it that he can know that the innocent man will not make a greater contribution to social utility than the ten who will die -- maybe the innocent man is an expert on cancer about to bring about a cure

whereas the other ten people are merely drunk hoboes; and so on.) But Hare will respond to such an example by saying[52]: McCloskey's example does not represent a situation that we are at all likely to come across and hence to say that it is counter-intuitive that we ought to do what the utilitarian says we ought to do is to misapply our intuitions -- they cannot be appealed to in such wayout, improbable situations. Of course, Hare must grant that there may be *real life*, if very rare, situations where utilitarianism can be brought into apparent conflict with our intuitions. But it suffices to point out that the anti-utilitarian has not as yet presented such an example and in any case such a situation ought not to be treated at the intuitive level: such situations are, if there are any, very rare, and as such are properly treated at the critical level. In short, arguments that employ examples like McCloskey's in an attempt to demonstrate the inadequacy of utilitarianism are unsuccessful because their appeal to intuition is inappropriate.

However, there is a crucial flaw in Hare's argument. Note that Hare gives the following rationale or justification for our moral intuitions. We mere mortals suffer from weaknesses which make it impossible for us to choose as an archangel. Nonetheless, by employing the above mentioned intuitions we will choose, given the sorts of situations we are *likely* to come across, in a way that

most nearly approximates the choices of an archangel. That is, by adopting these intuitions or principles we will choose in a way that most nearly approximates how we would have chosen had we not suffered from the normal human weaknesses, i.e., how we would have chosen as a utilitarian with super-human knowledge, computational skills, and so forth. Just these intuitions or principles constitute the *best* set of intuitions or principles. Given *this* justification Hare is able to restrict the application of our moral intuitions to situations that we are *likely* to come across and not to situations which are highly improbable or merely logically possible. But while the *anti-utilitarian* will no doubt accept that our moral intuitions are a product of our moral education he will *reject* a *utilitarian justification* of those intuitions. He will agree that we attempt to inculcate certain intuitions in ourselves and others *because* we think they are the *best* set of such intuitions, but he will reject a utilitarian analysis of the term"best" in "best set of intuitions". The intuition that we have that we ought never to sentence the innocent is one that we have absorbed in our moral education, but such an intuition is included in our best set because actions in accordance with that intuition will be actions that properly respect the *rights* of other people and actions

that are *just*. But if the anti-utilitarian rejects a utilitarian justification of our moral intuitions, as surely he will, then there is no need for him to accept Hare's claim that such intuitions are restricted in their application to situations that it is likely we will encounter.

The point can be put in another and possibly clearer way. Hare identifies our intuitions with what we have called "rules of convenience" : this, I take it, is obvious from what I have said previously. Now if the intuition which we have that we ought never to sentence the innocent were merely a rule of convenience, then it would follow that there are certain logically possible situations, like those of McCloskey's sheriff, where we ought to act contrary to that rule/intuition. But this is precisely what McCloskey and other anti-utilitarians would deny, and they hope that this is a view shared by most of us. For if we acted contrary to that rule/intuition -- as the utilitarian says on occasion we ought -- then we would perform an injustice: we would sentence an innocent man. But McCloskey and other people of deontological persuasion take it as obvious that we ought never to perform an injustice, and it is precisely this which they take to underlie the rule/intuition that we ought never to sentence the innocent. Such a rule/intuition is not simply a "rule of con-

venience", it is not subject to utilitarian justification.

Clearly then, Hare's argument is inconclusive against the hard-nosed deontologist but equally clearly the argument of Hare's opponents is not very successful. First, the reason why the argument of Hare's opponents is not very successful: no doubt most of us, or at least most of us in our deontological moments, strongly believe that we ought never to perform an injustice, but this is not a view shared by Hare, and it would not be one shared by us in our teleological moments. If there really *was* a case where sentencing an innocent man was justified on utilitarian grounds then given a teleological view such an act could not be described as the performance of a wrong even though it was an injustice. As Hare says, supposing an opponent to utilitarianism came up with an actual case where on utilitarian grounds murder was justified:

> if he really did find one, we should have to do some critical thinking on it because it would be clearly so unusual as to be beyond the range of our intuition. ...if he did actually have sufficient evidence (a very unlikely contingency), murder would *in that case* be justified ... [54]

But Hare's argument against his opponents is likewise not very successful because it is not a good argument on Hare's part when attempting to rebut the argument of his opponents to presuppose that the intuitions which his opponent take to be obviously *not* susceptible to utilitarian justification *are* so susceptible. Hare's opponents claim that his theory, utilitarianism, must be inadequate as a moral theory because it can be brought into conflict with our intuition that we ought never to sentence the innocent, i.e., Hare's theory would require us on occasions to perform an injustice. Hare's reply is to say that if there really were a case where to sentence an innocent man would maximize utility then such an act could not *really* be a wrong even though it was an injustice, *ex hypothesi* the action would maximize utility. Now clearly Hare's view as to what is right or wrong is informed by his theory. But so too is his opponents view that it can never be the case that we ought to sentence an innocent man, no matter how good the consequences. We have here precisely the problem to which I alluded when discussing Watkins' argument: our intuitions as to what is right or wrong, just or unjust, are not independent of our deontological or teleological theories or views that are currently under test by reference to those intuitions.

What, then, is the proper response to make to the

question posed at the head of this section? The proper response, I think, is to say, "No", provided we interpret the question to be, "Can we, by making appeal to our moral intuitions, come to a conclusion as to the adequacy or inadequacy of utilitarianism?". I have argued in this section that the question as to the adequacy or inadeqacy of utilitarianism via an appeal to our moral intuitions is irresolvable. And hence that what has constituted the main and most common form of debate as to the adequacy of utilitarianism is irresolvable.

Should we conclude from this that the question as to the adequacy of utilitarianism is in general irresolvable? Clearly not: for there may be other ways to demonstrate the adequacy or inadequacy of utilitarianism. Indeed, some such attempts are to be found already in the literature. For example, some philosophers have argued that utilitarianism's insistence on the *impartiality* of moral judgements conflicts with our duties to particular persons and the moral praiseworthiness of affection and loyalty[55]. Others have argued that utilitarianism is, in some sense, self-defeating[56]. I will not examine these approaches here as I do not find any of them particularly convincing. A far more promising approach in my view is to take an approach which is, as it were, in the opposite direction. Instead

of looking at the consequences -- the *implications* -- of utilitarianism we should look at its *foundations*. And here I believe one of the most interesting approaches is to look at the claim expounded in Chapter II that utilitarianism can be *deduced* from the *correct theory of decision*. Suppose that we could show that this claimed *correct* theory of decision was *not* correct. We would have to be careful here for we could not then straightaway conclude that utilitarianism was inadequate: for if we did, we would be guilty of the fallacy of denying the antecedent. But at least we could say that the *grounds* for utilitarianism were *mistaken* grounds. And if we could then go on to show that from what is actually the correct theory of decision we can deduce a theory quite at variance with utilitarianism -- one that required the *denial* of utilitarianism -- we could conclude that utilitarianism is inadequate. I believe that this approach is viable and it is this approach which we shall examine in the remaining chapters.

Before moving on, however, what can we say in summary that we have learnt from the present chapter? We noted that the common form of attack on utilitarianism consisted in showing that it can be brought into conflict with our moral intuitions. We then argued that a common form of defence to this attack, namely, the defence that consists in drawing a distinction between RU and AU does not

suffice, for a significant distinction between RU and AU cannot be maintained. But then we considered whether the fact that utilitarianism can be brought into conflict with our moral intuitions really mattered. We concluded that it did not, not because the utilitarian has an adequate defence, but because the question as to the adequacy or inadequacy of utilitarianism via an appeal to our moral intuitions was irresolvable. And finally, we suggested that this provided a motivation for ignoring the *common* approach which is to look at the implications of utilitarianism and that we should instead take the approach which is to look at utilitarianism's foundations.

Footnotes:

1. W D Ross (1930), Chapter II especially pp.17-18.
2. H J McCloskey (1957), especially pp. 468-469; (1963), p.599; (1965) especially pp. 252-255. Of course, Ross and McCloskey were not the first to criticise utilitarianism, but their objections are extremely well-known and have been extremely influential.
3. McCloskey (1963), p. 599.
4. Smart (1973), p. 30.
5. John Austin (1954), p. 47.
6. Austin (1954), p. 38.
7. I say "acted in accordance with" rather than "followed" for it seems right that a RUian is not concerned that individuals should actually *follow* a rule, it is enough that they should merely act in *accordance* with it. This supposition fits the consequentialist spirit of his theory.
8. Bernard Mayo (1957), p. 167-172.
9. We have to be careful here. First, the consequences deducible from a given act may make up a potentially infinite and undecidable set, and as such it would be logically impossible to work these out. The point is somewhat trivial however for the claim can be rephrased by understanding the term

"consequences" to denote finite, decidable, although very large sets. But even then the computation of *these* consequences may outstrip human capacities. Second, what consequences it is possible for one individual to compute may not be possible for another individual to compute: the possibility here will be relative to a person's computational skills, knowledge, etc.. Hence, when we later say that a utilitarian can only require that an individual consider the consequences it is possible for him to predict, just what those consequences are may well vary from individual to individual. Even so an individual does not have an automatic defence, when he performs what turns out to be the wrong act, that he did not have the requisite computational skills or knowledge. For we may want to say that he *ought* to have had those skills or knowledge. This is a difficult problem and we here simply presuppose that it is subject to utilitarian analysis.

10. Smart (1973), pp. 42-43, and T L S Spriggs (1965), pp. 287-288.

11. J L Mackie (1977), pp. 137-138.

12. T L S Spriggs (1965), pp. 288-289.

13. Harsanyi (1977), especially pp. 30-41; Harsanyi (1979), especially pp. 311-314.

14. See R D Luce & H Raiffa (1957), pp. 170-177.

15. Harsanyi (1977), p. 37.

16. Harsanyi (1977), p. 35.

17. Harsanyi (1977), p. 38.

18. Harsanyi (1977), p. 38.

19. Harsanyi (1977), p. 37.

There is a point that I should note here. For ease of exposition I have spoken of the "decision maker" whereas Harsanyi makes his point about the expectation and incentive effects in terms of what he calls "primary agents". A particular moral agent (i.e., utilitarian) *A* in a given class *C* of possible decision situations is known as the *decision maker*. The agent *A* as well as all other moral agents who are or will be in a situation belonging to class *C* are known as *primary agents*. All moral agents whose interests are directly or indirectly effected by what the primary agents do in situations of class *C* are known as the *secondary agents*. With this terminology we can now say that any utilitarian decision maker *under RU* must choose his strategy on the assumptions that *all* primary agents will employ the same strategy. Of course, the classes of primary and secondary agents are not necessarily disjoint. However, there is a heuristic advantage in Harsanyi's talk of *primary agents*. For now it is

very clear that any decision maker must not only consider what would be the effects on the expectations and incentives of other people in society supposing that they knew that *he* had adopted a certain strategy (which could give rise to effects which were quite trivial) but rather he must consider what would be the effects supposing that they knew that all decision makers in that class of situation (i.e., all primary agents) adopted that strategy (which would have far from trivial effects). For example, to revert to McCloskey's sheriff: under RU he must not only consider what would be the effects on the other individuals in society supposing that they knew that *he* had adopted a strategy which permitted the sentencing of the innocent, but he must also consider what would be the effects given that they knew that all other sherrifs (or law officers) had adopted that strategy -- supposing that there is more than one sheriff in society (in the small town). Clearly, in McCloskey's example we are supposing that the small town is an isolated community with only one sheriff. Hence for the purposes of that example the distinction between the decision maker and primary agents is one of no consequence; we suppose that there is and will be only one individual who can make a decision in the sort of decision situation that the sheriff

is in. The class of primary agents is, in this example, assumed to be unitary. Even so it would not follow in this example that the expectation and incentive effects would be trivia-; if so-and-so is the only sheriff -- the only individual to dispense justice -- then to know that he has adopted a strategy which permits the sentencing of the innocent would have effects on one's expectations and incentives that were far from trivial.

20. Harsanyi (1977), p. 38.

21. Harsanyi (1977), p. 38.

22. It may be thought that we can in fact present a stronger argument here by distinguishing between the situations that involve sentencing the innocent in the following way. We *drop* the assumption that the situations where to sentence the innocent maximizes utility are very rare, and instead talk of the first n such situations (where n is some small number) and the rest. We then say that a RUian ought to adopt a strategy which required the sentencing of the innocent in the first type of situation, and to adopt a strategy which forbid the sentencing of the innocent in the second type of situation. This would not have distastrous expectation and incentive effects. But the trouble with this argument is that it does depend on a rather unnatural and I'm not sure totally plausible criterion for distinguishing types of situations. For consider: we have to

allow that for any decision maker, he will confront, say, 30 situations that involve sentencing the innocent and which are identical in all respects except temporal precedence, and yet we are going to say that one (type of) action is to be assigned to the first, say, 3 such situations and not to the rest simply on the grounds that the first 3 are distinguishable from the rest because they *are* the first 3 such situations. This problem is completely avoided in the next argument I present against Harsanyi in the following paragraph of the text, and as such I prefer to advance that argument rather than the one adumbrated above. In that argument we *also* drop the assumption that the situations where to sentence the innocent maximizes utility are very rare, and we allow that if situations are identical in all respects (except temporal precedence) then the same action must be assigned to each, but we simply point out that a genuinely utilitarian theory ought to allow *some*, but not *all*, decision makers to employ a certain strategy if that maximizes utility. All this argument requires is that we be able to distinguish between decision makers and there is surely no problem in that.

23. Gertrude Ezorsky (1968), especially pp. 538-540.

24. Ezorsky (1968), p. 538.
25. Ezorsky (1968), p. 539.
26. Ezorsky (1968), p. 539.
27. Ezorsky (1968), p. 540 footnote.
28. Ezorsky (1968), p. 540.
29. Aristotle *Nichomachean Ethics* (VII) (J A K Thomson (trans.)).
30. Bernard Mayo (1957), especially pp. 162-165.
31. Mackie (1977), p. 138-139.
32. Or, more accurately, attributed to Smart by, for example, McCloskey (1963), p. 599.
33. Sprigge (1965), p. 276-280.
34. R M Hare (1981), especially Chapter 8.
35. J W N Watkins (1963), especially pp. 97-113.
36. Watkins (1963), pp. 103-104.
37. Watkins (1963), p. 101.
38. Karl R Popper (1968), pp. 273-274.
39. Watkins (1963), pp. 109-110.
40. Watkins (1963), p. 111.
41. The point made in this paragraph has, of course, already been made with respect to basic statements. For a fuller statement of the argument I sketch here see J J Kupperman (1970), pp. 57-60.
42. Smart (1965), p. 347.
43. Smart (1965), p. 348.
44. Watkins (1963), p. 112.

45. Watkins (1963), p. 113.

46. Hare (1981), pp. 46-47.

47. Hare (1981), p. 50.

48. Hare (1981), p. 40.

49 An indication of how this is so is given by Hare in his (1981), pp. 42-43.

50. Hare (1981), pp. 44-45.

51. Hare (1981), p. 48.

52. Hare (1981), pp. 134-135.

53. For a statement of Hare's position on the AU/RU debate see his (1981), p. 43.

54. Hare (1981), p. 135.

55. For example, see Rescher (1975), especially pp. 70-72, and the opening paragraph of section 3, Chapter II.

56. For a discussion and rejection of this idea see Peter Singer (1972), pp. 94-104.

CHAPTER IV

THE INADEQUACY OF ORTHODOX DECISION THEORY

1. An Outline of the General Approach and the Underlying Assumptions taken to demonstrate the Inadequacy of Orthodox Theory.
2. Falsifying Orthodox Theory: The Tversky-Kahneman Experiment.
3. Orthodox Theory from the Normative Standpoint.
4. "Subjective", "Ethical" and "Overall" Preferences.
5. The First Argument for the Empirical Vacuity of Decision Theory: Decision Theory is "Soft-edged".
6. The Second Argument: The Holism of Reasons, Actions and Rationality.

1. *An Outline of the General Approach and the underlying Assumptions taken to demonstrate the Inadequacy of Orthodox Theory*

As we saw in Chapter II it is possible, given orthodox decision theory and a minimal characterization of what it is to make a moral judgement, to deduce utilitarianism as a moral theory. The purpose of this

chapter will be to show that orthodox decision theory is inadequate as a theory of rational decision. But the question that arises is, "In what *sense* is orthodox theory an inadequate theory?". The response we give to this question very much depends on what meta-level status we ascribe to orthodox theory. I mention three possibilities here.[1] We could regard orthodox theory as a purely *formal* theory, i.e,, as a theory which simply offers a stipulative definition of "rationality" in its axioms (i.e., the rationality postulates (1) to (4) mentioned in Chapter II section 2) and what is deduced from those axions (in particular, the expected utility theorem) are simply the logical implications of those axions.[2] The only question of adequacy that can arise with respect to such a theory concerns whether the deductions have been properly carried out. I don't doubt the logical abilities of orthodox theorists and I grant that orthodox theory is perfectly adequate from the formal point of view. But to take this stance towards orthodox theory renders it inadequate for the deduction of a moral theory. For if the theory does not tell us how we *ought* as rational individuals to act, or how we *do* as rational individuals act, then the theory is powerless to tell us how as rational individuals we ought or would act in situations like that envisaged under the equiprobability model. These remarks suggest

that the appropriate meta-level stances for our purposes are the *normative* and *empirical* stances. As a normative theory, orthodox theory would be regarded as a system which yields *prescriptions* for rational choice, i.e., it would tell us how, as rational individuals, we ought to act even if in fact we do not act that way: in particular, it would tell us that we ought to choose or act so as to maximize expected utility. As an empirical theory orthodox theory would be regarded as a system which yields empirical hypotheses about actual human behaviour: in particular, it would tell us that individuals actually do choose or act so as to maximize expected utility. It will be my intention in this paper to show that it is at least *problematic* that orthodox theory is adequate from the normative standpoint, although I believe we cannot come to any definite conclusions as to the adequacy of orthodox theory conceived of as a normative theory, and I will also show that orthodox theory is definitely inadequate as an empirical theory. It is the question of the adequacy of orthodox theory conceived of as an empirical theory that will be of prime concern in this chapter for it is this question, I believe, which has a *definite* answer.

To show that orthodox theory is inadequate as an empirical theory we have to do two things. We not only have to indicate what empirical data apparently shows

orthodox theory is inadequate, we also have to show that orthodox theory or decision theory generally is capable of being shown to be inadequate by such data, i.e., that it has *empirical content*. As we shall see it has been *denied* that decision theory has empirical content. So what we need here is some characterization of when a theory does or does not have empirical content. I will *assume* that a theory has empirical content if and only if there are specifiable empirical conditions such that if they obtain then this falsifies the central hypotheses of the theory. As we shall see, however, we do not require of *every* hypothesis that should these conditions apparently obtain then the hypothesis must be abandoned. For we allow that there may be very good reasons - even reasons that are consistent with a falsificationist point of view - for clinging to a hypothesis in the face of apparently conflicting evidence, and in such cases we simply revise our estimate of that evidence.[3] The assumption that a theory has empirical content if and only if it is falsifiable is, I realize, a very large assumption but it would take us too far afield to establish it in this thesis. In any case, as will become apparent, it appears to be an assumption shared by our opponents.

Our argument will also require two theses which are of crucial importance and which will be argued for more

fully in the following interpolation. Until then I ask the reader to merely assume their truth. Each thesis requires the rebuttal of a distinct version of a venerable argument that has become generally known as the "Logical Connection Argument". This argument, usually addressed against the causal theorist of action, asserts that the relation between a *reason* and its *action* is of a "logical" or "conceptual" nature. The first of the two theses required is that we may have *independent access* to a person's reason for acting. A reason we will understand in the usual way to consist of a belief and a preference (or a desire or a want). To say that we have independent access to a person's reasons is to say that we may positively identify a person's beliefs and preferences without inferring them from the performance of the action for which those beliefs and preferences are the reason. This thesis of independent access should not be confused with another quite distinct thesis. Our thesis does not deny, what is no doubt true, that by and large we identify a person's beliefs and preferences via their actions (where we understand verbal behaviour to be in the category of action). The thesis of independent access merely denies that our *only* way of knowing a person's beliefs and preferences is to infer them from the performance of the action *which those beliefs and preferences predict or explain* (i.e., for which they are

the reason). Now as we will be claiming that we may genuinely explain or predict human action on the basis of a person's beliefs and preferences we clearly require the thesis of independent access, otherwise such explanations or predictions are, as we shall later say (section 5), *epistemically circular*.

The second thesis required is that the relation between a person's reason and his action is in some sense *contingent*. More specifically we require that where *R* is the reason for the action *A*, then the conditional "If *R* occurs then *A* occurs" may be false, i.e., the antecedent be true and the consequent false. We require that such a conditional may be false even when there is no *countervailing* reason *R'* that occurs: *R'* is a countervailing reason to *R* just when *R'* is the reason for *A'* and in performing *A'* the agent cannot perform *A*. That is, we require that the conditional "If *R* occurs then *A* occurs" may be false and not only when it is true that the individual had the reason *R* but fails to perform the action *A* because he changed his mind. We will also require that such a conditional may be false even when *R* occurs under *optimal conditions*. We may not be able to fully specify these conditions but we have in mind that the individual does not suffer a sudden attack of paralysis, a sudden attack of amnesia, does not suffer a sudden death, and so on. This second thesis I will refer

to, for short, as the thesis of *contingent reason-action conditionals*. This thesis should not be confused with another quite distinct thesis - its *converse* - which we can call the thesis of contingent *action - reason* conditionals. The thesis of contingent action - reason conditionals would assert that the conditional "If *A* occurs then *R* occurs" is contingent. We will allow that such a thesis may be false because the sentence "*A* occurs" entails the sentence "*R* occurs". But, as we shall see, the truth of this thesis is not required by our argument, and even if it is false that does not show that the thesis required by our argument - the thesis of contingent *reason - action* conditionals - is false. To be sure, if the thesis of contingent *action - reason* conditionals is false then that may be sufficient to show that reasons cannot be the *causes* of actions. But our argument that decision theory has empirical content does not require that reasons be the causes of actions. For while we will say that some appropriately modified and universalized conditional like "If *R* occurs then *A* occurs" may function as a law-like statement in the explanation and prediction of human action, as Carl Hampel has argued there are true laws employed in scientific explanations and predictions which are *not* causal in character[4]. Our only concern in attempting to determine whether decision

theory has empirical content is whether such a conditional is falsifiable, not with whether it expresses a causal connection.

Moreover, even if the thesis of contingent *action - reason* conditionals is false, i.e., that the conditional "If *A* occurs then *R* occurs" *cannot* be false, this does not require that we abandon the thesis of independent access. The thought that this may be otherwise is prompted by the following idea. Suppose we establish by independent means that for some person *R* occurs. And also suppose that we observe his behaviour and assert that *A'* occurs. If we grant that "*A'* occurs" entails "*R'* occurs", and *R'* is a countervailing reason to *R* , then we must revise our original reason ascription based on independent means: we must say that his reason was not really *R* but *R'*. However, note that for it to be the case that we must revise our original reason ascription it must be the case that we describe or count his behaviour as the performing of the action *A'*. We need not so describe his behaviour: we could say that he failed to perform any action. And this is quite consistent with his having the reason *R*, *provided* the thesis of contingent *reason - action* conditionals is true, for then "*R* occurs" may be true and "*A* occurs" be false (and "*A* occurs" will, of course, be false if the individual has failed to perform any action). These points will be

referred to again in section 5 and an amplification of them will be given then.

Finally, we will distinguish between three types of rationality of three senses of "rational". We will speak of formal practical rationality, epistemic rationality and ends rationality.[5] An individual exhibits *ends rationality* when his preferences are consistent, i.e., when they are complete and transitive (see Chapter II section 2). A person exhibits *epistemic rationality* when his *beliefs* are consistent and correct by our standards (see section 6 of this chapter). A person exhibits *formal practical rationality* when he *acts* in a maximizing fashion *relative* to his beliefs and preferences. Notice that for a person to exhibit formal practical rationality it is not necessary that he have any particular beliefs or any particular values. Although, of course, as we saw in Chapter II section 2, if an individual is to choose in a maximizing fashion relative to his beliefs and preferences then his preferences must at least be consistent. We will call the claim that all individuals exhibit formal practical rationality the *Rationality Principle*: our terminology here follows that of Popper[6] and Watkins[7]. Our primary concern in this chapter will be to argue that the Rationality Principle has empirical content, for a theory of decision is simply an attempt to give a more precise

specification of that principle. Let us now look at the adequacy of the specification of that principle given by orthodox decision theory: in the first instance we consider this question from the empirical standpoint.

2. *Falsifying Orthodox Theory: The Tversky-Kahneman Experiment*.

As we have already noted (Chapter II section 2) if we assume that a person's preferences obey certain conditions, i.e., that they are *consistent* and that they obey the *Probablistic Equivalence Postulate* and the *Sure-thing Principle*, then we may deduce the *expected utility theorem*. This theorem says that the utility of a lottery is equal to the sum of the products of the utility of each of the (mutually exclusive and exhaustive) outcomes times its probability. Thus for some lottery L consisting of the outcomes O_r and O_s at probabilities p_r and $(1 - p_r)$ respectively, we have:

$$U_i(L) = p_r . U_i(O_r) + (1 - p_r) . U_i(O_s)$$

This says something very important about a person's *preferences*: it says that the *value* any person ascribes to a lottery will simply be equal to the lottery's mathematically expected utility. But, as was remarked in

Chapter II, orthodox decision theory is not simply a theory that characterizes a person's preferences; it also says something about how people *act*. As Donald Davidson says, "The second part of the theory relates action to preferences"[8]; it says that a person with a set of preferences so characterized always chooses that action (alternative) - from among those available to him at the moment - such that no other has a higher expected utility. That is, orthodox theory claims that an individual *acts* so as to maximize expected utility. Thus orthodox decision theory can be seen as providing a more specific formulation of the Rationality Principle. An individual is supposed to have a certain set of beliefs - he believes that certain alternatives are available to him at the moment with outcomes at certain probabilities; he has a certain set of preferences - in particular, they are consistent and obey the Probablistic Equivalence Postulate and the Sure-thing Principle; and it is claimed that the individual will choose in a maximizing fashion relative to those beliefs and preferences, i.e., he will choose the alternative with the highest expected utility.

But a problem for orthodox decision theory arises from the fact that it does not seem to be true that individuals act so as to maximize expected utility. Consider the following favourite example of Harsanyi's. Suppose that an individual is willing to pay $5 for a

lottery ticket that will give him a 1/1000 chance of winning $1,000. Harsanyi then asks, "How will the theory of vNM utility functions explain the fact that he is willing to gamble at such highly unfavourable odds?"[9] Harsanyi says, "The explanation will be obviously in terms of the *relative importance* he assigns to the possibility of winning $1,000 as against the relative importance he assigns to the possibility of losing the $5 he will invest." That is, we explain the apparent fact that this individual is not maximizing expected utility by supposing that this particular individual's marginal utility for money sharply increases at or about $1,000. This explanation may carry some weight with respect to certain cases, for example, when, as Harsanyi points out, there are important complementarities among the commodities the individual could buy at or about $1,000. I surmise, however, that there will not always be such complementaries and as a consequence an appeal to marginal utility will, in these cases, have little persuasive force. It would appear more plausible to say that the individual likes gambling.

Be that as it may we can put the matter to one side for the power and importance of the experiment to which I shall now refer resides in the fact that it demonstrates that an appeal to decreasing or increasing marginal utility will not suffice for a defence of orthodox theory.

In the Tversky-Kahneman[10] experiment subjects were asked to choose between A and B, and C and D in the following situation;

Choice I : A = (1,000, 1/2, 0) B = (400)

Choice II : B = (1,000, 1/10, 0) D = (400, 1/5, 0)

Without going into the details of the experimental design we can simply note that a variety of subjects were asked in Choice I to choose between receiving $1,000 or $0 with probability ½ or $400 with a probability of 1. A similar interpretation applies to Choice II. But apparently nearly all subjects chose *B* over *A*, and *C* over *D*, and analogous results were obtained using different payoffs and probabilities. These results seem to falsify orthodox theory. How so?

Initially suppose that utility is a linear function of monetary payoff and let the utility of $0 = 0, i.e., u(0) = 0. Then the expected utility of *A* is:

$$u(A) = \tfrac{1}{2} \times u(1,000) + \tfrac{1}{2} \times u\ (0) = 500$$

and

$$u(B) = 1 \times u(400) = 400$$

On the assumption then that the utility of money is a linear function of monetary payoff when the individuals choose *B* over *A* they do not choose the alternative that

maximizes expected utility. The obvious response on the part of the orthodox theorist is to point out that the utility of money is *not* a linear function of monetary payoff, rather it is *marginally decreasing*. (Notice that it is this sort of approach that is taken by Harsanyi in the case discussed above). But then if in choosing *B* rather than *A* the individuals are to choose the alternative that maximizes expected utility it will have to be the case that u(1,000)<800 for then the expected utility of *A* is *less than* 400 while that of *B* is *equal to* 400. However we then have:

$$u(C) = 1/10 \times u(1{,}000) + 9/10 \times u(0) < 80$$
$$u(D) = 1/5 \times u(400) + 4/5 \times u(0) = 80$$

That is, the expected utility of *C* is *less than* the expected utility of *D* , and yet subjects choose *C* over *D* . In other words, if we say that individuals in choosing *B* over *A* choose the alternative that maximizes expected utility because utility for money is marginally decreasing, then this implies that when individuals choose *C* over *D* they do *not* choose the alternative that maximizes utility. That is, it does not seem that the orthodox theorist can claim that in *both* the choice of *B* over *A* and in the choice of *C* over *D* individuals choose so as to maximize expected utility. Hence if the

orthodox theorist puts his theory forward as an empirical theory it seems to have been falsified: it is not the case that individuals act so as to maximize expected utility.

I now want to consider three possible responses that the orthodox theorist who still wishes to present his theory as an empirical theory could make to the above experiment and argument. None of these responses will, I believe, suffice: the first because it is irrelevant, and the second and third because they involve the introduction of seriously *ad hoc* auxillary hypotheses. The first response involves a device commonly favoured by supporters of the expected utility approach that was introduced by Milton Friedman and L J Savage: this is the idea that marginal utility curves have *inflexion points*[11], i.e., that a marginal curve may decrease and increase (i.e., be *S*-shaped). But this device is not sufficient here, for the Tversky-Kahneman experiment shows that if we stick to the claim that individuals are maximizing expected utility in the choice of *A* over *B* and *C* over *D* then we must admit (using the numerical values previously mentioned) that \$1,000 has a value *less than* 800 (in Choice I) and *greater than* 800 (in Choice II). But this is not even possible for utility curves with inflexion points. However, the mention of \$1,000 having two different values in Choice I and Choice II suggests a

second response, and this is to introduce the auxillary hypothesis that the subjects' preferences changed over time from Choice I to Choice II. Well this is certainly *possible* but this hypothesis does seem to be purely *ad hoc*: it is merely introduced, without further evidence to save orthodox theory from strong counter-evidence. The third response involves the subjects' beliefs: we introduce the hypothesis that we incorrectly assessed the subjects beliefs concerning which alternatives were available and at what probabilities. But once again, for similar reasons, this too seems an *ad hoc* hypothesis.

Nonetheless, I do not want to suggest that it is *always* unjustifiable to introduce apparently *ad hoc* hypotheses to protect the central hypotheses of a theory from apparently falsifying instances to those central hypotheses. I suggested as much in the opening section of this chapter. The idea here rests on the distinction between the *content* of a theory or hypothesis and how we *treat* that theory or hypothesis: a theory or hypothesis may have empirical content, i.e., it may be falsifiable, but we may treat that hypothesis, for good reasons, even reasons consistent with a general falsificationist view point, as unfalsifiable. Here I follow Watkins in claiming that we are sometimes justified, even on falsificationist grounds, in clinging to the central hypotheses of our theory in the face of conflicting

evidence[12]. Watkins calls such hypotheses *principles*: a principle is a privileged component of a theory that is treated as unfalsifiable in the interest of the falsifiability of the whole system. As examples of principles in science we could cite the conservation laws of mass-energy and that space-time is continuous. Notice that as a matter of psychological fact such principles may be such that we come to say that we cannot imagine them to be false, we find it inconceivable that there should be a falsifying instance to them: for example, it is surely an empirical (i.e., falsifiable) hypothesis that space-time is continuous, and thus that it is possible that an object is in one place up until time t, and then is just in another place after time t, and yet if we came across such a case we would say that this just cannot be - or measuring instruments or something else must have been in error. And to treat such principles in this way is quite consistent with a general falsificationist approach to science: for to give up such a principle would play havoc with our ability to explain and, in particular, predict empirical phenomena that could falsify our system - our physical theories - in any way at all. In effect, to give up such a principle, unless we can find another to replace it - and with these principles it just is very difficult to imagine what *could* replace them - would be to give up doing science.

I want to suggest that some formulation of the Rationality Principle is just such a principle; it is *falsifiable* but it is *treated* as unfalsifiable for without it our ability to explain and predict human action would be seriously diminished. Nonetheless, such principles are not sacrosanct; they may be replaced if something better comes along. And in my view we have just such a situation with respect to orthodox decision theory; its formulation of the Rationality Principle cannot accommodate the results of the Tversky-Kahneman experiment and its formulation is replaceable by another which *can* accommodate the results.

If we want to say, as I think we should, that the Rationality Principle in some formulation is not to be abandoned, i.e., we wish to keep the general idea that all individuals act in a maximizing way relative to their beliefs and preferences, and as it seems clear from the Tversky-Kahneman experiment that individuals do *not* act so as to maximize *expected* utility, then it must be the case that the *utility* they ascribe to an alternative is *not* equal to its *expected* utility. But as we saw we may deduce the expected utility theorem, i.e., the theorem that says that the utility of an alternative, in particular, a lottery, is equal to its expected utility, if an individuals' preferences are consistent and obey the Probablistic Equivalence Postulate and the Sure-thing

Principle. In which case, by *modus tollens*, it must be that an individual's preferences do not obey all these conditions. Now experimental tests to determine whether people's preferences are actually consistent have been, to say the least, inconclusive.[13] And more importantly, as we shall see in section 6, there is a strong argument to the effect that if we are to be able to ascribe preferences to an individual at all we must *presuppose* that his preferences are consistent. Hence, any problem with our characterization of the relation of preference must reside with The Probablistic Equivalence Postulate and/or The Sure-thing Principle. Of these two conditions it is interesting to note that Harsanyi says they:

> presuppose that the *decision maker has no specific utility or disutility for gambling as such* ... (they) assume that the decision maker will take a purely *result-orientated* attitude towards lotteries, and will derive all his utility or disutility from the prizes he may or may not win through those lotteries, rather than from the act of gambling itself.[14]

We can see this when we consider once again what is required by these two principles. Recall that the Probablistic Equivalence Postulate required, in particular,

that an individual be indifferent between a one stage and a two stage lottery if they were probabilistically equivalent. But this condition will not hold for an individual who likes gambling: such an individual will not be indifferent between a one stage lottery and a two stage lottery even if they are probabilistically equivalent just because the latter presents him with the opportunity of gambling twice whereas with the former he can only gamble once. A parallel argument can be advanced for an individual who does not like gambling. The Sure-thing Principle also presupposes that individuals do not like or dislike gambling, or rather, more accurately, it presupposes that they will take a purely *result-orientated* attitude towards lotteries. In particular, it presupposes that individuals are not *risk averse*, i.e., that they value certainty as such. For, as we noted, the Sure-thing Principle requires that if an individual prefers a non-risky alternative to a risky alternative, then he will not prefer a lottery involving the latter to a lottery involving the former if the lotteries are otherwise equivalent. But this will not hold for an individual who values certainty as such just because it may have been the non-risky alternative's advantage of *certainty* that made it attractive relative to the risky alternative, and this advantage disappears when *both* alternatives are imbedded in a lottery.

It is easy enough to demonstrate that there is a failure of at least one of the above conditions for the individuals in the Tversky-Kahneman experiment. (Tversky offers a similar argument but in terms of the Reduction Assumption and the Substitution Assumption,which, as we noted, also required what the Probablistic Equivalence Postulate and the Sure-thing Principle required respectively). According to the Probabilistic Equivalence Postulate the compound lottery (*A*, 1/5,0) is indifferent to the simple lottery *C*. Similarly, according to that condition (*B*, 1/5,0) is indifferent to *D*. But now, by the Sure-thing Principle, if *B* is preferred to *A*, as seems to be the case for the individuals in the Tversky-Kahneman experiment, then any probability mixture of *A* with 0 will not be preferred to the same probability mixture of *B* with 0. In particular, (*A*, 1/5,0) will not be preferred to (*B*, 1/5,0). But as the subjects choose *C* over *D* it *does* seem that the individuals prefer (*A*, 1/5, 0) to (*B*, 1/5,0).

What is going on here? Tversky suggest that we are witnessing what he calls "the certainty effect". That is, for the subjects in the Tversky-Kahneman experiment.

> the utility of a positive outcome appears greater when it is certain than when it is embedded in a gamble,[15]

Let us entertain this hypothesis that individuals value certainty as such, i.e., that they are risk averse. Given this hypothesis we explain the choice of *B* over *A* not by saying that u(400) > ½ x u(1,000), but by saying that *B* enjoys a certainty advantage over *A*. In other words, we drop the claim that u (1,000) < 300 - although we grant that the utility of money is marginally decreasing such that u(1,000) < 1,000 but reject the idea that the rate of decrease is rapid enough to ensure that u(1,000) < 800 - and instead simply say that *B* is chosen over *A* because *B* has a certainty advantage over *A*. This implies that in choosing *C* over *D* the individuals choose that alternative which maximizes expected utility, but that, of course, in choosing *B* over *A* they do not.

Pretty clearly, however, orthodox decision theory cannot entertain this hypothesis, for then the orthodox theorist must allow that individuals do not always act so as to maximize expected utility. Note that we have not, however, had to abandon the idea that individuals act in a maximizing fashion relative to their beliefs and preferences. We still subscribe to the view that individuals act in a way that maximizes utility relative to their preferences (and beliefs), but we now suppose that they have a preference for certainty as such, i.e., that they place a higher utility on a (positive) outcome when it is certain than when it is embedded in gamble. But orthodox

theory cannot suppose this for, as we already argued, the Sure-thing Principle rules out that a (positive) outcome should have a higher utility when it is certain than when it is embedded in a gamble: The Sure-thing Principle precludes an individual from valuing certainty as such. Hence, while we have not had to abandon the Rationality Principle we have seen that the orthodox theorist's formulation of it is empirically inadequate.[16]

Davidson also considers the Tversky-Kahneman experiment but he comes to a conclusion seemingly quite at variance with the one reached above: Davidson thinks that the Tversky-Kahneman experiment does *not* falsify decision theory, but rather highlights the *unfalsifiability* of decision theory. Davidson argues[17] that what is crucial is what *description* we give to the alternatives. We can describe the alternatives in such a way that decision theory is not necessarily falsified by the results of the Tversky-Kahneman experiment. He suggests that:

> ... the reason subjects shy away from *A* is that the zero outcome should be described as missing out on the prospect of getting $400 for certain. Put differently, the lack of risk in *B* had a value of its own. Given *this* assumption, decision theory is not necessarily falsified by the results.[18]

I agree with Davidson that we can describe the alternatives in such a way that given the assumption that people disvalue risk the results of the Tversky-Kahneman experiment do not falsify decision theory. But such a theory would not be *orthodox* theory, i.e., a theory which maintains that individuals maximize expected utility and that their preferences obey the Sure-thing Principle. For orthodox decision theory such descriptions are irrelevant because that theory specifically *excludes* the possibility that individuals value lack of risk: it is just not open for the orthodox theorist to assume that "the lack of risk on B had a value of its own."

So while the Tversky-Kahneman experiment may not have falsified *some* theory of decision, *viz.*, some *unorthodox* theory which assumes that individuals value certainty as such, it would be mistaken to conclude from this that decision theory and, in particular, orthodox theory is unfalsifiable. To do so would be to confuse two incompatible theories both of which might be properly called a theory of decision making. On the one hand we have orthodox theory - this theory we've argued is false. On the other hand we have a theory which claims that individuals act so as to maximize utility and that they ascribe utility to lack of risk - such a theory has not been shown to be false. It would thus be a mistake to confuse these two theories and if we do confuse them then

it would not be surprising that we should erroneously conclude that decision theory is unfalsifiable.

The point can be put in another way. The Rationality Principle has not been falsified by the Tversky-Kahneman experiment, i.e., we have not shown that it is false that individuals act in a maximizing way relative to their beliefs and preferences, i.e., that they act so as to maximize utility, if we allow that individuals value certainty as such. Given that individuals value certainty as such then the individuals in the Tversky-Kahneman experiment can still be said to maximize utility when they choose the alternatives they do choose. However, the formulation of the Rationality Principle given by Orthodox theory *has* been falsified. But we should obviously not conlcude from *this* that the Rationality Principle is unfalsifiable, i.e., that it lacks empirical content. Of course, there may be other reasons for thinking that the Rationality Principle lacks empirical content and we shall consider these after we have considered whether orthodox theory can be regarded as adequate from the *normative* standpoint.

3. *Orthodox Theory from the Normative Standpoint*.

Many decision theorists would not be greatly concerned that orthodox decision theory has been shown to be false, for they regard orthodox theory as a

normative (or prescriptive) theory. They would grant that orthodox theory may well be empirically inadequate but this does not show that it is normatively inadequate. Hence Howard Raifla has said in response to an experiment similar to the Tversky-Kahneman experiment;

> But no one claims that most people *do* behave as they *ought* to behave. Indeed, the primary reason for the adoption of a prescriptive or normative theory (that is, an "ought to do" theory) for choice behaviour is the observation that when decision making is left solely to unguided judgement, choices are often made in an internally inconsistent fashion, *and this indicates that perhaps the decision maker could do better than he is doing*. If people always behaved as this prescriptive theory says they ought to, then there would be no reason to make a fuss about a prescriptive theory. We could then just tell people, "Do what comes naturally".[19]

Now I think that we can establish a *prima facie* case against orthodox decision theory conceived of as a normative theory. That is to say, we can give some reason for thinking that it is inadequate and show that some ways

which attempt to argue for its adequacy are unsuccessful. However, in so doing we have not definitely demonstrated that orthodox theory is inadequate as a normative theory. We have only shown that we have some reason to doubt the orthodox theorist's claim that his theory is normatively adequate and that some attempts to show that it is adequate are unsuccessful: this still leaves open the possibility that orthodox theory *is* normatively adequate - that there *is* some, as yet unexamined, argument which can show its adequacy. I will give a reason, however, as to why the orthodox theorist should not hold out much hope for such an argument.

The basic problem here revolves around the notion of *formal practical rationality*. As we said, an individual exhibits formal practical rationality when he acts in a maximizing fashion relative to his beliefs and preferences, i.e., when he acts so as to maximize utility relative to his beliefs and preferences. Now this is a very "thin" sense of rationality: it does not require that an individual have any particular beliefs or that he value or disvalue anything in particular, although, as we noted, it does require that an individual's preferences at least be consistent (i.e., that they be complete and transitive). It is this sense of "rational" which is endemic to the contemporary Western characterization of rational man.[20] And it is not difficult to see why this should be so.

First, this sense is *so* thin that we feel inclined to say that if it isn't rational to act in a maximizing way, relative to one's beliefs and preferences, what is? Second, this thinner sense is less problematic than any thicker sense: it is far less problematic to say how an individual, qua rational man, is to act *given* his preferences, rather than to say what an individual, *qua* rational man, is to value. (We ignore the problem of what an individual, *qua* rational man, is to believe as this is not germane to our present discussion.) Hence decision theorists are, as Tversky remarks, "eager to tell people how to act, *in the light of their values*, but they are very reluctant to tell people how to feel, *or what values they should have.*"[21] In other words, decision theorists have been keen to tell people how to act if they are to act *rationally* in the formal practical sense.

How does the orthodox theorist tell people to act if they are to act rationally? As we have seen he tells them to act so as to maximize expected utility. But if in so doing an individual is to maximize utility relative kto his preferences, i.e., if he is to act in the formal practical rational manner at all, then his preferences must obey certain conditions, in particular, the Sure-thing Principle, which ensure that the utility of an alternative (lottery) is equal to its expected utility. But as we saw the Sure-thing Principle *precluded* an individual from having certain values, in particular, it

precluded an individual from valuing certainty as such. Hence, when the orthodox theorist tells people how to act if they are to act rationally, the sense of "rational" here goes well beyond the formal practical sense alluded to above.

Suppose the orthodox theorist now advances his theory as a normative theory, i.e., he admits that individuals are not actually rational, in his sense, but nonetheless they ought to be. But as the orthodox theorist's sense of "rational" goes beyond the formal practical sense and requires that an individual not have certain values, in particular, that he not value certainty as such, it is clear that the orthodox theorist is now advancing a theory which *prescribes* certain values on the grounds that it is rational to have certain values and not others. But what grounds could there be for supposing that it is rational to not value certainty as such? Clearly, the grounds do not come from a consideration of what it is to be rational in the minimal and exceptionally compelling sense of formal practical rationality, for this sense of rationality just does not require that an individual not have certain values.

A hint of a justification is given by the emphasized part of the quotation from Raiffa given above: that if individuals do not value certainty as such then they will do better in their choice behaviour than if they

do value certainty as such. But this will only be true, for example in the case of the Tversky-Kahneman experiment, if it is true that the utility of an alternative can be simply identified with the monetary payoffs. That is, individuals will do better in situations like the Tversky-Kahneman experiment *at least in terms of expected monetary gains* if they do not allow their choice to be affected by their attitude towards certainty. But the point demonstrated by the Tversky-Kahneman experiment is precisely that individuals do *not* simply identify utility with monetary payoffs. In other words, individuals would do better in terms of results if they took a purely result-orientated approach to choice, but individuals do not take such an approach because it is not just the results to which they ascribe utility, they also ascribe utility to the amount of risk involved in achieving those results. So the question still remains, how can we show that individuals ought to take a purely result-orientated approach to lotteries, i.e., that they ought not to value certainty?

Another way that we might attempt to justify the view that individuals ought to take a result-orientated attitude towards risk taking is suggested by Harsanyi[22]: he appeals to paradigm cases of rational decision makers - individuals whom we would regard as the epitome of wise and prudent men. Here Harsanyi cites responsible

business executives using their shareholders money and responsible political leaders acting on behalf of their constituents. As he says, such individuals are expected to achieve the best possible results. But I do not find this argument at all convincing just because it is not clear that such individuals are *simply* expected to achieve the best possible results. Normally, of course, such individuals are judged simply on the results they achieve and this because normally shareholders and constituents are not aware of the day-to-day decisions (and certainly not the details thereof) being made by their leaders: all they have to go on is the results achieved. But this fact is surely quite incidental. Suppose that a business executive was required to make a choice like that in the Tversky-Kahneman experiment and his shareholders knew the details of the decision situation. Would they not regard a choice of *A* over *B* as irresponsible? For afterall, in his choosing *A* rather than *B* his shareholders would realize that they are missing out on the prospect of getting $400 for certain. Would not the shareholders regard such an executive as a fellow far too willing to take risks? Certainly if the shareholders were anything like the ordinary men and women who were the subjects in the Tversky-Kahneman experiment they would. Now my argument here, of course, presupposes two things. First, not

only that shareholders (and constituents) *will* value certainty – a safe enough assumption in virtue of the fact that a variety of subjects were chosen for the Tversky-Kahneman experiment – but also that it is not the case that they *ought* to take a result-orientated approach. About this assumption I remark that it is difficult to see, much as before, what argument could be offered for the claim that shareholders (or constituents) ought to take a result-orientated attitude towards risk-taking. Second, I have assumed that by a "*responsible* business executive" we mean an individual who would be so regarded by his shareholders if they knew the details of his decision making. Now if Harsanyi does not mean by the term "responsible" what I mean by that term (in the present context) then he at least owes us some account of what he does mean by that term. And obviously if his examples are going to do the job he has set them, i.e., justifying the claim that decision makers ought to take a result-orientated approach, then we had better have a grip on the notion of responsible decision maker which is independent of the notion of a decision maker who takes a purely result-orientated approach: we cannot justify anything merely by definition.

The reader may grant that no argument has been offered as to why rational individuals ought not to value certainty, but now ask, "What argument is there

to show that such a claim is mistaken?" My response is to say that I have no conclusive argument and I don't see how there could be, but I also don't see how there could be a conclusive argument to the contrary. All I can offer is a *prima facie* case against the claim that individuals ought not to value certainty by pointing out that individuals whom we intuitively regard as perfectly rational - and here I have in mind not only the subjects of the Tversky-Kahneman experiment but also decision theorists such as Watkins and Maurice Allais who have thought that it is perfectly reasonable not to take a result-orientated approach to risk-taking - *do* value certainty as such. But I can offer no conclusive argument that the claim that individuals ought not to value certainty is mistaken. Equally no conclusive argument has been offered to the contrary. And the reason for this is, as I've said, that it is problematic what an individual , *qua rational man*, ought to value: it is difficult to see what argument *could* be offered here. To be sure attempts have been made in the Aristotelian and Kantian tradition to determine what ends a man simply as a rational man must have. And an orthodox theorist might now attempt such an approach himself. Now my point is not simply that such approaches are notoriously difficult of execution but also that it is difficult to see how these approaches are relevant to

the preference for certainty: is it at all plausible, for example, to suppose that not valuing certainty as such is necessary if a man is to achieve *eudaimonia?* To be sure, we may be able to get much further in specifying what an individual, *qua moral man*, ought to value, and we will return to this issue in Chapter V. In the meantime I suggest that the most fruitful approach to decision theory is to regard it as an empirical theory: questions of adequacy are, it seems to me, on this approach capable of definite answers. But before going on to consider some arguments which attempt to show that decision theory lacks empirical content, I want to *briefly* consider an argument which asserts that the claim that individuals act so as to maximize utility relative to their behaviour is obviously false, even allowing that individuals' preferences may include a preference for certainty as such. A consideration of the reply we make to this argument may give some added plausibility to the claim of those who maintain that the Rationality Principle lacks empirical content.

4. *"Subjective", "Ethical" and "Overall" Preferences*

We start with the observation that it just seems to be a fact that there are many cases where a person chooses a certain alternative *A* over another alternative

B and yet, as far as we can ascertain, he prefers *B* to *A*. Here is a simple and hackneyed example. A doctor goes to work in a leper colony somewhere in Northern Africa, but he prefers the good life to one of relative deprivation, he prefers to treat the typical middle class diseases to treating leprosy, he prefers a cool climate to a hot climate, and so on. Nonetheless, he chooses to work among the lepers because he believes he *ought* to do so. As a consequence it seems to be false to say that all individuals act in a maximizing fashion relative to their beliefs and preference, i.e., the Rationality Principle seems to be false.

However, the decision theorist will reply that the above argument founders on a too narrow conception of preference: decision theory should not be identified with the naive and now discredited theory of psychological egoism.[23] Rather, we should, with Harsanyi, distinguish between a person's *subjective* and *ethical* preferences. A person's subjective preferences are basically self-interested - they concern his welfare as he conceives it. On the other hand, a person's ethical preferences are defined by the equiprobality model (see Chapter II section 3). That is, they "express what this individual prefers (or, rather, would prefer) on the basis of impersonal social considerations alone".[24] Harsanyi gives an indication of what we are to do with this distinction

when commenting on a paper by Kurt Baier.[25] He asks the question of whether people can act contrary to their preferences (as is suggested by the example mentioned in the first paragraph)? His reply suggests that we make the following sort of response to that question. We distinguish between a person's *subjective* preferences and his *truly overall* preferences which we define as those that govern his actual behaviour. A person's truly overall preferences represent a *compromise* between his subjective and ethical preferences. The question as to whether people act contrary to their preferences can now be answered by saying that of course a person may act contrary to his subjective preferences but that they cannot act contrary to their truly overall preferences. (Of course, the force of the "cannot" here is of prime importance in our argument concerning the empirical status of decision theory, but we put that problem to one side for the moment). In other words, the sense of "preference" we are after when we say that all individuals act in a maximizing way relative to their preferences (and beliefs), is more or less the sense of "want" alluded to by William Alston:

> a sense of "want" that extends more widely than just the concept of a direction towards goals that are inherently attractive. For often ...

> one intentionally brings about a state of affairs that is not attractive in itself ... we are fishing for something like the notion of a disposition to "strive for" S, regardless of *what* the source of this striving is ...[26]

I grant that even the distinction between subjective preferences and ethical preferences may not give sufficient structure to the notion of preference to enable us to account for all the behaviour which is counter to subjective preference. As Amartya Sen has pointed out it is unclear what Harsanyi would say of those cases where an individual "departs from his personal welfare maximization ... not through an impartial concern for all, but through a sense of commitment to some particular group, say to the neighbourhood or to the social class to which he belongs."[27] Nonetheless, the distinction is sufficient to defuse the sort of counter-example we have mentioned above and we will suppose, for ease of argument, that in talking of subjective and ethical preferences we have exhausted the category of preference. The important point for our purposes is to simply note that the notion of preference employed in decision theory is wider than that found in ordinary discourse: the utility ascribed to an outcome is supposed to represent the value placed on that outcome by an individual from *all* viewpoints -

the selfish, self-interested, social, ethical and so on.

5. *The First Argument for the Empirical Vacuity of Decision Theory: Decision Theory is "Soft-edged"*.

The argument I will consider in this section has been succinctly put by Philip Pettit in an article where he claims that decision theory, or what he calls "Rational Man Theory" is "soft-edged". By which he means that any time an action occurs which is not in accord with the beliefs and(overall) preferences which we have determined independently[28] in advance of that action we can always amend the ascription of preference and/or belief that we originally gave to the individual whose action we are attempting to explain. As Pettit says:

> ... this is to say that we put down what seemed like an anomalous observation as being merely a mistake: we discount it and cling to the theory. The move is always at our disposal and, more than this, it is one to which we are a always forced. For we cannot begin to imagine what it would be like to resort to any of the deeper responses in squaring rational man theory with observed behaviour.[29]

Now it *completely* misses the point to respond to Pettit's argument by saying, "Its perfectly easy to imagine the sort of case to which Pettit refers - medicine is packed with theories about physiological reactions. It is perfectly possible that a person should have certain beliefs and preferences and then act in a non-maximizing way relative to those beliefs and preferences". The quick and obvious reply to this argument is to say that what these physiological theories describe are *reactions*, mere *bodily movements*, not *actions*. It *is* inconceivable that an individual should *act* in a non-maximizing way relative to his beliefs and preferences although, of course, it is not inconceivable that he should *react* in a non-maximizing way. The distinction here is the common place one in the philosophy of action between bodily movements on the one hand and actions on the other. It is the distinction between my arm rising and my raising my arm. The distinction is often blurred by the fact that the term "action" (and, in particular, the term "behaviour") is quite often used to refer to bodily movements *and* genuine action. Roughly the full-blooded sense of "action" refers to a bodily movement done for a reason (or done intentionally). I will not be concerned in this thesis to defend the distinction nor to explicate it in any greater detail. We will simply take the distinction as given and will rest content with a fairly

rough and intuitive understanding of it. The important point is that given this full-blooded sense of action then it certainly does seem impossible that a person should *act* contrary to his reasons, i.e., that he should *act* so as to *not* maximize utility relative to his beliefs and preferences. In which case Pettit's argument stands and it would seem that the Rationality Principle - the claim that all individuals act in maximizing way relative to their preferences and beliefs - lacks empirical content. However, as I will now attempt to show, it does *not* follow from Pettit's argument that the Rationality Principle is empirically vacuous.

Note that Pettit's argument is basically as follows. If we describe some individuals bodily movement as the performance of an action (or as certain behaviour) then it follows that the individual must have had certain beliefs and preferences (in particular, beliefs and preferences relative to which the action was utility maximizing). But if *these* beliefs and preferences are contrary to the ones which we ascribed to the individual by independent means in advance of that action then we are forced to amend the original ascription of belief and/or preference. In other words, Pettit's argument crucially depends on the claim we mentioned in section 1 that "If *A* occurs then *R* occurs" cannot be false. As we noted we would allow that this conditional cannot be

false. But does it follow from the fact that the thesis of contingent action-reason conditionals is *false* that the Rationality Principle is *unfalsifiable*? To see that there is no such implication consider what is required for there to be a falsifying instance to the Rationality Principle. We require that it may be the case that an individual has certain beliefs and preferences (determined by indepdedent means) and then for it not to be the case that he acts in a way that is maximizing relative to those beliefs and preferences. In short, we will have a falsifying instance to the Rationality Principle when an individual has certain beliefs and preferences and fails to act on them. Thus, we can agree with Pettit that *if* we describe some bodily movement as the performance of an action then it must have been done for a reason relative to which that action was utility maximizing (this follows from the putative fact that the thesis of contingent action - reason conditionals is false). And if that reason is contrary to the reason we originally ascribed, i.e., if that reason is such that a certain action is utility maximizing relative to it and in performing that action the individual cannot perform the action which is utility maximizing relative to the originally ascribed reason, then we must revise our original ascription. But note that we have said *if* we describe the bodily movement as the performance of a certain

action - it is *not necessary* that we so describe the bodily movement. And if we do not describe the bodily movement as the performance of an action - we describe it merely as a bodily movement - then we are not required to revise our original ascription of belief and preference. Providing, that it is possible that an individual should have certain beliefs and preferences and that he should fail to act in a utility maximizing manner relative to them, i.e., despite the occurence of this reason all that should occur is some bodily movement - no action occurs. In other words, provided it is possible the conditional "If *R* occurs then *A* occurs" is false, i.e., that the thesis of contingent reason-action conditionals is true.

There are two important points that must be made here. First, we must be willing to maintain that the individual failed to perform *A* not simply because he performed *A'* for which *R* is the reason and *R'* is a countervailing reason to *R*. For if we allowed that the only way the conditional "If *R* occurs then *A* occurs " could be false is that the agent performs some other action *A'* for which *R'* was the reason, then it would still be the case that *some* reason-action conditional was true, *viz.*, "If *R'* occurs then *A'* occurs", and hence, in such a case, we would not have a falsifying instance to the claim all individuals act in a utility maximizing

fashion relative to their beliefs and preferences, i.e., we would still not have a falsifying instance to the Rationality Principle The second point arises from the following natural thought. If we establish that a certain reason *R* has occurred, and *R* occurred in *optimal conditions* i.e., we ascertain that the individual is not paralysed or that the action *A* is for some other reason physically impossible to perform, and that the individual is not suffering a sudden attack of amnesia or that the action *A* is for some other reason mentally impossible to perform, then *A must* (logically) occur. But then surely the Rationality Principle lacks empirical content. Note that this idea that given *R* occurs under optimal conditions and no countervailing reason occurs then the action *A* for which *R* is the reason must necessarily occur has some added plausibility in virtue of the fact, as noted in section 4, that a person's preferences are taken to include his subjective and ethical preferences: maybe a person can fail to act according to his own self-interest, but surely no-one can fail to act according to his overall preferences. But if what we have just previously said is correct then we are saying that no genuinely falsifying instance can be advanced against the Rationality Principle: any time we have an apparently falsifying instance this will simply be due to a failure of certain background conditions (auxillary hypotheses). Hence, as we remarked

in section 1, if we are to claim that the Rationality Principle has empirical content we require that the conditional "If *R* occurs then *A* occurs" may be false even when no countervailing reason occurs *and* even when *R* occurs under optimal conditions. We will return to argue for this claim in the following Interpolation.

In the meantime I wish to emphasise that we have good reasons not to admit of a falsifying instance to the rationality principle, i.e., the Rationality Principle is just one of the principles mentioned in section 2 which we will treat as unfalsifiable. I suggest that whenever we are presented with an apparently falsifying instance to the Rationality Principle we will respond in either of the following two ways. Either we will attempt to modify our original ascription of preference and/or belief, or we will suppose that the reason has not occurred in optimal conditions. Suppose we admit a falsifying instance to the Rationality Principle, i.e., we admit that an individual has certain beliefs and preferences (under optimal conditions) and fails to act in a utility maximizing fashion relative to those beliefs and preferences. If so, then any individual's behaviour becomes inexplicable in terms of the individual's reasons. Of course, behaviour *may* remain explicable at some "deeper" level, say at the physiological level, but whether this is so or not we can certainly admit that our ability to

explain and predict nearly all human behaviour would be practically impossible in such terms and until such time as an appropriately formulated hypothesis is advanced to replace the Rationality Principle we need the Rationality Principle too badly to be free to discard it in the face of some putatively falsifying instance. In other words, we may admit that, for example, physicalism (a term used to denote the family of reductionist theories of the mind) remains a metaphysical possibility, i.e., that the prediction that empirical inquiry will show that the mind is identical to the brain makes *sense* whether or not the prediction turns out to be *true*[30], but this does not leave us free to abandon the Rationality Principle if we are to be actually able to explain and predict behaviour. Hence in response to an apparently falsifying instance to the Rationality Principle we have good reason to protect it from falsification by revising our original ascription of belief and/or preference. Hence I agree with Pettit that there is a sense in which we are *forced* in the face of any apparently anomalous observation to regard it as being merely a mistake and to cling to our theory, in particular, its central hypothesis, the Rationality Principle. But this is a sense quite consistent with the principle having empirical content. We are not forced to revise our original ascription of belief and preference because there is some "conceptual connection" between a

person's beliefs and preferences and his utility maximizing behaviour: we are forced to revise our original ascription on strong pragmatic grounds.

The second type of response, namely, where we suppose that the reason has not occurred in optimal conditions, may be required because, on occasions we are not in a position to revise our original ascription of belief and/or preference. I will discuss why we may be in such a position more fully in the next section. The point briefly is that given what we know of an individual (his life history) and his situation it is totally implausible to suppose that he has a set of beliefs and preferences other than those which we originally ascribed to him. In which case, if we wish to protect the Rationlity Principle from falsification, as I have said we have good reason to do, then on occasions we will be forced to say that at least one, and we may not be able to specify *which* one, of the background optimal conditions fails to obtain. That is, we will say things like, "the individual must have have been drunk, suffered a sudden attack of amnesia or paralysis, or be in some way mentally deranged". In other words, we say that there is something wrong with the physical or mental make-up of the individual such that despite the occurrence of the reason this fails to give rise to the normal maximizing behaviour. Now normally we will be very loathe to take this sort of

response to an apparently falsifying instance to the Rationality Principle. For although with this response we have protected the Rationality Principle from falsification we have bought this protection at a price, for we are now not in a position to explain this *particular* individual's behaviour. It is instructive here to look at a case study involving Vice-Admiral Tryon which Watkins has examined.[31]

In 1893 Admiral Tryon, "a brilliant officer, energetic, resourceful, imaginative, and destined for the top" ordered a manoeuvre which resulted in the drowning of 356 officers and men. Now the beliefs and preferences normally ascribed to Tryon on this occassion - beliefs and preferences which at least on the face of it Tryon must have had given what we know about Tryon and his situation - are such that his behaviour was certainly not utility maximizing relative *them*. The response of many of his contemporaries was to suppose that Tryon was drunk, suffering from fever, or as one naval officer put it "Though bodily he was present on the afternoon of June 22 the guiding brain that made him so dear to us was absent." Such a response may protect the Rationality Principle from falsification but only at the expense of rendering Tryon's behaviour inexplicable. For the only grounds we have, *in Tryon's case*, for supposing that some background optimal condition fails to obtain, the failure

of which we are offering as an explanation of his non-maximizing behaviour, is just that he failed to act in a maximizing way relative to the beliefs and preferences we originally ascribed to him. Watkins calls these "explanations" offered by Tryon's contemporaries "pseudo-explanations" and Carl Hempel has referred to this type of "explanation" as being "epistemically circular"[32]. Of course, in relatively rare cases we may have some independent evidence for a breakdown in the optimal conditions, typically this will be the case, for example, with respect to the inmates of psychiatric institutions. It is worth remarking that, as Watkins notes, Tryon quickly became in the eyes of many a suitable candidate for just such an institution, but we wonder, how could the brilliant, energetic, imaginative officer suddenly become as one with the inmates of asylums?

6. *The Second Argument: The Holism of Reasons, Actions and Rationality*.

The second argument I shall consider which attempts to establish that the Rationality Principle lacks empirical content makes reference to the *holistic* nature of the ascriptions of belief, action and rationality. This argument has been presented by Davidson. He maintains that:

> if we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behaviour, then we are committed to finding, in the pattern of behaviour, belief and desire, a large degree of rationality ... [33]

I will not consider here the claim that if we are to describe certain motions as behaviour then we must presuppose that the individual had certain beliefs and preferences: we've already argued that the denial of this claim is not required by our argument. Thus the claims that we will examine are those which concern the attribution of beliefs and preferences.

It is important to note that the idea that the ascriptions of beliefs and preferences is holistic does not deny the thesis of independent access. That is, it does not claim that we are restricted to inferring a person's beliefs and preferences from the action to be explained or predicted. For example, one way of determining a person's beliefs and preferences independently of the action to be explained or predicted is to ask the person what his beliefs and preferences are. Holism does not deny that we can ascertain a person's beliefs and preferences by such means nor that such means are unreliable. Rather it claims that any such verbal responses, if we are to take them as indicative of that

person's beliefs and preferences, require that we presuppose that the individual is rational. But if so a claim that people are rational is not an empirical hypothesis, rather it is presupposed in any attempt to determine what their beliefs and preferences are. Hence, the Rationality Principle, which claims that all people are rational, cannot be an empirical hypothesis but is rather a necessary presupposition of our even being able to determine what people's beliefs and preferences are. The argument proceeds along the following lines.

In the first instance let us restrict our attention to the ascription of belief. And suppose we do so by noting a person's verbal "reports". But Davidson argues that "we cannot understand what a man means by what he says without knowing a good deal about his beliefs".[34] The reason is that in order to understand a person's verbal behaviour we must be in a position to know when a speaker holds a sentence he uses true. Now a speaker will hold a sentence to be true partly in virtue of what he believes and partly in virtue of what he means by that sentence. So the problem of understanding or interpreting a person's verbal behaviour comes down to the problem of simultaneously determining the person's beliefs and meaning from the sentences to which a speaker subscribes. That is, we cannot determine what a person means by what he says without at the same time inferring

that he has certain beliefs. If we are to understand or interpret what a person says we cannot do so by simply considering an isolated unit (a word or a sentence), for it is only within the context of a system (a language) that we can specify the role of the units. The obvious strategy then to effect an understanding or translation of what a person says will be to assume that the system of beliefs underlying, as it were, the system of sentences, is consistent and correct according to our standards. By so doing we can pair up the sentences the speaker uses with the sentences we use and hold true in similar circumstances. This is the familiar thesis of the indeterminacy of translation.

Now I do not propose to deny this argument nor do I think it is necessary to do so in order to claim that the Rationality Principle has empirical content. Notice that the above argument establishes that if we are to ascribe beliefs to an individual on the basis of verbal behaviour then we must presuppose he is rational *in the sense that* he has a system of consistent and correct beliefs (by our standards). But this is *not* the sense or "rational" which is employed by the Rationality Principle when it asserts that all individuals are rational. And this is not to deny that there is a common, everyday sense in which to say that a person is rational *is* to say that they have a consistent and correct

system of beliefs. But we must make a distinction between two senses of "rational", namely, those which I called in section 1, *epistemic* rationality and *formal practical* rationality. Davidson's argument establishes that we must presuppose that individual's are epistemically rational, i.e., that they have a set of consistent and correct beliefs if we are to ascribe beliefs to them on the basis of their behaviour, in particular, their *verbal* behaviour. It does not establish that we must presuppose that individuals are rational in the formal practical sense, i.e., that we must presuppose that relative to a person's beliefs (and preferences) he acts in a way so as to maximize utility. But the Rationality Principle only asserts that individuals are rational in the formal practical sense, and hence Davidson's argument does not establish its empirical vacuity.

In point of fact I believe that those who wish to maintain that the Rationality Principle has empirical content, e.g., the decision theorist who advances his theory as an empirical theory, should welcome Davidson's argument. For now determining the initial conditions upon which we are to base our potentially falsifying predictions becomes that must less problematic. Indeed the argument allows us to supplement our argument for the thesis of independent access. According to Davidson when we ask an individual what his beliefs are with

respect to any decision situation then we must presuppose, if we are to understand what he is saying, that he has correct and consistent beliefs about that situation by our standards and, in particular, that the beliefs about what alternatives are available at what probabilities are correct and consistent by our standards. In other words, we seem to have an argument *a priori* for reading off a person's beliefs from the decision situation:- we must presuppose that he has that set of beliefs we would have in that situation, i.e., that he has that set of beliefs which we would count as correct and consistent for that situation.[35]

It might be objected that Davidson's argument *does* establish that we must presuppose that individuals are rational in the formal practical sense, in that we must assume that individuals in their verbal behaviour act in a way that is maximizing relative to their beliefs and their desires to express the belief communicated in the verbal utterance. I think this is correct, but the point is not damaging for the empirical status of the Rationality Principle. For what needs to be established to demonstrate that claim is that we must make a *global* assumption as regards the rationality of an individual. That is to say, even granting the above point, namely, that we must presuppose that the individual acts rationally in the formal practical sense as regards the verbal

expression of his beliefs, we have not thereby established that we must assume that an individual will act rationally relative to the beliefs (and preferences) which we have determined that he has on the basis of his verbal behaviour. Only if that is established have we shown by this argument the Rationality Principle lacks empirical content.

My argument also holds for Davidson's analogous argument as regards the attribution of preferences. Here we conclude that in order to ascribe preferences on the basis of a person's verbal behaviour we must presuppose that his preferences are consistent. Consider, for example, the presupposition that they are transitive. Davidson is no doubt correct when he says:

> I do not think we can clearly say what should convince us that a man at a given time (without change of mind) prefered a to b, b to c, and c to a. *The reason for our difficulty is that we cannot make good sense of an attribution of preference except against a background of coherent attitudes.*[36]

Now it is true that the Rationality Principle assumes, as we have already noted, that an individual's preference

ordering is consistent - in particular, that it obeys the transitivity condition. But as I also remarked in section 2 experimental tests to determine whether people's preferences *are* transitive have been inconclusive. Given this fact and the fact that the Rationality Principle *requires* that preferences are transitive, then I think the supporter of the Rationality Principle should be happy with Davidson's argument to the effect that this is *not* a matter for experimental test, rather preferences *must* be assumed to be transitive if we are to ascribe preferences at all. But in welcoming this argument the supporter of the Rationality Principle need not admit that this hypothesis, the central hypothesis of decision theory, is empirically vacuous for we have only established that we must assume *ends* rationality.

* * *

We seem entitled to conclude, then, from the arguments examined in this chapter that decision theory does have empirical content and that orthodox decision theory is inadequate from the empirical viewpoint. Our conclusions with respect to orthodox theory from the normative standpoint were somewhat less definite although we did claim that a *prima facie* case could be made against orthodox theory. But throughout our argument for its empirical inadequacy we have had to assume the truth of two theses, what I called the thesis of

independent access and the thesis of *contingent reason-action conditionals*. We must therefore, given the crucial role played by these two theses, provided some defence of them. This requires that we venture into an examination of an argument that has figured prominently in the philosophy of action - the so-called *Logical Connection Argument*. The literature associated with this argument is vast and I wish to be relatively brief in my treatment of it. So I have selected those issues and arguments which seem to me to be of the most importance for our general argument.

FOOTNOTES

1. I will ignore the possibility that the hypotheses of decision theory are *synthetic a priori* (see S J Latsis (1976) pp. 3-7).
2. A similar view was held by Anatol Rapaport at least with respect to the extension of decision theory to interdependent decision making, i.e., game theory. See Rapoport (1970) pp. 49-52.
3. For an exposition of this view and its application to the matters discussed in this chapter see Watkins (1970) pp. 173-174.
4. For example, see his (1966) p. 99. And as Hempel also says, while it is true that rational explanation (explanation in terms of reasons) conforms "essentially to one or the other of our two basic types of scientific explanation ... (this) result and the arguments that led to it do not in any way imply a mechanistic view of man, of society and of historical processes ..." p.26.
5. The names for the first two types of rationality and their characterization I have taken from G W Mortimore (1976) pp. 96-97.
6. Karl Popper (1973) p.179.
7. J W N Watkins (1970) especially p. 172.
8. Donald Davidson (1980a) p. 268.

9. J C Harsanyi (1979) p. 299.
10. See Amos Tversky (1975) especially pp. 164-168.
11. Milton Friedman and L J Savage (1948) especially pp. 293-297.
12. Watkins (1970) pp. 173-174.
13. For example, see and compare Amos Tversky (1969) and F Mosteller and P Nogee (1951). See also section 6.
14. Harsanyi (1978) p. 224, first emphasis mine.
15. Tversky (1975) p. 166.
16. The detailed construction of an alternative unorthodox theory of decision is beyond the scope of this thesis. It is sufficient for our purposes that we indicate where orthodox theory is mistaken and what any correct theory must take account of.
17. Davidson (1980a) p. 272. Davidson also quotes Tversky as being in support of this argument. As I read Tversky he argues that orthodox theory *has* been shown to be empirically (or descriptively) inadequate. But Tversky also argues that neither his experiment nor Allais' thought experiment demonstrate the *normative* inadequacy of orthodox theory. However, neither Davidson nor I are concerned at the moment with the adequacy of orthodox theory from the *normative* standpoint.
18. Davidson (1980a) p. 272.

19. Howard Raiffa (1970) pp. 81-82, my emphasis.
20. See David Garthier (1975) pp. 412-413 and S.I.Benn and G W Mortimore (1976) especially pp. 268-282.
21. Tversky (1975) p. 172, my emphasis.
22. Harsanyi (1978) p. 225. Harsanyi thinks that it is "probably a reasonably realistic descriptive prediction" that individuals, at least in "serious" decisions, will take a result orientated approach to risk-taking. About this we can see Harsanyi is mistaken, unless he wants to claim - but on what grounds? - that the decisions of the subjects in the Tversky-Kahneman experiment were not "serious". He also thinks it is an "obvious normative rationality requirement" but, as will be seen from the text, I think the adjective "obvious" is unwarranted here.
23. For a discussion of psychological egoism see Richard B Brandt (1959) especially pp. 371-372 and John Hospers (1963) especially pp. 141-157. A similar comment is also made by Harsanyi (1977a) p.27.
24. Harsanyi (1955) p.315.
25. Harsanyi (1977b) pp. 443-445.
26. William Alston (1974) p.79.
27. Amartya Sen (1979) p. 103. Although Harsanyi's later work, I think, indicates that committment is included in one's subjective preferences, see his (1979) p. 292.

28. This independent determination is possible according to one of the "rational man postulates": "Beliefs, desires and decision principles can sometimes be construed in advance of action from past behaviour - assuming personal consistency - and from present circumstances - assuming commonness of response. Other factors, such as emotional expression, may also facilitate this sort of interpretation." 1978 p. 45. I will have something more to say about the assumption of commonness of response in section 6.
29. Pettit (1978) p. 51.
30. The distinction is due to Richard Rorty (1965) p. 24.
31. Watkins (1970) pp. 211-216.
32. Carl Hempel (1965) p. 373.
33. Davidson (1980b) p. 237.
34. Davidson (1980b) p. 238.
35. This is the argument I mentioned in section 5 to the effect that it may not always be plausible to revise our ascriptions of belief and/or preferences on the face of some putatively falsifying instance to the Rationality Principle. I also remark that the argument seems extendable to the attribution of preferences to an individual, i.e., if we are to understand what a person says with respect to his preferences in some situation then we must presuppose that his preferences are more or less the ones we

would have in that situation given his life history. In other words, interestly, we seem to have an argument for the *similarly postulate* mentioned in Chapter III section 4.

INTERPOLATION

Two Theses about Reasons and Actions

As I noted and as we have seen in the previous chapter there were two theses that we required if we were to be able to plausibly argue that the Rationality Principle, some formulation of which will be the central hypothesis in a theory of decision, has empirical content. These two theses I referred to as, the thesis of *contingent reason-action conditionals*, and the thesis of *independent access*. To establish the truth of these two theses requires that we make an excursion into a minefield area in the philosophy of action that involves a venerable argument known as the Logical Connection Argument. I do not propose to give an exhaustive treatment of this argument - I do not propose to examine all the many formulations of it - it will suffice for my purposes if we simple examine those formulations that are necessary for me to establish the two theses mentioned above. Now the Logical Connection Argument is normally addressed against the causal theorist of action. Roughly, it goes as follows: since the relation between reason and action is *logical* in nature, and the relation between cause and

effect is *contingent*, reasons cannot be construed as the causes of actions. But, as I have previously remarked, we are not concerned to show that reasons are causes, hence we may allow that there are *other* formulations of the Logical Connection Argument not examined in this interpolation which establish some sort of "logical connection" between reasons and actions which is sufficient to show that reasons are not the causes of actions. (For example, maybe it can be shown that the thesis of contingent action-reason conditionals is false and *this* is sufficient to show that reasons cannot be causes.)

I shall consider the thesis of contingent reason-action conditionals first. This thesis asserts, it will be recalled, that if a reason *R* occurs then it is logically possible that the action *A* for which *R* is the reason should not occur, even if no countervailing reason *R'* occurs and even if *R* occurs the optimal conditions. This thesis we said was required if the Rationality Principle was to have empirical content. For consider, to summarize what I said earlier, if it is the case that should *R* occur under optimal conditions and no countervailing reason *R'* occurs *entails* that *A* occurs, then there would be no possibility of presenting a falsifying instance to the Rationality Principle. However, the thesis of contingent reason-action

conditionals has been denied.

The argument against the thesis of contingent reason-action conditionals has been nicely presented by Raziel Abelson[1]. He begins by considering the views of people like Bruce Goldberg to the effect that the thesis is obviously true. Goldberg rightly points out that even if (as some have maintained) the description of a reason necessarily "includes" a description of the action it does not follow that the occurrence of the reason entails the occurrence of the action:

> If I want to go to the theatre, does it follow that I go to the theatre? There are at least some occasions when we don't do what we want to do.[2]

Abelson grants that Goldberg is right in this, but denies that Goldberg has successfully shown thereby that there is not some sort of "logical bond" between reason and action. To demonstrate this point Abelson asks us to consider the following case:

> Assume that Jones wants, intends, desires, or in some sense has a motive to open the window. What does this entail about what he will do? Well, it entails that he will

> open the window, but it does not entail this *tout court*. It entails that he will open the window *provided* that no reason arises for his not doing so (e.g., a hurricane is blowing outside) and provided nothing prevents him (e.g., he is not paralyzed, and the window isn't stuck).[3]

In other words, Abelson claims that if R occurs and no countervailing reason R', occurs and R occurs under optimal conditions, then this *entails* that A occurs (where R is the reason for A). This, of course, is precisely what we denied when we asserted that the thesis of contingent reason-action conditionals was true. Hence the truth of a claim such as Goldberg's is not sufficient for the truth of that thesis and as I've already argued, it is the truth of *that* thesis rather than some such claim as Goldberg's that is required if the Rationality Principle is to have empirical content.

However, we must now ask, what *argument* does Abelson offer for his claim? Surely, the hard-nosed causal theorist - and we in our defence of the thesis of contingent reason-action conditionals - will simply say that Abelson is mistaken: it *is* logically possible that R occur under optimal conditions and R' not occur, *and* A not occur. That is, we deny, as our thesis asserts,

that if *R* occurs under optimal conditions and *R'* does not occur, then this *entails A* occurs. As far as I can see the only argument Abelson offers for his claim is as follows:

> To say "I want to open the window; nothing prevents me, and I have no reason or motive not to, not even the motive of laziness; but still, I won't open the window" is senseless. What on earth could I mean by 'want'?[4]

Now if by "senseless" Abelson means that such a sentence is contradictory (and this seems warranted in virtue of the fact that he talks of the *meaning* of "want") and if in his previous claim he means by "entails" that it would be contradictory to assert the antecedent and deny the consequent of "If *R* occurs under optimal conditions and *R'* does not occur, then *A* occurs", then the passage quoted above is not an *argument* for his original claim but is merely a *restatement* of it.[5] To be sure the restatement does make more specific in what Abelson thinks the entailment relation between "*R* occurs, *R'* does not occur, *R* occurs under optimal conditions" and "*A* occurs" consists; the entailment relation supposedly holds in virtue of the very meaning of the word "want" or "reason".

But is Abelson right in this? His rhetorical question at the conclusion of the quoted passage seems a bit lame: the obvious reply is, "Well, that sense of 'want' such that the sentence 'If I want to open the window...' is *not* contradictory".

But perhaps Abelson has some (unstated) reason for supposing that the sense of "want" or "reason" that he envisages is the correct or legitimate sense. Notice that this reason certainly requires to be spelt out for it certainly does not seem to be the case that "*R* occurs, *R'* does not occur, *R* occurs under optimal conditions, and *A* does not occur" is contradictory: at least my linguistic intuitions are such that to say this sort of thing is not the same as saying, "Joe is a bachelor and (the same) Joe is married". A natural thought is to suppose that evidence for Abelson's view is to be had from the fact that we never *would* admit to the situation where *R* occurs, etc. and *A* does not occur. But this is no evidence for saying that such a situation is logically impossible: that requires that we never *could* (logically) admit to such a situation, which still, of course, needs to be established. Similarly, even if we admit that such situations are unimaginable - where we understand "unimaginable" in some non-question begging sense, i.e., as not being equivalent to "contradictory" - this still does not establish what is required for Abelson's argu-

ment. Our position here is strengthened, I think, by the fact that we can explain why we never *would* admit to a situation where *R* occurs, etc. and *A* does not occur, and why we have come to find such situations in some sense unimaginable. As we noted we have good reason to discount any apparently falsifying instance to the Rationality Principle; it is also true that the thesis of contingent reason-action conditionals is needed if the Rationality Principle is to be falsifiable, i.e., if the Rationality Principle is to be falsifiable then it must at least be logically possible that *R* occurs, etc. and *A* does not occur; but if we actually admitted that there *was* a case, not merely that it was *logically possible* that there was a case, where *R* occurs, etc. and *A* does not occur, then we would have admitted a falsifying instance to a principle that lies at the very heart of our attempts to explain and predict human behaviour.

Let me now turn to the thesis of independent access. This is the thesis, it will be recalled, that said that we may know a person's reason without inferring it from the action for which that reason is the reason. That is, it asserts that it is not the case that the *only* way we can know a person's reason is to infer it from the performance of the action. Against those who think otherwise Alvin Goldman has mounted the following powerful objection: it is not the case that our only evidence for

knowing an agent's reasons (or wants) is from the performance or occurrence of the action, there are in addition

> such items as (1) other acts of the agent, including verbal avowals in particular; (2) antecedent events that may be causally relevant to wanting to do *A*, including, for example, other wants of the agent; and (3) want- manifestations that are not acts - e.g., facial expressions.[6]

(To appreciate Goldman's point about facial expressions we must remember that not all bodily movements are actions in the full-blown sense of "action".) There, one might have thought, was an end to the matter: the thesis of independent access is clearly established.

However, consider the following idea. It may be granted that of course it is not the case that the *only* way that we may determine a person's reason for action is from their action. Nonetheless, the occurrence of the action is surely our most *reliable* guide to their reasons - what action they perform is, as of were, the final arbiter to their reason. Thus, if someone acts in such a way that this is not consistent with the reason we have described to him by the means outlined by

Goldman above then we will, indeed must, revise our original ascription. So on this view while we may at least provisionally determine an individual's reason in advance of his action, any such ascription is always revisable in the light of what action he actually performs. This view seems to me quite correct, although somewhat misleading, but it in no way requires an abandonment of the thesis of independent access. I do not want to deny that an ascription of a reason to an individual by the means outlined by Goldman is only provisional - but then this in no way reflects on the suitability of such ascriptions for the purposes of empirical investigation (see the comments on basic statements in Chapter III section 6). And I do not want to deny that *if* we assert that an individual has *acted* in a certain way that is inconsistent with the reason ascribed to him by the means outlined by Goldman then we must revise our original reason ascription: this will follow from the fact, which we've granted, that the thesis of contingent *action-reason* conditionals is false. But this could only prove an embarrassment to the thesis of independent access if it is logically necessary that we describe the individual's bodily movement as behaviour or action. For then it would not be that independent ascriptions are just revisable, but that they are necessarily so. But as I have argued above it is not logically necessary that we describe an individual's

bodily movement as certain behaviour or action even if we allow that he had a certain reason for acting, the reason occurred in optimal conditions, and no counter-vailing reason occurred. That is, it is logically possible that even so an individual should fail to act - his bodily movement is just bodily movement - and in that case, even granting that the thesis of contingent *action-reason* conditionals is false, there is no necessity that we should revise our original reason ascription. Of course, as before, while such situations may be logically possible this is not to say that we would actually admit of such a case: we have good reason, to protect the Rationality Principle from falsification, to revise our original reason ascription, and in this sense we are *forced* to revise our original ascription. Moreover, it is in this way that we can say that a person's actions will be the most *reliable* guide to their reasons: they will, in fact, be heavily relied upon.

Hence I claim that if the thesis of contingent reason-action conditionals is true then the thesis of independent access is true. Or rather, more accurately, if the thesis of contingent reason-action conditionals is true (which I've argued is the case or, at least, that the arguments addressed against it are very weak) then, if we are to be able to ascribe reasons to an individual at all (which we seem able to do by the means

outlined by Goldman), then the thesis of independent access is true.

The consideration of the above objection enables us to more easily present an argument in reply to an objection to the thesis of independent access suggested by some remarks of James Otten.[7] In commenting on Goldman's argument Otten makes the point that the occurrence of *A* is the only *criterion* we have for the occurrence of *R* : the items mentioned by Goldman are merely *symptoms* for *R*. This point rests on the Wittgensteinian notion of a criterion which Otten puts as follows: if the occurrence of X is a *criterion* for the occurrence of Y, then it is a conceptual truth that the occurrence of X is evidence for the occurrence of Y; whereas if the occurrence of X is a *symptom* for the occurrence of Y, then it is a contingent truth that the occurrence of X is evidence for the occurrence of Y. Of the items mentioned by Goldman - other acts of the agent, antecendent events involving the agent, want-manifestations - Otten says:

> In the case of each of these proposed behavioural criteria the crucial question is whether the fact that they count as evidence for the existence of (*R*) is a conceptual truth or a merely contingent

> truth. And in each case the answer, it will be found, is that the fact is merely a contingent truth. There simply is no contradiction involved in saying, for instance, that salivating or licking one's lips is not evidence for wanting to eat; even though in fact these are manifestations of the want to eat. So, the only behavioural criterion, and indeed the only criterion at all, for a person's want to perform a certain action is the actual performance of that action.[8]

Now the last sentence of this quotation certainly sounds ominous for the thesis of independent access, but this is only because in ordinary discourse we do not always use the term "criterion" in the Wittgensteinian sense. Notice that when Otten says that the occurrence of the action *A* is our only criterion for the occurrence of the reason *R* he merely means (and must only mean this in the light of his account of the notion of a criterion) that the occurrence of *A* is the only behaviour which is such that it is a *conceptual* truth that the occurrence *A* is evidence for the occurrence of *R*. And it seems we should agree with Otten on this point in that we have granted that the thesis of contingent *action-reason* conditionals is false. For we have granted that if the action occurs

then this entails that a certain reason occurred, and thus it seems that it will be a conceptual truth that the occurrence of *A* is evidence for the occurrence of *R*. But this is not to say that the occurrence of *A* is the *only* evidence for the occurrence of *R*: Otten has neither claimed nor shown that the items mentioned by Goldman are not evidence *at all* for the occurrence of *R*; they may not be *criteria* (in the Wittgensteinian sense) but they are still evidence and quite reliable evidence. In general when we cite evidence for such-and-such being the case we do not cite *criteria*, but only, to use the terminology at hand, *symptoms*. (Thus, if I cite my evidence for you being in the room next door I say things like, "Well his light is on, there are sounds like his moving around in there ...", but of course it's quite consistent with that evidence that you should not be in your room.) Now the sort of point being made by Otten could only cause us to give up the thesis of independent access if it could be established not merely that the occurrence of *A* is criterial evidence for the occurrence of *R*, but also that it is the *only real evidence* we have for the occurrence of *R*: that our symptomatic evidence must must always be revisable in the light of the criterial evidence. But as I've argued it is quite possible that *R* should occur and that *A* not occur, in which case we have *no* criterial evidence for the individual's reason, but

only symptomatic evidence. And such evidence *is* evidence and there is no *necessity* that we should revise our reason ascription based on such evidence.

* * *

I trust now that we are in a position where we need not merely *assume* the truth of the thesis of contingent reason-action conditionals and the thesis of independent access, but that we actually have some good reason to believe them to be true. Let us then now return to our main argument and see what we can say about moral theory in the light of what we now know about decision theory.

FOOTNOTES

1. Raziel Abelson .(1969) especially pp. 183-184.
2. Bruce Goldberg (1965) p. 72.
3. Abelson (1969) p. 183.
4. Abelson (1969) p. 183.
5. In so saying, of course, I presuppose a *non-essentialist* interpretation of Abelson's argument. This, as I've intimated, seems justified and in any case for a discussion and rejection of an essentialist form of this sort of argument see William G Dean (1975) especially pp. 352-354.
6. Alvin Goldman (1970) p. 111.
7. James Otten (1977) pp. 734-736.
8. Otten (1977) p. 736.

CHAPTER V

UNORTHODOX DECISION THEORY AND THE DEFENCE OF CONTRACTARIANISM

1. Alternative Conceptions of Justice and Rationality.
2. Unrestricted and Restricted Applications of Maximin: Towards an Adequate Version of the Maximin Principle.
3. Maximin and the Original Position.
4. The Special Features of the Original Position.
5. A Reconsideration of the Objections to Maximin.
6. The Nature and Status of Preference Contractarianism.

1. *Alternative Conceptions of Justice and Rationality*

In Chapter II I pointed out that from orthodox decision theory together with a minimal characterization of what it is to make a moral judgement we can deduce utilitarianism. There we said that to make a moral judgement was to make a judgement that was impersonal and impartial, and this we further defined in terms of the equiprobability model. If an individual is to make such a judgement and he is to do so in a manner which, according to the orthodox theorist, is rational, i.e.,

he chooses so as to *maximize expected utility*, then he will choose as a utilitarian.

In contrast, John Rawls in his derivation of a contractarian theory of justice claims that a rational individual when making a judgement in a certain sort of situation which has certain features among which are those that will ensure that his judgement is a moral judgement, will *not* choose so as to maximize expected utility, rather he will employ what is known as the *maximin principle*. Not surprisingly Rawls' theory has been attacked by the orthodox decision theorists and their preference utilitarian brethren. It is the object of this chapter to defend Rawls against these attacks.

Now it is important to note the general thrust of the criticisms made by the orthodox theorists against Rawls. They do not dispute Rawls' idea that we can gain an important insight into what constitutes a just society by invoking the notion that it is that arrangement of the basic social structure that rational individuals would choose in situations like, what Rawls calls, "the Original Position": (A further account of the Original Position will be given in sections 3 and 4). As Harsanyi, an orthodox theorist and leading critic of Rawls, remarks, "In my opinion, the concept of the original position is a potentially very powerful analytical tool for clarifying the concept of justice and other aspects of morality."[1]

Indeed, Harsanyi notes that he used essentially the same idea himself in his work on the analysis of moral judgements when he employed the equiprobability model to determine when a judgement was a moral judgement. That is, both Harsanyi and Rawls agree that the hallmark of moral judgement is that it be impartial and impersonal, and a judgement is impartial and impersonal when it is made under the constraints imposed by the equiprobability model or the Original Position. Where the orthodox theorist disagrees with Rawls is in Rawls' claim that in the Original Position rational individuals will choose according to the maximin principle. To quote Harsanyi again:

> ... the usefulness of this concept (the concept of the Original Position) crucially depends on its being combined with a satisfactory decision rule. Unfortunately, Rawls chooses the maximin principle as the decision rule for the participants in the original position.[2]

For the orthodox theorist the satisfactory decision rule for the individuals in the Original Position is that of *expected utility maximization*. This rule combined with the notion of the Original Position gives rise to a theory quite distinct from Rawl's own theory. Rawls has

called it "average utilitarianism"[3] and we have called it preference utilitarianism. So we have two theories of justice (or morality more generally) divided by their differing views on what constitutes a satisfactory decision rule for individuals in situations like the Original Position; that is, the dispute between them is *decision theoretic* in character. It is this fact which determines the nature of my argument in the remainder of this chapter.

My defence of Rawls will *not*, in a number of respects, amount to a complete defence of Rawls' theory of justice. My intention is to provide a limited defence of Rawls by focussing upon the decision theoretic problem of what, if anything, can be said for Rawls' insistence on the maximin principle as the rule for rational choice for the individuals in the Original Position. Now in so doing there are aspects of Rawls' theory which by and large I will simply take for granted, but as we shall see these are not aspects that effect the substance of the dispute between Rawls and the orthodox theorist. I should hasten to add that my argument is not intended to be a piece of Rawlsian exegesis, indeed there are a number of quite significant departures from Rawls' theory that I will have to make in order to provide my defence. Nonetheless, I trust that the theory I defend is recognizably Rawlsian in spirit. In rough outline my argument will be as follows:

a *correct* theory of decision has the consequence that individuals in a situation with the "qualitative anatomy"[4] of the Original Position will choose according to the maximin principle, i.e., they will choose a society organized according to the Difference Principle.

Having briefly stated the two differing conceptions of justice let me outline the differing conceptions of rationality they presuppose. As we have said (Chapter II section 2) it hardly seems problematic that a rational individual when choosing under certainty will choose that alternative with the outcome that he most prefers. But there is less agreement about what account we are to give of rational choice in situations of risk or uncertainty. The two best known contenders here are that individuals will choose that alternative which *maximizes expected utility* and that individuals will choose that alternative which *maximins*.

The mark of the orthodox decision theorist is that he claims that rational individuals under conditions of risk and uncertainty will choose so as to maximize expected utility. In those situations where there are *no* objective probabilities, i.e., the situation is one of uncertainty, the individual is assumed to employ *subjective* probabilities and then to proceed to choose as in a situation of risk. On the other hand, as intimated above, it is claimed that rational individuals

under conditions of risk or uncertainty will choose that alternative which maximins. That is, they will choose that alternative which is such that it has the outcome with the *maximum minimum* utility. Now while *sometimes* that alternative which is maximin will also maximize expected utility[5], this is not, as we shall soon see, always the case. That is, there is only a partial extensional equivalence between the maximin principle and the maximization of expected utility principle. There is, then, a significant dispute between those decision theorists who champion the former principle and those who champion the latter. Moreover, up until recently the dispute has largely gone in favour of the orthodox decision theorist. However, the work of Tversky and Kahneman which we examined in Chapter IV (and also the work of others) has cast serious doubts on the adequacy of orthodox theory.

Let me very briefly recap the points I mentioned in the discussion of the Tversky-Kahneman experiment. As we noted a problem arises for orthodox theory because it cannot account for a certain attitude towards risk evident in what Tversky called *the certainty effect*. In their experiments Tversky and Kahneman showed that for a large number of subjects:

> the utility of a positive outcome appears

> greater when it is certain than when it is embedded in a gamble.

In other words, the experiments indicate that individuals value certainty as such, i.e., that they are risk averse. More fully, the experimental results suggest the hypothesis that an alternative which offers a *surety* of a particular outcome is valued more highly than some *risky* alternative *not* because the former has a higher expected utility than the latter, but rather because the former has a certainty advantage over the latter. Now as I argued orthodox theory cannot accommodate the experimental results: we can show that the results require the failure of at least one of the assumptions from which the expected utility theorem is deduced unless some *ad hoc* hypotheses are introduced. Moreover, pretty clearly, orthodox theory cannot entertain the hypothesis which *does* accommodate the experimental results, namely, that individuals value certainty as such. For in so saying we claim that an individual may choose one alternative rather than another not because it has a higher expected utility but because it has a certainty advantage.

Now the response that the orthodox theorist makes to the results of the Tversky-Kahneman experiment will vary according to the meta-level stance he takes towards decision theory. If he takes an *empirical* stance

towards decision theory then his theory has been falsified by the Tversky-Kahneman experiment. On the other hand, if he takes a *normative* stance towards decision theory then there is no question of a falsification or, more accurately, no possibility of demonstrating inadequacy by way of empirical data. However, I did argue that we can at least present a *prima facie* case against orthodox theory conceived of as a normative theory. And this because individuals whom intuitively we regard as perfectly rational do value certainty as such and that no argument has been presented - and there seems little hope of such an argument being presented - which can show that these individuals ought not to value certainty as such.

So the existence of the certainty effect gives us reason to doubt the adequacy of orthodox decision theory whether it is conceived of as a normative or empirical theory. However, having given reason to suspect the claim that rational individuals under conditions of risk or uncertainty choose so as to maximize expected utility it should not be thought that this straight-forwardly opens the way for maximin. As I shall show below there are problems with the maximin principle.

2. *Unrestricted and Restricted Applications of Maximin: Towards an Adequate Version of the Maximin Principle.*

In this section I propose to show the *inadequacy* of maximin if it is deemed to have *unrestricted* application; and then by a consideration of an example where maximin appears to be adequate to give a rationale for an a characterization of the types of decision situations to which maximin is to be restricted. This should also enable us to account for the putative counter-examples to maximin. The examples which I shall consider are, for reasons of clarity, quite simple but the argument is not adversely affected thereby.

There are two types of counter-example that I will present to demonstrate the inadequacy of the maximin principle if it is supposed to have unrestricted application. They are both cases of decision under risk, but the first is a case where according to maximin two alternatives are equivalent and yet where it is clear that it is rational to choose one alternative rather thatn the other. The second is a case where according to maximin it is rational to choose one alternative rather than the other and yet where it is obvious that the contrary choice is the rational choice.

The first counter-example we may represent by way of the following matrix;

	S_1	S_2
a_1	4	1
a_2	10	1

Fig. 1

Here which outcome will obtain depends on whether the agent chooses a_1 or a_2 and which of the equiprobable states of nature s_1 or s_2 obtains. The entry in each cell is the utility entry for each possible outcome. From the maximin point of view a rational individual could choose either a_1 or a_2: they have an identical maximum minimum utility of 1. And yet clearly if is rational to choose a_2 and *not* a_1.

It is possible to overcome this difficulty if we introduce the notion of a *dominant* alternative. The notion of dominance is familiar in decision theory and we can give an informal expression of it as follows. A dominant alternative is one which is such that no matter which state of nature obtains the decision maker does at least as well, and in at least one case does better, with that alternative than with any other alternative. Hence, in Figure 1 a_2 dominates a_1. This suggests a restriction on maximin by way of a lexical ordering of dominance and maximin: for any decision situation, if there is a dominant alternative a rational individual will choose that alternative, if not he will choose the maximin alternative.

However, this is not sufficient to fully overcome the problem for maximin as can be seen by way of the following example:

	S_1	S_2
a_1	4	1
a_2	1	10

Fig. 2

Here a_1 does *not* dominate a_2, nor vice versa. Also the alternatives have the same maximum minimum utility of 1. Hence, even given the above modification to the maximin principle a rational individual could choose either a_1 or a_2, and yet clearly it is rational to choose a_2 and *not* a_1

The second type of counter-example is representable as follows:

	S_1	S_2
a_1	4	1
a_2	0	10

Fig. 3

In this case a_1 is maximin relative to a_2 and hence given an unrestricted application of maximin we would require that individuals choose a_1 over a_2; and yet it seems clear that it would be rational to choose a_2 over a_1.

The presentation of the above counter-examples may

prompt the following two thoughts. First, as seems to be the case in the above counter-examples, rational individuals *always* choose the alternative that maximizes expected utility; and second, if individuals choose the maximin alternative that is only when the maximin alternative happens to be the alternative that maximizes expected utility.[7] However, both thoughts will be dispelled if we consider the following example.

Imagine a situation - one that is just the same as one of the situations employed by Tversky and Kahneman in their experiments and which we looked at in Chapter IV - where you are offered the choice between a_1 and a_2 given that the states of nature are equiprobable and the entry in each cell is the monetary payoff in dollars:

	S_1	S_2
a_1	400	400
a_2	1000	0

Fig. 4

As we saw most individuals choose, and I conjecture the reader would choose, a_1 rather than a_2. Now notice that in choosing a_1 rather than a_2 one has chosen the maximin alternative. But, most importantly, the question for us to consider is whether in choosing the maximin alternative one has chosen the alternative that maximizes expected utility? Well, suppose you are presented with the following choice:

a_3: a 1/10 chance of \$1000 or nothing

a_4: a 1/5 chance of \$400 or nothing

and you choose, as the subjects of the Tversky-Kahneman experiment choose, a_3 rather than a_4. As we noted it cannot be that in both these choices one is choosing to maximize expected utility. Rather we suggested that in choosing a_1 over a_2 individuals choose a_1 not because it maximizes expected utility, but because it enjoys a certainty advantage over a_2: it offers the *certainty* of receiving \$400 as against the possibility of getting considerably less, *viz.*, \$0. Thus, while we granted that the utility of money is marginally increasing so that u(1000) < 1000 (where u(0) = 0) we rejected the idea that the rate of decrease is rapid enough to ensure that u(1000) < 800. And this leaves us in a position to account for the choice of a_3 over a_4: a_3 maximizes expected utility relative to a_4. Notice also that a_3 and a_4 are *equivalent from the maximin point of view* - they both offer the identical maximum minimum utility of 0. This point will assume some importance in our later discussion.

So in the example of Figure 4 we seem to have a case where choice according to the maximin principle is appropriate, i.e., where it is rational to choose the maximin alternative, and where the maximin alternative is *not* the alternative that maximizes expected utility.

But there are other important lessons to be learnt from this example.

First, we have claimed that individuals choose the maximin alternative in the situation depicted in Figure 4 because it has a certainty advantage: *that* is the rationale we give for choosing according to the maximin principle in the above case. In general, of course, the maximin alternative will not offer the certainty of receiving the *same* payoff no matter which state of nature obtains, but only the certainty of receiving *at least* some amount, namely, the minimum possible payoff of the maximin alternative. But the rationale for choice according to the maximin principle will remain essentially the same. An individual chooses according to the maximin principle because the maximin alternative has a certainty advantage over its rivals: it offers the certainty of receiving, at least, a particular payoff as against the possibility of receiving less. As a consequence, if two alternatives are such that an individual can be sure that he will receive at least the same amount of utility with either alternative then neither alternative has a certainty advantage over the other and therefore the very rationale for choosing according to the maximin principle, in such a case, has disappeared. The maximin principle can only have plausible application where the certainty effect can operate, *ex hypothesi* this is not the case in

Figures 1 and 2: in both cases a_1 *and* a_2 offer the certainty of receiving at least one unit of utility. Similarly with respect to the choice between a_3 and a_4 above: they both offer the certainty of receiving at least \$0. In these cases where neither alternative has a certainty advantage over the other, there seems little else that the individuals can do other than choose that alternative which has the best prospects, i.e., to choose the alternative that maximizes expected utility.

Second, while we have claimed that rational individuals would choose the maximin alternative rather than the alternative that maximizes expected utility in Figure 4, we would not expect, and we would not think it reasonable *in some sense*, for individuals to choose a_1 rather than a_2 if the following were the case. Either a_2 offers \$1500 should s_1 obtain (and \$0 should s_2 obtain) or a_2 offers \$200 should s_2 obtain (and \$1000 should s_1 obtain). And this for the reason that we think that while a_1 has the advantage of offering \$400 for certain, this advantage can be nullufied if a_2 has possible gains (relative to the maximin alternative) that are sufficiently large or possible relative losses that are sufficiently small. But here we presuppose a *normal* degree of risk aversion. It is quite possible that an individual should be *so* risk averse that even in the altered situation he still prefers a_1 to a_2. Such individuals

as a matter of fact might be extremely rare, but surely they are not inconceivable. And we cannot say that such individuals choose irrationally - at least if we are using the term "rational" in the formal practical sense. Of course, we could go so far as to say of such individuals that they have an odd, bizzare or maybe even pathological aversion to risk, and hence that their choices are unreasonable in some sense. This would explain why we would find any choice of a_1 over a_2 in Figure 3 "irrational": given a normal degree or risk aversion one would choose a_2 over a_1 - only someone with a quite abnormal degree of risk aversion would choose otherwise.

A more precise statement of the above idea can be had by the introduction of the notion of *gambler-indifference maps*. I take the notion from Watkins who in turn acknowledges G L S Shackle[8]. A gambler-indifference map will be a measure of an individual's attitude towards certainty. *Stakes* or what an individual has or will receive for certain are measured along the x-axis and *prizes* or what an individual will receive at a certain probability are measured along the y-axis. Both stakes and prizes are in monetary terms or are representable in monetary terms. (This accords well with the idea which we will mention later that individuals choose between various allotments of the primary goods on the basis of

an *index* of the primary goods.) To draw a gambler-indifference curve we select some probability value, and for each point on the x-axis we seek to discover the prize such that an individual is indifferent between the gamble, i.e., the probability of receiving the prize, and the stake, i.e., what he has for certain. So for some given probability P we can obtain the monetary value of the prize O_r and the monetary value of the stake O_s such that an individual i is indifferent between $p.\ O_r$ and the stake O_s, i.e., we can determine:

$$u_i(O_s) = u_i(p.\ O_r)$$

We now draw for each probability value the straight-line passing through the origin that represents *numerically* "fair" combinations of stakes and prizes (thus for $p = \frac{1}{2}$, the prize is always twice the stake). Thus we might have something like the following:

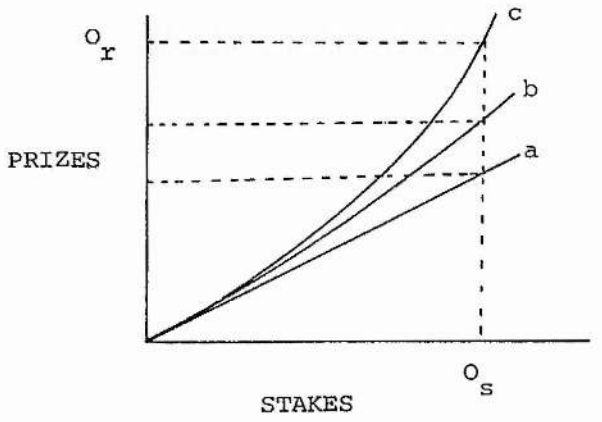


Fig. 5

The straight-line a represents the numerically "fair" combination of stake and prize; the curve b represents the curve we would expect given orthodox decision theory combined with the doctrine of marginally decreasing utility for money; and the curve c represents the curve we would expect given our examination of the Tversky-Kahneman experiment. Thus returning to Figure 4, and using a gambler-indifference map like that in Figure 5 we would note that the certainty (the stake) of \$400 is worth something less than, say, a ½ chance of \$1500 or nothing. Hence when the prize is increased to \$1500 (i.e., the situation is altered such that a_2 offers \$1500 should s_1 obtain and \$0 should s_2 obtain) then if the individual is to maximize utility he must choose a_2. When the possible relative loss is reduced we first obtain from the gambler-indifference map that value of the stake such that the utility of the stake for i is equal to u_i (½. 1000), and then the value of the stake such that the utility of the stake for i is equal to u_i (½. 200); we then add these together and if the value exceeds \$400 then the individual, if he is to choose the alternative that maximizes utility, must choose a_2. I should hasten to add that the gambler-indifference map in Figure 5 is a gross simplification: a more realistic map would have curves for a range of probability values between 0 and 1, and the curves for some probabilities and stakes, e.g., those involving very low probabilities and

very small stakes, may significantly depart from the shape of the curve in Figure 5. Nonetheless, for the range of values germane to our discussion (we are, for example, only concerned with relatively large stakes and maybe, too, only relatively large probabilities - see footnote 30) the shape of the curve in Figure 5 will suffice as an approximation.

For the remainder of our argument the following three points are especially worthy of note. First, the reason we suppose that individuals will choose the maximin alternative rather than an alternative that maximizes expected utility is that the former has a certainty advantage over the latter and that individuals value certainty as such. Second, that the degree to which individuals are risk averse, i.e., how much they value certainty, may vary from individual to individual: all we can talk of is a *normal* degree of risk aversion. And third, assuming a normal degree of risk aversion (or, indeed, any degree of risk aversion) there will be some limit to the size of the relative possible gains and losses of the alternative that maximizes expected utility if the maximin alternative is to be chosen: in particular, the gains must be sufficiently small or the losses sufficiently large.

We have, then, some idea of when and why the maximin principle is applicable. Let us now look at its role in the Original Position.

3. *Maximin and the Original Position.*

The Original Position is a purely hypothetical situation that replaces the historical or quasi-historical situations of earlier social contract theories wherein individuals agree on the arrangement of the basic social institutions. The Original Position has certain features such that any agreement reached in such a situation would satisfy at least the minimal conditions for being fair or just. Basically, the Original Position insures that the agreement reached is impartial and impersonal.[9] The impartiality and impersonality of the agreement is effected by supposing that the individuals in the Original Position choose between alternative arrangements behind the "veil of ignorance". That is, the individuals are assumed to make a choice under certain epistemic restraints. They do not know what their place will be in the society agreed upon - their class position or social status. They do not know particular facts about their own psychology, e.g., to what extent they are risk averse; although it is assumed that they know certain general facts, e.g., the laws of human psychology.[10]

When individuals make a choice between alternative arrangements of the basic social institutions they make a choice between alternative ways of distributing the *primary goods*, for it is those institutions which determine that distribution. A primary good is something which

any man will want (no matter what else he turns out to want once the veil of ignorance is lifted). The chief primary goods are rights and liberties, powers and opportunities, income and wealth.[11] It is assumed that the individuals want more rather than less of a primary good and that they are rational in the sense that they will choose accordingly. That is, it is assumed that the individuals in the Original Position are rational in the formal practical sense.

Now the important point for us to note at this juncture is that the veil of ignorance not only ensures that the agreement reached is impartial and impersonal, it also ensures that the choice is a decision under *uncertainty*. The veil of ignorance even precludes the individuals from knowing the *probability* of their occupying a particular social position in the future society.[12] On what basis, then, will an individual make his decision?

According to Rawls he will choose that arrangement of the social institutions which is an instantiation of the *Difference Principle*. That is, he will choose that arrangement of the social institutions which is such that it is to the greatest benefit of the least advantaged.[13] In other words, he chooses that arrangement which ensures that any distribution of the primary good maximizes the minimum allotment of the primary goods.

As Rawls notes there are two ways that one might proceed to argue for such a principle.[14] First, one might attend to what sort of social structure would result from the employment of this principle and then compare such a structure with our "considered judgements of justice". That is to say, we could attempt to justify the Difference Principle by that process which gives rise to a *reflective equilibrium*[15] for the principle. Alternatively, a "conclusive argument" (Rawls' phrase) could be had by noting that in a situation like the Original Position *rational* individuals would choose according to the Difference Principle. Says Rawls:

> In order to see how this might be done, it is a useful heuristic device to think of the (Difference Principle) as the maximin solution to the problem of social justice. There is an analogy between the (Difference Principle) and the maximin rule for choice under certainty.[16]

As I understand Rawls the important reason why there is only an analogy between the maximin principle and the Difference Principle is that the former only applies to situations where utilities are known whereas the latter is meant only to apply to situations where utilities are *not* known. The Original Position is a situation where

the utilities are not known: the veil of ignorance precludes the individuals from being able to assign a utility level to some allotment of a primary good. When choosing between alternative arrangements the individuals do so by means of an index of the primary goods expressed, say, in hundreds of dollars. Rawls regards this as a distinct advantage for his theory. If utilities were to be employed in choosing between differing institutional arrangements these utilities would have to be interpersonally comparable. Says Rawls of these interpersonal comparisons

> Simply because we do in fact make what we call interpersonal comparisons of well-being does not mean that we understand the basis of these comparisons or that we should accept them as sound. ... For questions of social justice we should try to find some objective grounds for these comparisons, ones that men can recognize and agree to. At the present time, there appears to be no satisfactory answer to these difficulties....[17]

However, Rawls is quick to note that while the Difference Principle is framed to circumvent these difficulties he does not wish to stress its relative merits

on that score. Primarily because he does not argue and he does not want to assume that the difficulties alluded to above for interpersonal comparisons of utility cannot be overcome.[18] This is a significant concession on Rawls' part and one that I think he was wise to make especially in the light of the comments made in Chapter II section 4. There we noted that the philosophical objectives to interpersonal comparisons of utility essentially involve the problem of other minds, and that as a point of strategy in philosophical argument we do not want our primary objection to a first order ethical theory to depend on such a general and apparently intractable problem. Of course, there is also the practical problem involved with interpersonal utility comparisons, and Rawls does allude to this, but I don't think that this is sufficient to persuade us that utilitarianism is inadequate as a moral theory. In any case Rawls seems to see the real difficulty with the use of such utilities residing in the fact that the utilities reflect values which "it does not make sense to pursue" or which are morally irrelevant. Of a utilitarianism which does use such utilities he says: "The controversy about interpersonal comparisons tends to obscure the real question, namely, whether the total (or average) happiness is to be maximized in the first place."[19] Now there is surely something right in what Rawls says: not *every* preference which a person has and which defines

his utility function should enter into a calculus of what constitutes a just society. Harsanyi has also recognised this problem:

> Common sense distinguishes between *sensible* preferences (sensible wants) and *foolish* preferences (foolish wants). It would be absurd for any ethical theory to disregard this distinction: nobody can seriously assert that we are just as duty-bound to help other people to satisfy their utterly foolish preferences as we are duty-bound to help them to satisfy their very sensible ones.[20]

But an ethical theory which bases itself on individuals' preferences seems to be in danger of losing this distinction. Harsanyi, however, argues that this need not be the case for:

> we may distinguish between a person's *explicit* preferences, i.e., his preferences as they actually *are*, possibly distorted by factual and logical errors - and his 'true' preferences, i.e., his preferences as they would be under 'ideal conditions' and, in particular, after careful reflection and in possession of all the

relevant information.[21]

Our ethical theory is then supposed to only take cognizance of an individual's true preferences. Indeed, Harsanyi thinks that we will have to make a further qualification, for our ethical theory should disregard "not only preferences distorted by factual or logical errors, but also preferences based on clearly antisocial attitudes, such as sadism, resentment, or malice."[22]

Now the distinctions here between explicit, true, and antisocial preferences raise a plethora of philosophical problems, not least of which is how Harsanyi is to *justify* the exclusion of, say, anti-social preferences: will this and can this be done simply on utilitarian grounds - if not, how can the theory be classified as genuinely utilitarian? Be that as it may I will not pursue the point further and instead I will mention the following three points which are especially worthy of note.

First, Rawls, as we've already noted, assumes that the individuals in the Original Position are rational in the formal practical sense. But he makes a "special assumption" about these individuals: they do not suffer from *envy*.[23] This assumption, whatever its merits, seems to be on a par with Harsanyi's assumption that the anti-social attitudes of sadism, resentment, etc., are to be

disregarded.

Second, while we've had to raise the veil of ignorance to some extent to allow the individuals in the Original Position to compute interpersonal utilities this does not seem to destroy the impartiality or impersonality of any agreement reached. It is still the case that no individual knows what his position will be in any future society nor the probability that he will occupy any particular position. Moreover, he does not, according to Harsanyi (see Chapter II section 3), compute utilities by supposing that he is in some social position with *his* preferences: rather, he supposes that he is the occupant of some social position with *that* individual's preferences. In more ordinary language, an individual ascribes a utility for each person in society for some given institutional arrangement by putting himself in the other fellow's shoes. Hence any agreement reached by the employment of such utilities would still seem to leave us with a decision situation whereby no individual could choose in a partial and personal manner. The veil of ignorance was designed to ensure that no individual could choose in a way that favoured his "particular condition". Our slight lifting of the veil of ignorance has not jeopardized this "fairness" of the choice in the Original Position. And it would seem that we can alter the features of the choice situation in any way we like

provided that fairness is not thereby jeopardized.

Third, having raised the veil of ignorance to give the individual's access to the preferences of individuals in society we have also given them, in particular, access to the preferences for certainty as such. The natural question is whether such preferences should be allowed to figure in our theory of justice. Let us in the first instance look at this question from Harsanyi's viewpoint. It would seem that if Harsanyi is to answer this question in the negative he would have to show either that such preferences were "utterly foolish". i.e., explicit preferences distorted by factual or logical error, or that they were antisocial. But where is someone who values certainty as such guilty of factual or logical error? Is such an attitude really to be put on a par with sadism, resentment, and malice? I think not. Moreover, we have already considered whether we can plausibly say that as rational individuals, individuals ought not to value certainty (see Chapter IV section 3). Let us now look at the question from the standpoint of some of the remarks that Rawls has made about the role of risk aversion in the Original Position. Rawls does not object to risk aversion affecting the choice of the individuals in the Original Position. Indeed, he thinks that given the nature of decision it is perfectly justified. Given that it is a decision where a whole life is at stake, that the

parties are responsible to later generations, and that the situation has other qualitative features (to be discussed in the next section) to choose in a risk averse manner is absolutely reasonable. This is not to say, of course, that the choice of an individual in the Original Position is simply to be a function of his own personal (and possibly idiosyncratic) attitude towards risk. That is why the veil of ignorance is deemed to preclude an individual from knowing his own attitude towards risk.[24] Now the approach I shall suggest has certain affinities with Rawls' approach such that I think I am justified in saying that it captures the essence of his position with respect to the role of risk aversion in the choices of the Original Position. I will argue that given that the Original Position has certain qualitative features, then assuming a normal attitude towards risk, the individuals will choose in a risk averse manner, and any such choice will not depend on each individual's own personal attitude towards risk.

To return to our main argument. The idea we want to pursue is that we can, apparently, offer a "conclusive argument" for the Difference Principle by noting that rational individuals will choose according to the maximin principle in the Original Position and that there is an identity, or at least a strong analogy, between the maximin principle and the Difference Principle.

But, as we've seen in section 2, maximin is not universally applicable. However, this is where it is crucial to note that Rawls does *not* maintain that the maximin principle is generally applicable, even in situations of uncertainty.[25] Rather it is applicable in "special circumstances", namely, those circumstances which have *three* features each of which is evident in the Original Position. These I call "the special features of the Original Position".

4. *The Special Features of the Original Position.*

The first feature of the choice in the Original Position which Rawls thinks makes application of the maximin principle particularly appropriate is that the individuals are not in a position to estimate the probability of their occupying any particular social position. Now I should note that I do not believe that it is *essential* to allude to this feature in order to defend Rawls' use of the maximin principle: that can be achieved with the remaining two features. Nonetheless, as I shall indicate below, I think there is more to be said for this feature than I suspect the orthodox theorist would allow.

As we noted, according to orthodox theory in a situation of uncertainty an individual is assumed to employ subjective probabilities. What value should be given to

these probabilities in such a situation as the Original Position? Harsanyi claims that individuals should assign an *equiprobability* to their occupying any particular social position. The basis for this assignment is the Laplacean principle of insufficient reason. Rawls thinks probabilities arrived at in this way should be discounted as a basis for rational decision in the Original Position. Such probabilities are required for the expected utility principle whereas, in contrast, the maximin principle does not require any assignment of probabilities: individuals simply choose that alternative, irrespective of the probability of the various outcomes, which maximizes the minimum utility.

This brings to mind the following complex problem in probability theory: the assignment of equiprobability on the grounds of insufficient reason seems to be *baseless*. Afterall, because of insufficient reason we seem to have no grounds for supposing that the probability of any particular outcome has any particular value - including that of equiprobability with every other outcome. We seem to have derived a conclusion upon which we are going to base a rational decision out of a state of ignorance. How is this possible?[26]

Rather than get embroiled in this difficult problem I propose to make the following two points. I have argued - in section 2 - that there is a case to be made

for the maximin principle as a rule for choice in situations of risk. If anything I think this case is stronger in situations of uncertainty. Consider once again the situation depicted in Figure 4, only this time suppose that the situation is one of uncertainty rather than risk. That is, suppose that the individual does not know the probability of s_1 obtaining or the probability of s_2 obtaining. Here it seems to me an individual who was risk averse would have additional reason for choosing a_1 rather than a_2. For the only reason he could have for not choosing a_1 was that he had good reason to believe that the probability of s_1 obtaining was at least ½. But does he have good reason to believe that, when he does not know the objective probability of s_1 obtaining? Or, at least, does he have *as good* a reason to believe that the probability of s_1 obtaining is ½ in the situation of uncertainty, as he would if he knew that whether s_1 obtained would depend on the result of the toss of a fair coin? If we're inclined to answer "No" here, on the grounds that probabilities estimated on the basis of insufficient reason are at least problematic or in some way less "firm", then it would seem that an individual has additional reason to choose a_1 rather than a_2. He seems entitled to *ignore* these somewhat problematic probability estimates and simply choose that alternative which offers him the surety of $400 as against the risk

of getting $0 - that seems quite reasonable.

This raises the second point that I wish to discuss with respect to the first feature of the Original Position. Harsanyi has argued against Rawls by claiming that "a rational decision maker simply *cannot help* using subjective probabilities".[27] Hence it would be false to suppose that by choosing according to the maximin principles individuals are avoiding the use of such probabilities. His argument proceeds by way of the following example.

	X wins	X doesn't win
Bet that X wins	100	0
Bet that X doesn't win	0	100

Fig. 6

Here one chooses between two bets (that X wins the next election and that X doesn't win the next election) such that if you bet that X wins and he wins you get $100 and nothing if he doesn't, and if you bet that X doesn't win and he doesn't you get $100 and nothing if he does. You pay nothing for either of these bets, so as Harsanyi remarks it would be irrational for you not to accept either bet - some chance of getting $100 is better than no chance at all. Then Harsanyi argues:

> if you choose the first bet then I can infer that (at least implicitly) you are assigning a subjective probability of ½ or *higher* to Mr X's winning the next election.

(A similar argument can be adduced if you choose the second bet.) But this seems to me to be an unwarranted conclusion. Suppose I know nothing about Mr X or his electorate, in short, I have no way of estimating that he will or will not win. Further, suppose I choose the first alternative (I bet that X will win) - does it follow that I have assigned a subjective probability of ½ or higher to his winning? Of course not. Maybe I chose the first alternative because it *was* the first alternative or, more realistically, because I liked the sound of Mr X's name. Certainly we could then say that the abstraction of the decision situation in Figure 6 does not adequately capture that decision situation as I perceive it: the bet that X will win has some additional utility for me that is not simply a matter of its monetary payoff. That can be granted but it does not effect the point that I choose the first alternative without assigning a subjective probability of ½ or higher to Mr X's winning. By betting that Mr X wins, then whether he wins or not, at least I have the satisfaction that I bet that he will win.

Harsanyi might object to this argument by saying, "But surely you wouldn't have chosen the first alternative on such trivial grounds unless you thought that the probability of his winning was at least ½." However, this misses the point: I do *not* assign any probability value to Mr X's winning; as I have no idea of the probability of whether Mr X will win or not I ignore all probability estimates; but I have to choose (a chance of $100 is better than no chance at all) and I have to choose on some basis; so I decide to choose on the basis that I like Mr X's name.

Orthodox theory, with its insistence that an individual will employ subjective probabilities in situations of uncertainty is in danger of being rendered a trivial theory whether it is conceived of as an empirical or normative theory. This is particularly apparent in Harsanyi's later remarks:

> if a decision maker follows the maximin principle, he is not really avoiding a choice of subjective probabilities, *at least implicitly. Of course, he may not think explicitly in terms of probabilities at all. But, whether he likes it or not, his behaviour will really amount to assigning probability one* (or nearly one) to the worst possibility in any given case. (My emphasis).

But now what appeared to be a substantive claim seems to be no more than a trivial truth. Whatever an individual may think, his *behaviour* will be taken to show that he employs subjective probabilities, and, moreover, subjective probabilities of a certain value. There is no way that his choice behaviour can be inconsistent with the claim that he employs subjective probabilities of a certain value, and theoretical objections and reports by the agent of his own decision processes are simply swept aside by *fiat*. I don't think that Rawls, or anyone else who objects to the use of subjective probabilities in situations of uncertainty, will be very impressed by this argument.

However, let us now turn to the remaining two features which, as I've said, will bear the burden of our defence of the maximin principle in the Original Position. I call these two features the "maxima and minima conditions".

The *maxima condition* is put by Rawls as follows: an individual in the Original Position

> cares very little, if anything, for what he might gain above the minimum stipend that he can, in fact, be sure of by following the maximin rule.[28]

The *minima condition* states that the individuals strongly

disvalue what they might loose relative to what they can be sure of by following the maximin rule:

> the rejected alternatives have outcomes that one can hardly accept.

Now these features are the features we mentioned in section 2 when discussing the applicability of the maximin principle to the decision situation in Figure 4. We said there that the maximin alternative (i.e., a_1) would be chosen provided the relative possible gains of a_2 were not large or that the relative possible losses were not small. As we noted, however, whether the possible gains *were* sufficiently large or whether the possible losses *were* sufficiently small would very much depend on how much the individual valued certainty as such: in our example we presupposed that individuals had a *normal* attitude towards risk. The question, then, is what is the degree of risk aversion of the individuals in the Original Position? Rawls has said, "From the standpoint of the original position, the parties will surely be very considerably risk-averse; if we ask how risk averse, we might say not less than that of most any normal person".[29] Let us proceed with something like this idea, i.e., we suppose that there is a certain value for certainty as such which the vast majority of individuals in the

Original Position place on certainty (the results of the Tversky-Kahneman experiment certainly indicate that there is such a value).

We now imagine an individual choosing between two alternative social arrangements: one is the maximin alternative and the other maximizes expected utility (we assume that the individuals assign equiprobability to their occupying any particular social position and hence our argument does not require that the individuals do not use subjective probabilities). We further assume that the maxima and minima conditions obtain. So we have the following situation:

	S_1	S_2	S_i	S_n
a_1	u_1^1	u_1^2	u_1^i	u_1^n
a_2	u_2^1	u_2^2	u_2^i	u_2^n

Fig. 7

Here the states of nature are the social positions in society and as there are n individuals in society (or, following Rawls, n "representative men"[30]) there are n such positions. The alternatives are a_1 and a_2 with utility entries as indicated. The sign "u_1^i" represents the utility for individual i (i.e., the utility for the individual in the s_i position) of the alternative a_1.

That is, it denotes the utility of some allotment of a primary good for the individual in the s_1 position under the social arrangement a_1. Suppose that the minimum utility entry in a_1 is greater than the minimum utility entry in a_2, i.e., a_1 is the maximin alternative, and also suppose that a_2 maximizes expected utility. Now an individual choosing between a_1 and a_2 does not know which of the $\{1, \ldots, n\}$ individuals he will be, although he does know - as we mentioned in section 3 - what each individual's attitude towards risk is. More importantly, supposing the maxima and minima conditions obtain, he knows that it is most likely he prefers a_1 to a_2. For to say that the maxima and minima conditions obtain is just to say that given a normal attitude towards risk (i.e., the attitude most people have) the relative possible gains/losses are not sufficiently large/small to overcome the advantage of certainty offered by the maximin alternative. Therefore, in choosing *as if* he had a normal attitude towards risk and thus choosing a_1 rather than a_2 he is more likely to choose the alternative he actually prefers. The situation is somewhat like the following: you are offered a choice between *A* and *B* but you don't know whether you prefer *A* to *B* or *vice versa*. However, you are reliably informed that the *probability* that you prefer *A* to *B* is very much *greater* than the probability that you prefer *A* to *B*. Hence, you choose *A*.

Of course, we would have to presuppose that *how much* utility might accrue to you, given you choose *B*, is not that much greater than the utility that might accrue to you, given you choose *A*. That is, we would have to presuppose that how much you might prefer *B* to *A* is not so large as for you to be willing to accept a considerably low probability of you preferring *B* to *A*. This presupposition is guaranteed in the case of the decision in the Original Position; for we assume that only a normal degree of risk aversion is required to find a_1 attractive relative to a_2, and, as we saw in the case of Figure 4, a normal degree of risk aversion will only suffice to outweigh a relatively marginal gain in expected utility offered by some more risky alternative. In other words, an individual will *not* choose the maximin alternative as the relative possible gains of the alternative that maximizes expected utility increase and/or as the relative possible losses decrease, i.e., as the expected utility of the alternative that maximizes expected utility increases to a value which exceeds the value of the advantage of certainty offered by the maximin alternative: for most people, i.e., those with a normal attitude towards risk, the value of the advantage of certainty offered by the maximin alternative is only sufficient to outweigh a marginal gain in expected utility. Hence, even if an individual does not value certainty as such,

a_2 will only offer him a marginal gain in (expected) utility relative to a_1.

There are two important questions raised by the above sort of approach.

First, it might be objected that as the individual in the Original Position making the choice between a_1 and a_2 in Figure 7 does not know whether he values certainty as such at the time of making his decision, and that by the time he *does* know (i.e., by the time the veil of ignorance is lifted) he will occupy one of the social positions, why does he not simply choose between the alternatives on the basis of which offers the highest expected utility? That is, given that the decision maker is ignorant of his attitude towards risk, why not take a purely *result orientated* approach to the choice between a_1 and a_2 in Figure 7? This is a serious objection, but not one that is decisive, although it does highlight certain issues that ought to be mentioned.

One thing we should note immediately. In saying that the individuals in the Original Position choose between alternatives without *knowing* their own personal attitude towards risk is not to say that they do not *have* an attitude towards certainty as such (even at the time of making the choice). If we have some difficulty in making sense of unknown but actual preferences we cah

always give a less *literal* reading of the Original Position which nonetheless preserves its essential character. We can require of the individuals that even though they know their own personal attitude towards risk they should choose *as if* they did not (see Chapter II section 3). Moreover, we do want the *actual* preferences of the individuals in society for certainty as such to figure in the determination of which alternative arrangement ought to be chosen from the viewpoint of justice. The reason for this is that any theory, including that being proposed here, which attempts to determine what constitutes a just society for some group of individuals by making reference to the preferences of those individuals should take account of *any* preferences those individuals may have *unless*, using Harsanyi's terminology, they are not *true* or *social* preferences. So, as the preference for certainty is neither "false" nor anti-social, we seem perfectly justified in making a modification to the description of the Original Position to ensure that this attitude figures in the choice of the individuals in the Original Position. The only thing we must observe when making such a modification is that we do not thereby disturb the impartially and impersonality of the decision reached. Both objectives - that the attitude towards certainty should figure in the choice and that the choice should be impartial - can be achieved in the following way.

We suppose that the individuals choose between the alternatives without knowing their personal attitude towards certainty, but knowing that having made the choice and before occupying a social position the veil of ignorance will be lifted to such an extent as to allow the individuals to know their personal attitude towards certainty. This implies that an individual choosing between a_1 and a_2 in Figure 7 knows that it is most likely that should he choose a_2 he will, when the partial lifting of the veil of ignorance occurs, *regret* not having chosen a_1 for he will discover that the latter *actually has a higher utility* for him: it enjoys a certainty advantage over a_2.

The second important question is this: what reason have we to suppose that the maxima and minima conditions obtain? That is, supposing that an individual in the Original Position will choose as if he had a normal attitude towards risk, how can we be sure that the maximin alternative will be such that any other alternative (in particular, that which maximizes expected utility) will have outcomes which represent relative gains for which he "cares very little" and relative losses which he "can hardly accept"? To come up with some sort of answer to this question we would have to look more closely at the notion of a primary good, how much of them is available for distribution, what distributions of them are possible, and what we would expect individuals' utility functions

for those primary goods to look like. Fortunately, however, these are problems that we can ignore for the purposes of this thesis.[31] For the orthodox theorist, and in particular the preference utilitarian, claims that the individuals in the Original Position should choose that alternative which maximizes expected utility and *not* the maximin alternative *whether or not* the maxima and minima conditions obtain. In other words, it may be that in the final analysis the sort of theory being put forward here - a theory which claims that individuals will choose the maximin alternative in a situation like the Original Position - fails because there is no good argument for the claim that in the Original Position the maxima and minima conditions obtain. But the orthodox theorist does not and cannot object to this sort of theory on those grounds: his is a theory which claims that individuals will *not* choose the maximin alternative *even if* the maxima and minima conditions obtain. The primary object of this thesis, and especially this chapter, has been to show that the orthodox theorist is mistaken in this claim. Whether the maxima and minima conditions obtain in the Original Position has not been and could not be the object of the dispute between the orthodox theorist and Rawls: rather theirs is the decision theoretic dispute that centres on whether the maximin principle is ever an appropriate selection rule for rational choice.

Having argued that under certain conditions the maximin principle *is* an appropriate selection rule and that it has been claimed by Rawls that the Original Position is a situation that satisfies those conditions let us now go on to consider the objections that have been raised by the orthodox theorist against the maximin principle.

5. *A Reconsideration of the Objections to Maximin.*

Here I propose to concentrate on the objections that Harsanyi has brought against the maximin principle.[32] His argument proceeds by way of counter-example and these may be divided into two types. First, there are those counter-examples which attempt to show that the maximin principle is inadequate as a selection rule for decision situations in everyday life - it leads to decisions which are clearly *irrational*. Second, there are those counter-examples which attempt to show that it is inadequate in the Original Position - it leads to decisions which are clearly *morally* unacceptable.

Consider Harsanyi's most well-developed counter-example of the first type. Suppose you live in New York and are offered two jobs at the same time. The first is a boring, low paid job in New York; the second is an exciting, well-paid job in Chicago. But to get to Chicago to take up the job there you must take a plane trip for which there

is a very small but positive probability that you will be killed in a plane crash. The situation can be summarised as follows:

	Plane accident	No plane accident
Choose N.Y. job	poor job, but stay alive	poor job, but stay alive
Choose Chicago job	death	good job, and stay alive

Fig. 8

As Harsanyi notes the *worst* you can do choosing the New York job is *better* than the *worst* you can do choosing the Chicago job. So, concludes Harsanyi, if you are to choose according to the maximin principle you must choose the New York job, and yet that choice seems to be clearly irrational.

There are two points that Harsanyi makes about the situation depicted in Figure 8 to which I would like to draw attention. First he says, "I am assuming that your chances of dying in the near future for reasons other than a plane accident can be taken to be zero"; and second, he says that if you are to follow the maximin principle "you must choose the Chicago job *under any conditions* - however unlikely you might think a plane accident would be, and however strong your preferences

might be for the excellent Chicago job".

It is obvious enough why Harsanyi must assume that the probability you will die for reasons other than a plane accident is zero, for otherwise it might not be that the maximin principle clearly recommended that you choose the New York job. Now it might just be a fact that this assumption holds - there might really be no possibility that you will die by any other means than a plane crash. But this hardly sounds like a decision situation of "everyday life": there would be in reality a very small but positive probability that you would die, say, from being struck by lightning or being run over by a bus if you stayed in New York. In which case we might ask why Harsanyi has picked on the *way out* possibility that you will die in a plane crash from the *miriade* way out possibilities associated with your choice, in particular, those associated with your staying in New York? To make a point similar to that made by Hare in another connection (see Chapter III section 6), can we be sure that our commonsense judgements - our judgements about what it is clearly rational to do in *everyday* situations are applicable to the *completely uncommon* situation envisaged by Harsanyi? Moreover, if we do side with Harsanyi that it is clearly irrational to not choose the Chicago job, may be this is because we have (unconsciously) tempered our judgement by associating

with staying in New York the very possibility that we will die by means other than a plane crash which would be implicit in any *normal* choice between staying in New York and moving to Chicago.

But the more serious problem for Harsanyi's putative counter-example is this: it mignt be a counter-example to an *unrestricted* version of maximin, but it is not a counter-example to the restricted version of maximin being put forward by Rawls and myself. Neither Rawls nor I would claim that you must choose the New York job "under any conditions" - no matter how unlikely you might think a plane accident would be, and no matter how much you valued the Chicago job. The certainty advantage enjoyed by choosing the New York job is not sufficient to overcome the advantages of choosing the Chicago job, viz., the very good prospects of an excellent job with only a minute possibility of a catastrophe. That is, the relative possible gains of the alternative that maximizes expected utility *are* sufficiently large and the relative possible losses *are* sufficiently small to outweigh the advantage of certainty offered by the maximin alternative. Of course, in saying this we assume a normal degree of risk aversion: it is at least conceivable that there *is* a person who would find the choice of the Chicago job too risky.

Let us now turn to the second type of counter-example I mentioned. Consider a society consisting of two indiv-

iduals, both of them critically ill, and suppose there is an antibiotic for which there is enough to cure one of the individuals. Individual *A* is a basically healthy person (apart from the present illness which can be cured by the antibiotic) but individual *B* is suffering from a terminal disease in addition to the present illness - nonetheless, the antibiotic will prolong the life of *B* for another couple of months. To whom ought the antibiotic be given? As Harsanyi notes, Rawls seems committed to the view that the antibiotic ought to be given to *B* on the grounds that this is to the greatest benefit of the least advantaged member of society, i.e., giving the antibiotic to *B* is recommended by the Difference Principle. But, as Harsanyi notes:

> In contrast, utilitarian ethics - as well as ordinary common sense - would make the opposite suggestion. The antibiotics should be given to *A* because it would do "much more good" by bringing him back to normal health than it would do by slightly prolonging the life of a hopelessly sick individual.

Now we will ignore the sorts of replies that might be prompted by a consideration of certain remarks by Rawls. For example, Rawls says that when judging the

principles of justice we should not suppose that they apply "to distributions of particular goods to particular individuals who may be identified by their proper names. ... They are meant to regulate basic institutional arrangements. ... Our common sense intuitions for the former may be a poor guide to the latter."[33] Instead we will tackle this putative counter-example head on and in much the same way as with the previous counter-example depicted in Figure 8. The point can be made as follows. A just society is one organised according to the Difference Principle because that is the sort of society that rational individuals would choose in the Original Principle. We know this because it is rational to choose according to the maximin principle in a situation with the qualitative anatomy of the Original Position. But in so saying we presuppose that the maxima and minima conditions obtain. Clearly in the sort of situation envisaged by Harsanyi they do *not*: there *is* an alternative arrangement to that recommended by the maximin principle where an individual (viz., *A*) can do *much* better. Now it might be thought that Harsanyi has a ready reply to this objection. After-all, Harsanyi has come up with an example - and one that seems quite possible - where it is not the case that the maxima and minima conditions obtain. Surely, if a counter-example can be so readily and easily constructed against Rawls' theory this demonstrates the inadequacy of Rawls'

theory. But this is not right. In the first place we allow that it is logically possible that in *some* situation the maxima and minima conditions do not obtain: this is obviously correct. But the important point is whether in *the Original Position* the maxima and minima conditions obtain. It may be that it is only a contingent fact about ourselves and our world that they *do* obtain in the Original Position and in which case this says something important about our theory of justice which I will discuss in the next section. Be that as it may, with the mere presentation of the above example Harsanyi has not given us reason to suppose that the maxima and minima conditions do not obtain in the Original Position which is what is required if we are to demonstrate the inadequacy of Rawls' theory along these lines. In the second place, it cannot be that Harsanyi's objection to Rawls' theory *is along these lines*. For, as we remarked at the conclusion of section 4, Harsanyi, as an orthodox theorist, must argue that *even if* the maxima and minima conditions obtained it would not be rational for individuals to choose according to the maximin principle in the Original Position.

I conclude therefore that neither type of counter-example presented by Harsanyi is successful in demonstrating the inadequacy of the maximin principle either from the rational or moral point of view. Our argument for the adequacy of the maximin principle, in particular, for its

proper place in a theory of justice, still stands.

6. *The Nature and Status of Preference Contractarianism.*

In this concluding section I wish to make a few brief clarificatory remarks concerning the nature and status of the sort of theory advanced and defended in this chapter.

There is at least one important respect in which the theory I have attempted to defend is similar to the theory advanced by Harsanyi and which we have referred to as "preference utilitarianism": they are similar in that they both make reference to the *preferences* of the individuals in society in order to determine what constitutes the just society. Hence I call the sort of theory I have defended, *preference contractarianism*. There are also other respects in which our theories are similar. We would both agree that we can determine what constitutes the just society by noting that it is that sort of society that would be chosen by *rational* individuals in a situation like the Original Position which ensured the *impartiality* and *impersonality* of the choice made. Where we disagree is over what is the rational selection rule for the individuals in the Original Position. I maintain that provided the maxima and minima conditions obtain individuals will choose according to the maximin principle, whereas Harsanyi maintains that whether or not the maxima and

minima conditions obtain individuals will choose according to the maximization of expected utility principle. It is here that preference contractarianism and preference utilitarianism essentially differ. And it is here that preference contractarianism is essentially similar to Rawlsian contractarianism: we both agree that the individuals in the Original Position will choose according to the maximin principle and hence that a just society will be one that is organised according to the Difference Principle. It is this principle, rather than the average utility principle, which is the principle of justice.

But now notice that our argument for the Difference Principle involved us in saying, *inter alia*, that(most) individuals in the Original Position value certainty as such and that the maxima and minima conditions obtain. Consider the idea that the individuals value certainty as such. Now it might be said, and quite rightly, that it is only a *contingent* fact that the individuals value certainty: surely it is logically possible for some society of individuals that they *not* value certainty as such. As we noted in our argument for the correctness of unorthodox decision theory we could only *definitely* say that it was adequate as an *empirical* theory of actual human behaviour, and from the *normative* viewpoint that there was no good argument to show that individuals ought not to value certainty, which is *not* to say that we have

a good argument to show that they *ought* to value certainty. But if so, it seems that it is possible and rationally permissible that there should be a society - *not our own* - for which the Difference Principle is *not* the principle of justice. Thus it may be that in a society where the individuals are not risk averse they should choose in a situation like the Original Position not according to the maximin principle but according to the expected utility principle; and for *these* individuals the Difference Principle would most certainly not be the principle of justice. One may be tempted to see this "relativization" of justice inherent in the sort of approach we have taken as a weakness in that approach. But I side with Rawls in rejecting the idea put by some philosophers

> that ethical first principles should be independent of all contingent assumptions ... (and that) moral concepts should hold for all possible worlds.[34]

Rather

> the fundamental principles of justice quite properly depend upon the natural facts about men in society.

It is required of those who think otherwise to advance a coherent theory - in that sense of "coherent" made explicit in Chapter II section 1 - which is such that it makes clear how we come by or how we come to an understanding of these immutable ethical first principles. Preference contractarianism *is* a coherent theory and it *does* make abundantly clear how we come to an understanding of first principles. And recall that we listed coherency and understanding as among the advantages of utilitarianism: in advancing a theory quite at variance with utilitarianism we do not want, if possible, to loose these advantages. Of course, if ethical first principles were necessary and self-evident this would give us immutable first principles that were known; but this is just the *a priorism* which we discarded in Chapter III section 6. None of this amounts to saying, however, that the approach we have taken is not completely general in that it may be used to determine what constitutes a just society for any possible society of individuals, at least, for any possible possible society where the individuals can be properly regarded as subjects of a theory of rational decision (whether this is conceived of as a normative or empirical theory). The just society, i.e., a society organized according to principles which are properly called just, for any group of individuals will be simply that which would have been rationally chosen by those individuals in

a situation which ensured that their choice was impartial and impersonal. With respect to *us* in *our* world we have reason to believe that the just society is one organized according to the Difference Principle.

So it would seem that by looking at the foundation of first order ethical theories, in particular, their foundation in a theory of rational decision making, we can reach some important conclusions with respect to the inadequacy of those ethical theories: namely, we have seen that we have good reason to reject utilitarianism and good reason to embrace its most viable rival, contractarianism.

FOOTNOTES:

1. Harsanyi (1975), pp.594-595.
2. Harsanyi (1975), p. 595. And as Harsanyi goes on to say, "By the very nature of the maximin principle, this choice cannot fail to have highly paradoxical implications."
3. See especially Rawls (1972) sections 27 & 28.
4. The phrase is Rawls', see his (1972), p. 157.
5. See also footnote 7. Here is an example (for an interpretation of the matrix the reader is referred to section 2):

	S_1	S_2
a_1	5	3
a_2	2	1

6. I have taken the example from Tisdell (1968), pp. 34-35. Tisdell also mentions the idea of Milnor (1954), pp. 49-59 that this sort of counter-example can be overcome by introducing the notion of dominance.
7. The thought is Harsanyi's: "Of course, Rawls is right when he argues that in *some* situations the maximin principle will lead to reasonable decisions ... But closer inspection will show that this will happen only in those situations where the maximin principle is essentially *equivalent* to the expected-utility

maximization principle (in the sense that the policies suggested by the former will yield expected-utility levels as high, *or almost as high*, as the policies suggested by the latter would yield)." (1975), p. 595, last emphasis mine. The phrase I have emphasised is interesting in the light of my later arguments. Harsanyi appears to be suggesting that on occasions it would be reasonable (rational?) to choose the maximin alternative even when it offers marginally less expected utility. Why would it be reasonable to choose thus, and how is that consistent with his claim that the expected utility maximization principle is the rational selection rule? On the other hand, as I shall argue, assuming a normal degree or risk aversion and a theory which takes cognizance of such an attitude we can easily account for such choices.

8. Watkins (1970), p.188.
9. See Rawls (1972), p. 12.
10. Rawls (1972), especially, pp. 136-137.
11. Rawls (1972), p. 62.
12. Rawls (1972), p. 155.
13. Rawls (1972(, p. 83.
14. Ralls (1972), p.152.
15. Rawls (1972), p. 20.
16. Rawls (1972), p. 152.

17. Rawls (1972), pp. 90-91.
18. Rawls (1972), p. 91.
19. Rawls (1972), p. 91.
20. Harsanyi (1977), pp. 29-30.
21. Harsanyi (1977), p. 29.
22. Harsanyi (1977), p. 30.
23. Rawls (1972), p. 143.
24. Rawls (1972), p. 137.
25. Rawls (1972), p. 155.
26. For a discussion of the issues raised here see, for example, Lucas (1970), pp. 109-125.
27. Harsanyi (1975), p. 599.
28. Rawls (1972), p. 154.
29. Rawls (1974), p. 143.
30. For this notion which has not played a role in our argument see Rawls (1972), p. 64. It is worth noting that there are certain advantages for our theory in employing this notion depending, of course, on how its defined. But suppose we identify representative men with representative individuals from social classes then this will render any calculation of whether a_1 or a_2 is to be adopted that much easier and it will also have the effect that the probability values involved in any calculation are quite large rather than very small.
31. For a discussion of the sorts of issues mentioned

here see James Fishkin (1975) especially pp. 616-620. Of course, it would be desirable that the maxima and minima conditions at least have some initial plausibility. Consider the primary good of income: it at least seems plausible that there should be some income such that people's utility curves for income should from a low plateau rapidly increase at that point (say, a point that represents a"decent standard of living") and that the curve should flatten out very considerably past that point. We suppose that the amount of income available is not infinite or, at least, is not such that it does not matter how we distribute the income (there is hardly a problem for justice with respect to income if no matter how we distribute the income everyone is assured of an income of at least $\$10^{10}$ per year) and we suppose that the maximin alternative ensures that each individual will get at least that amount which represents a "decent standard of living". Any other distribution, in particular, that which maximizes expected utility, will then have outcomes that represent gains for which everyone cares very little and losses that everyone can hardly accept.

32. See Harsanyi (1975), especially pp. 595-597.

33. Rawls (1972), p. 64.

34. Rawls (1972), p. 159.

SELECTED BIBLIOGRAPHY

(Books and articles referred to in the text)

R Abelson (1969), "Doing, Causing, and Causing to Do", Journal of Philosophy, 66, pp. 178-192.

W Alston (1974), "Conceptual Prolegomena to a Psychological Theory of Intentional Action", in Philosophy of Psychology, S C Brown (ed.), London.

Aristotle, Nichomachean Ethics (J A K Thomson translation).

J Austin (1954), The Province of Jurisprudence Determined, introduction by H L A Hart, first published 1832, London.

S I Benn & G W Mortimore (1976), "Can Ends be Rational? The Methodological Implications", in Rationality and the Social Sciences, S I Benn and G W Mortimore (eds), London.

R B Brandt (1959), Ethical Theory, New Jersey.

D Davidson (1980a), "Hempel on Explaining Action", in Essays on Actions and Events, Oxford.

D Davidson (1980b), "Psychologoy as Philosophy", in Essays on Actions and Events, Oxford.

G Ezorsky (1968), "A Defense of Rule Utilitarianism against David Lyons who insists tieing it to Act Utilitarianism, plus a brand new way of checking out general Utilitarian Properties", Journal of Philosophy, 65, pp. 533-544.

J Fishkin (1975), "Justice and Rationality: Some Objections to the Central Argument in Rawls's Theory", American Political Science Review, 69, pp. 615-629.

M Friedman & L J Savage (1948), "The Utility Analysis of Choices Involving Risk", Journal of Political Economy, 56, pp. 279-304.

D Gauthier (1975), "Reason and Maximization", Canadian Journal of Philosophy, 4, pp. 411-433.

B Goldberg (1965), "Can a Desire Be a Cause?", Analysis, pp. 70-72.

A Goldman (1970), A Theory of Human Action, New Jersey.

R M Hare (1963), Freedom and Reason, Oxford.

R M Hare (1981), Moral Thinking, Oxford.

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

J Harsanyi (1975), "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory", American Political Science Review, 69, pp. 594-606.

J Harsanyi (1977), "Rule Utilitarianism and Decision Theory", Erkenntnis, 11, pp. 25-53.

J Harsanyi (1977b), "Morality and the Prisoners' Dilemma Problem: Comments on Baier's Paper", Erkenntnis, 11. pp. 441-446.

J Harsanyi (1978), "Bayesian Decision Theory and Utilitarian Ethics", Americal Economic Review, Papers and Proc., 68, pp. 223-228.

J Harsanyi (1979), "Bayesian Decision Theory, Rule Utilitarianism, and Arrow's Impossibility Thoerem", Theory and Decision, 11, pp. 289-317.

C Hempel (1965), Aspects of Scientific Explanation, New York.

C Hempel (1966), "Explanation in Science and History", in Philosophical Analysis and History, William Dray (ed.), New York.

J Hospers (1963), Human Conduct, London.

J J Kupperman (1970), Ethical Knowledge, London.

S J Latsis (1976), "A Research Programme in Economics", in Methods and Appraisal in Economics, Spiro Latsis (ed.), Cambridge.

J R Lucas (1970), The Concept or Probability, Oxford.

R D Luce & H Raiffa (1957), Games and Decisions, New York.

J L Mackie (1977), Ethics: Inventing Right and Wrong, London.

J H McCloskey (1957), "An Examination of Restricted Utilitarianism", Philosophical Review, 66, pp. 466-485.

H J McCloskey (1963), "A Note on Utilitarian Punishment", Mind, p. 599.

H J McCloskey (1965), "A Non-Utilitarian Approach to Punishment", Inquiry, 8, pp. 249-263.

B. Mayo (1957), "Varieties of Imperative", Aristotelian Society Proceedings, (Supp.), pp. 161-174.

J Milnor (1954), "Games Against Nature", in Decision Processes, R M Thrall, C H Coombs and R L Davis (eds.), New York.

G W Mortimore (1976), "Rational Action", in Rationality and the Social Sciences, S I Benn and G W Mortimore (eds.), London.

F Mosteller & P Nogee (1951), "An Experimental Measure of Utility", Journal of Political Economy, 59, pp. 279-304.

J Otten (1977), "Reviving the Logical Connection Argument", Canadian Journal of Philosophy, 7, pp. 725-743.

P Pettit (1978), "Rational Man Theory", in Action and Interpretation, C Hookway and P Pettit (eds.), Cambridge.

K R Popper (1968), The Logic of Scientific Discovery, London.

K R Popper (1973), Objective Knowledge, Oxford.

H Raiffa (1970), Decision Analysis, Massachusetts.

A Rapoport (1970), N-Person Game Theory, Ann Arbor.

J Rawls (1972), A Theory of Justice, Oxford.

J Rawls (1974), "Some Reasons for the Maximin Criterion", American Economic Review, 64, pp.141-146.

N Rescher (1975), Unselfishness, Pittsburgh.

R Rorty (1965), "Mind-Body Identity, Privacy, and Categories", Review of Metaphysics, pp. 24-54.

W D Ross (1930), The Right and The Good, Oxford.

A K Sen (1979), "Rational Fools: A Critique of the Behavioural Foundations of Econimic Theory", in Philosophy and Economic Theory, F Hahn and M Hollis (eds.), Oxford.

P Singer (1972), "Is act-utilitarianism self-defeating?", Philosophical Review, 81, pp. 94-104.

J J C Smart (1965), "The Methods of Ethics and the Methods of Science", Journal of Philosophy, 62, pp. 344-349.

J J C Smart (1978), Utilitarianism For and Against, Cambridge, (with Bernard Williams).

T L S Sprigge (1965), "A Utilitarian Reply to Dr McCloskey", Inquiry, 8, pp. 264-291.

C Tisdell (1968), The Theory of Price Uncertainty, Production, and Profit, New Jersey.

A Tversky (1969), "The Intransitivity of Preferences", Psychological Review, 76, pp. 31-48.

A Tversky (1975), "A Critique of Expected Utility Theory: Descriptive and Normative Considerations", Erkenntnis, 9, pp. 163-173.

J W N Watkins (1963), "Negative Utilitarianism", Aristotelian Society Proceedings, (Supp.), pp. 95-114.

J W N Watkins (1970), "Imperfect Rationality", in Explanation in the Behavioural Sciences, R Borger and F Cioffi (eds.), Cambridge.