Whichever way we put it, whether in terms of what I am prepared to prescribe or permit universally (and therefore also for when I am the victim) or in terms of how to be fair as between the interests of all the affected parties, I conclude that the acts I have listed will come out wrong on the act-utilitarian calculation, because of the harms done to the interests of those who are cheated, or the non-fulfilment of prescriptions to which, we may assume, they attach high importance. If we add to this move the preceding one which rules out fantastic cases, and are clear about the distinction between judgements about the character of the agent, judgements about the moral rationality of the action, and judgements about its moral rightness as shown by the outcome, I think that this form of utilitarianism can answer the objections I have mentioned. Much more needs to be said; the present paper is only a beginning, and is not very original.[25] I publish it only to give some indication of the way in which ethical theory can help with normative moral questions, and to try to get the discussion of utilitarianism centred round credible forms of it, rather than forms which we all know will not do.

[25] Among many others from whose ideas I have learnt, I should like in particular to mention Dr Lynda Sharp (Mrs Lynda Paine), in whose thesis 'Forms and Criticisms of Utilitarianism' (deposited in the Bodleian Library at Oxford) some of the above topics are discussed in greater detail.

2 Morality and the theory of rational behaviour

JOHN C. HARSANYI

1 Historical background

The ethical theory I am going to describe in this paper is based on three different time-honoured intellectual traditions in moral philosophy. It also makes essential use of a great intellectual accomplishment of much more recent origin, namely, the modern Bayesian theory of rational behaviour under risk and uncertainty.

One of the three moral traditions I am indebted to goes back to Adam Smith, who equated the moral point of view with that of an impartial but sympathetic spectator (or observer).[1] In any social situation, each participant will tend to look at the various issues from his own self-centred, often emotionally biassed, and possibly quite one-sided, partisan point of view. In contrast, if anybody wants to assess the situation from a *moral* point of view in terms of some standard of justice and equity, this will essentially amount to looking at it from the standpoint of an impartial but humane and sympathetic observer. It may be interesting to note that modern psychological studies on the development of moral ideas in children have come up with a very similar model of moral value judgements.[2]

Another intellectual tradition I have benefited from is Kant's. Kant claimed that moral rules can be distinguished from other behavioural rules by certain formal criteria and, in particular, by the criterion of universality (which may also be described as a criterion of reciprocity).[3] For example, if I really believe that other people should repay me any money they have borrowed from me, then I must admit that *I* am under a similar moral obligation to repay any money I have borrowed from other persons under comparable circumstances. Thus, in ethical content, Kant's principle of universality says much the same thing as the golden rule of the Bible: 'Treat other people in the same way as you want to be treated by them.' Among contemporary authors, the Oxford moral philosopher Hare has

Reprinted from *Social Research*, Winter 1977, vol. 44, no. 4.

[1] Adam Smith 1976.

[2] See, for example Piaget 1962.

[3] Immanuel Kant 1785.

advocated a moral theory based specifically on the Kantian universality requirement (which he calls the 'universalisation' requirement).[4]

My greatest intellectual debt, however, goes to the utilitarian tradition of Bentham, John Stuart Mill, Sidgwick, and Edgeworth, which made maximisation of social utility the basic criterion of morality – social utility being defined either as the sum, or the arithmetic mean, of the utility levels of all individuals in the society.[5] (What these classical utilitarians called 'social utility' is often called the 'social welfare function' in modern welfare economics. But in many cases the term 'social welfare function' is now used in a less specific sense, without any utilitarian connotations.)

Though many details of the classical utilitarian position may be unacceptable to us today, we must not forget what basic political and moral principles they were fighting for. Basically, both in politics and in ethics, they fought for reason against mere tradition, dogmatism, and vested interests. In politics, they conceived the revolutionary idea of judging existing social institutions by an impartial rational test, that of social utility, and did not hesitate to announce it in clear and unmistakable terms if they felt that many of these institutions had definitely failed to pass this test. Likewise, in ethics, they proposed to subject all accepted moral rules to tests of rationality and social utility.

Their main opponents in moral philosophy were the intuitionists, who claimed that we can discover the basic moral rules by direct intuition, which, of course, made any rational evaluation of such moral rules both impossible and unnecessary. Apparently, these intuitionist philosophers were not particularly troubled by the well-known empirical fact that people's 'moral intuitions' seem to be highly dependent on accidents of their own upbringing and, more fundamentally, on the accident of being raised in one particular society rather than another. Though there were many notable exceptions, most people raised in a warlike society or a slave-holding society or a caste society always claimed to have the clear 'moral intuition' that the social practices of their society had full moral justification. It was this uncritical acceptance of existing social practices that the utilitarians fought against by their insistence on subjecting all moral beliefs to a rational test.

In our own time, these crude forms of obscurantism in ethics have largely disappeared. But it is still true, it seems to me, that the updated version of classical utilitarianism is the only ethical theory which consistently abides by the principle that moral issues must be decided by rational tests and that moral behaviour itself is a special form of rational behaviour. I think it can be easily shown that all nonutilitarian theories of

[4] Hare 1952.
[5] Bentham 1948; John Stuart Mill 1962; Sidgwick 1962; Edgeworth 1881.

morality, including John Rawls's very influential theory[6] and several others, at one point or another involve some highly irrational moral choices, representing major departures from a rational pursuit of common human and humane interests, which, in my view, is the very essence of morality.

Yet, notwithstanding its very considerable intellectual accomplishments, classical utilitarianism was open to some major objections. The most important step toward resolving most of these objections was taken by Keynes's friend, the Oxford economist Harrod, who was the first to point out the advantages of *rule* utilitarianism over *act* utilitarianism.[7] (But he did not actually use this terminology. The terms 'act utilitarianism' and 'rule utilitarianism' were introduced only by Brandt.[8]) Act utilitarianism is the view that each individual act must be judged directly in terms of the utilitarian criterion. Thus a morally right act is one that, in the situation the actor is actually in, will maximise social utility. In contrast, rule utilitarianism is the view that the utilitarian criterion must be applied, in the first instance, not to individual acts but rather to the basic general rules governing these acts. Thus a morally right act is one that conforms to the correct moral rule applicable to this sort of situation, whereas a correct moral rule is that particular behavioural rule that would maximise social utility if it were followed by everybody in all social situations of this particular type.

I will discuss the moral implications of these two versions of utilitarian theory in section 9. As I will argue, only rule utilitarianism can explain why a society will be better off if people's behaviour is constrained by a network of moral rights and moral obligations which, barring extreme emergencies, must not be violated on grounds of mere social-expediency considerations. Prior to the emergence of rule-utilitarian theory, utilitarians could not convincingly defend themselves against the accusation that they were advocating a super-Machiavellistic morality, which permitted infringement of all individual rights and all institutional obligations in the name of some narrowly defined social utility.

Virtually all the moral content of the ethical theory I am going to propose will come from these three intellectual traditions: Adam Smith's, Kant's, and that of the utilitarian school. Yet it would not have been possible to put all these pieces together into an intellectually satisfactory theory of morality before the emergence, and without an extensive use, of the modern theory of rational behaviour and, in particular, the modern

[6] Rawls 1971. For a detailed critique of Rawls' theory, see Harsanyi 1975a. For a discussion of some other nonutilitarian theories, see Harsanyi 1975c.
[7] Harrod 1936.
[8] Brandt 1959, pp. 369, 380.

theory of rational behaviour under risk and uncertainty, usually described as Bayesian decision theory. The Bayesian concept of rationality is a very crucial ingredient of my theory.

2 Ethics as a branch of the general theory of rational behaviour

I propose to argue that the emergence of modern decision theory has made ethics into an organic part of the general theory of rational behaviour. The concept of rational behaviour (practical rationality) is important in philosophy both in its own right and because of its close connection with theoretical rationality. It plays a very important role also in the empirical social sciences, mainly in economics but also in political science and in sociology (at least in the more analytically oriented versions of these two disciplines). What is more important for our present purposes, the concept of rational behaviour is the very foundation of the normative disciplines of decision theory, of game theory, and (as I will argue) of ethics.

The concept of rational behaviour arises from the empirical fact that human behaviour is to a large extent goal-directed behaviour. Basically, rational behaviour is simply behaviour consistently pursuing some well defined goals, and pursuing them according to some well defined set of preferences or priorities.

We all know that, as a matter of empirical fact, even if human behaviour is usually goal-directed, it is seldom sufficiently consistent in its goals, and in the priorities it assigns to its various goals, to approach the ideal of full rationality. Nevertheless, in many fields of human endeavour – for example, in most areas of economic life, in many areas of politics (including international politics), and in some other areas of social interaction – human behaviour does show sufficiently high degrees of rationality as to give a surprising amount of explanatory and predictive power to some analytical models postulating full rationality. (Of course, it is very possible that we could further increase the explanatory and predictive power of our theories if we paid closer attention to the actual limits of human rationality and information-processing ability, in accordance with Simon's theory of limited rationality.[9])

Moreover, whether people actually do act rationally or not, they are often interested in increasing the rationality of their behaviour; and they are also interested in the conceptual problem of what it would actually mean to act fully rationally in various situations. It is the task of the normative disciplines of decision theory, game theory, and ethics to help

[9] See, for example, Simon 1960.

people to act more rationally and to give them a better understanding of what rationality really is.

For reasons I will describe presently, I propose to consider these three disciplines as parts of the same general theory of rational behaviour. Thus one part of this general theory[10] will be:

(1) The theory of *individual* rational behaviour, which itself comprises the theories of rational behaviour

(1A) Under certainty,

(1B) Under risk (where all probabilities are known objective probabilities), and

(1C) Under uncertainty (where some or all probabilities are unknown, and may be even undefined as objective probabilities).

(1A), (1B), and (1C) together are often called utility theory while (1B) and (1C) together are called decision theory.

The two other branches of the general theory of rational behaviour both deal with rational behaviour in a *social* setting. They are:

(2) Game theory, which is a theory of rational interaction between two or more individuals, each of them rationally pursuing his own objectives against the other individual(s) who rationally pursue(s) his (or their) own objectives. Any individual's objectives may be selfish or unselfish, as determined by his own utility function. (A nontrivial game situation can arise just as easily among altruists as it can among egoists – as long as these altruists are pursuing partly or wholly divergent altruistic goals.)

(3) Ethics, which is a theory of rational behaviour in the service of the common interests of society as a whole.

I think it is useful to regard (1), (2), and (3) as branches of the same basic discipline, for the following reasons:

(i) All three normative disciplines use essentially the same method. Each starts out by defining rational behaviour in its own field either by some set of axioms or by a constructive decision model. In either case, this initial definition may be called the primary definition of rationality in this particular field. Then, from this primary definition, each derives a secondary definition of rationality, which is usually much more convenient than the primary definition in itself would be in its axiomatic or constructive form, both for practical applications and for further philosophical analysis. For example, in case (1A) the secondary definition of rationality is ***utility maximisation*** – which is for many purposes a much more convenient characterisation of rational behaviour under certainty than is its primary definition in terms of the usual axioms (the complete preordering requirement and the continuity axiom).

[10] The remaining part of this section will be somewhat technical, but it can be omitted without loss of continuity.

In cases (1B) and (1C), the secondary definition of rationality is *expected-utility maximisation* (with objective probability weights in case (1B) and with subjective probability weights in case (1C)).

In the game-theoretical case (2), the secondary definition is provided by various game-theoretical solution concepts.

Finally, in the case of ethics (case (3)), as we will see, the secondary definition of rationality (or of morality) is in terms of *maximising the average utility level* of all individuals in the society.

This common method that these normative disciplines use represents a unique combination of philosophical analysis and of mathematical reasoning. In each case, a movement from the primary definition of rationality to its secondary definition is a straightforward mathematical problem. But discovery of an appropriate primary definition is always essentially a philosophical – that is, a conceptual – problem (with the possible exception of case (1A), where the philosophical dimension of the problem seems to be less important). People familiar with research work in these areas know the special difficulties that arise from this unusual interdependence of philosophical and mathematical problems. These are definitely not areas for people who prefer their mathematics without any admixture of philosophy, or who prefer their philosophy without any admixture of mathematics.

(ii) The axioms used by decision theory, game theory, and ethics are mathematically very closely related. In all three disciplines they are based on such mathematical properties as efficiency, symmetry, avoidance of dominated strategies, continuity, utility maximisation, invariance with respect to order-preserving linear utility transformations, etc.

(iii) Yet the most important link among the three disciplines lies in the fact that recent work has made it increasingly practicable to reduce some basic problems of game theory and of ethics partly or wholly to decision-theoretical problems.[11]

3 The equiprobability model for moral value judgements

After the two introductory sections, I now propose to describe my theory of morality. The basis of this theory is a model for moral value judgements.

Any moral value judgement is a judgement of preference, but it is a judgement of preference of a very special kind. Suppose somebody tells us: 'I much prefer our capitalist system over any socialist system because

[11] In game theory, one step in this direction has been a use of probability models for analysing games with incomplete information (Harsanyi 1967–8). More recently, a decision-theoretical approach to defining a solution for noncooperative games has been proposed (Harsanyi 1975b). On uses of decision theory in ethics, see Harsanyi 1977.

under our capitalist system I happen to be a millionaire and have a very satisfying life, whereas under a socialist system I would be in all probability at best a badly paid minor government official.' This may be a very reasonable judgement of personal preference from his own individual point of view. But nobody would call it a *moral* value judgement because it would be obviously a judgement based primarily on self-interest.

Compare this with a situation where somebody would express a preference for the capitalist system as against the socialist system without knowing in advance what particular social position he would occupy under either system. To make it more precise, let us assume that he would choose between the two systems under the assumption that, in either system, he would have the same probability of occupying any one of the available social positions. In this case, we could be sure that his choice would be independent of morally irrelevant selfish considerations. Therefore his choice (or his judgement of preference) between the two systems would now become a genuine moral value judgement.

Of course, it is not really necessary that a person who wants to make a moral assessment of the relative merits of capitalism and of socialism should be literally ignorant of the actual social position that he does occupy or would occupy under each system. But it *is* necessary that he should at least try his best to disregard this morally irrelevant piece of information when he is making his moral assessment. Otherwise his assessment will not be a genuine moral value judgement but rather will be merely a judgement of personal preference.

For short reference, the fictitious assumption of having the same probability of occupying any possible social position will be called the *equiprobability postulate*, whereas the entire preceding decision model based on this assumption will be called the *equiprobability model of moral value judgements*.

We can better understand the implications of this model if we subject it to decision-theoretical analysis. Suppose the society we are considering consists of n individuals, numbered as individual $1, 2, \ldots, n$, according to whether they would occupy the 1st (highest), 2nd (second highest), . . . , nth (lowest) social position under a given social system. Let $U_1, U_2, \ldots, U_n$, denote the utility levels that individuals $1, 2, \ldots, n$ would enjoy under this system. The individual who wants to make a moral value judgement about the relative merits of capitalism and of socialism will be called individual i. By the equiprobability postulate, individual i will act in such a way as if he assigned the *same* probability $1/n$ to his occupying *any* particular social position and, therefore, to his utility reaching any one of the utility levels $U_1, U_2, \ldots, U_n$.

Now, under the assumed conditions, according to Bayesian decision

theory, a rational individual will always choose that particular social system that would maximise his expected utility, that is, the quantity

$$(1) \quad W_i = \frac{1}{n} \sum_{j=1}^{n} U_j,$$

representing the arithmetic mean of all individual utility levels in society. We can express this conclusion also by saying that a rational individual will always use this mean utility as his social welfare function; or that he will be a utilitarian, who defines social utility as the mean of individual utilities (rather than as their sum, as many utilitarians have done).[12]

Of course, this conclusion makes sense only if we assume that it is mathematically admissible to *add* the utilities of different individuals, that is, if we assume that interpersonal comparisons of utility represent a meaningful intellectual operation. I will try to show that this is in fact the case.

In describing this equiprobability model, I have assumed that individual i, who is making a moral value judgement on the merits of the two alternative social systems, is one of the n members of the society in question. But exactly the same reasoning would apply if he were an interested outsider rather than a member. Indeed, for some purposes it is often heuristically preferable to restate the model under this alternative assumption. Yet, once we do this, our model becomes a modern restatement of Adam Smith's theory of an impartially sympathetic observer. His impartiality requirement corresponds to my equiprobability postulate, whereas his sympathy requirement corresponds to my assumption that individual i will make his choice in terms of interpersonal utility comparisons based on empathy with various individual members of society (see section 5).

This equiprobability model of moral value judgements gives us both a powerful analytical criterion and a very convenient heuristic criterion for deciding practical moral problems. If we want to decide between two alternative moral standards A and B, all we have to do is ask ourselves the question, 'Would I prefer to live in a society conforming to standard A or in a society conforming to standard B? – assuming I would not know in advance what my actual social position would be in either society but rather would have to assume to have an equal chance of ending up in any one of the possible positions.'

[12] For most purposes the two definitions of social utility are mathematically equivalent. This is always true when n, the number of people in society, can be regarded as a constant. The two definitions, however, yield different decision criteria in judging alternative population policies. In this latter case, in my view, the mean-utility criterion gives incomparably superior results.

Admittedly, this criterion – or any conceivable moral criterion – will still leave each of us with the great moral responsibility, and the often very difficult intellectual task, of actually choosing between these two alternative moral standards in terms of this criterion. But by using this criterion we will know at least *what* the actual intellectual problem is that we are trying to solve in choosing between them.

My equiprobability model was first published in 1953, and was extended in 1955.[13] Vickrey had suggested a similar idea,[14] but my work was independent of his. Later John Rawls again independently proposed a very similar model, which he called the 'original position', based on the 'veil of ignorance'.[15] But while my own model served as a basis for a utilitarian theory, Rawls derived very nonutilitarian conclusions from his own. Yet the difference does not lie in the nature of the two models, which are based on almost identical qualitative assumptions. Rather, the difference lies in the decision-theoretical analysis applied to the two models. One difference is that Rawls avoids any use of numerical probabilities. But the main difference is that Rawls makes the technical mistake of basing his analysis on a highly irrational decision rule, the maximin principle, which was fairly fashionable thirty years ago but which lost its attraction a few years later when its absurd practical implications were realised.[16]

Our model of moral value judgements can also be described as follows. Each individual has two very different sets of preferences. On the one hand, he has his *personal preferences*, which guide his everyday behaviour and which are expressed in his utility function U_i. Most people's personal preferences will not be completely selfish. But they will assign higher weights to their own interests and to the interests of their family, their friends, and other personal associates than they will assign to the interests of complete strangers. On the other hand, each individual will also have *moral preferences* which may or may not have much influence on his everyday behaviour but which will guide his thinking in those – possibly very rare – moments when he forces a special impersonal and impartial attitude, that is, a moral attitude, upon himself. His moral preferences, unlike his personal preferences, will by definition always assign the same weight to all individuals' interests, including his own. These moral preferences will be expressed by his social-welfare function W_i. Typically, different individuals will have very different utility functions U_i but, as can be seen from Equation (1) above, in theory they will tend to have identical social-welfare functions – but only if they agree in their factual assump-

[13] Harsanyi 1953 and 1955.
[14] Vickrey 1945.
[15] Rawls 1957; 1958; and 1971.
[16] First by Radner and Marschak 1954, pp. 61–8. See also Harsanyi 1975a.

tions on the nature of the individual utility functions U_i and on the conversion ratios between different individuals' utilities (as decided by interpersonal utility comparisons) – which, of course, may not be the case.

By definition, a moral value judgement is always an expression of one's moral preferences. Any evaluative statement one may make will automatically lose its status of a moral value judgement if it is unduly influenced by one's personal interests and personal preferences.

4 An axiomatic justification for utilitarian theory

I now propose to present an alternative, this time *axiomatic*, justification for utilitarian theory. This axiomatic approach yields a lesser amount of philosophically interesting information about the nature of morality than the equiprobability model does, but it has the advantage of being based on much weaker – almost trivial – philosophical assumptions. Instead of using very specific philosophical assumptions about the nature of morality, it relies merely on Pareto optimality and on the Bayesian rationality postulates.

We need three axioms:

Axiom 1: Individual rationality. The personal preferences of all n individuals in society satisfy the Bayesian rationality postulates.[17]

Axiom 2: Rationality of moral preferences. The moral preferences of at least one individual, namely, individual i, satisfy the Bayesian rationality postulates.

Axiom 3: Pareto optimality. Suppose that at least one individual j ($j = 1, \ldots, n$) personally prefers alternative A to alternative B, and that no individual has an opposite personal preference. Then individual i will morally prefer alternative A over alternative B.

Axiom 3 is a very weak and hardly objectionable moral postulate. Axiom 1 is a rather natural rationality requirement. Axiom 2 is an equally natural rationality requirement: in trying to decide what the common interests of society are, we should surely follow at least as high standards of rationality as we follow (by Axiom 1) in looking after our own personal interests.

[17] Most philosophers and social scientists do not realise how weak the rationality postulates are that Bayesian decision theory needs for establishing the expected-utility maximisation theorem. As Anscombe and Aumann have shown (Anscombe and Aumann 1963), all we need is the requirement of consistent preferences (complete preordering), a continuity axiom, the sure-thing principle (avoidance of dominated strategies), and the requirement that our preferences for lotteries should depend only on the possible prizes and on the specific random events deciding the actual prize. (The last requirement can be replaced by appropriate axioms specifying the behaviour of numerical probabilities within lotteries. In the literature, these axioms are usually called 'notational conventions'.)

Axiom 1 implies that the personal preferences of each individual j ($j = 1, \ldots, n$) can be represented by a von Neumann–Morgenstern (=vNM) utility function U_j. Axiom 2 implies that the moral preferences of individual i can be represented by a social welfare function W_i, which mathematically also has the nature of a vNM utility function. Finally, the three axioms together imply the following theorem:

Theorem T. The social welfare function W_i of individual i must be of the mathematical form:

$$(2) \quad W_i = \sum_{j=1}^{n} a_j U_j \text{ with } a_j > 0 \text{ for } j = 1, \ldots, n.$$

This result[18] can be strengthened by adding a fourth axiom:

Axiom 4: Symmetry. The social-welfare function W_i is a symmetric function of all individual utilities. (That is, different individuals should be treated equally.)

Using this axiom, we can conclude that

$$(3) \quad a_1 = \ldots = a_n > 0.$$

Equations (2) and (3) together are essentially equivalent to Equation(1).[19]

I realise that some people may feel uncomfortable with the rather abstract philosophical arguments I used to justify my equiprobability model. In contrast, the four axioms of the present section make only very weak philosophical assumptions. They should appeal to everbody who believes in Bayesian rationality, in Pareto optimality, and in equal treatment of all individuals. Yet these very weak axioms turn out to be sufficient to entail a utilitarian theory of morality.

5 Interpersonal utility comparisons

In everyday life we make, or at least attempt to make, interpersonal utility comparisons all the time. When we have only one nut left at the end of a trip, we may have to decide which particular member of our family is in greatest need of a little extra food. Again, we may give a book or a concert ticket or a free invitation to a wine-tasting fair to one friend rather than to another in the belief that the former would enjoy it more than the latter would. I do not think it is the task of a philosopher or a social scientist to

[18] For proof, see Harsanyi 1955.

[19] There is, however, the following difference. Equation (1) implies that social utility must be defined as the mean of individual utilities rather than as their sum. In contrast, Equations (2) and (3) do not favour either definition of social utility over its alternative.

deny the obvious fact that people often feel quite capable of making such comparisons. Rather, his task is to explain how we ever managed to make such comparisons – as well or as badly as we do make them.

Simple reflection will show that the basic intellectual operation in such interpersonal comparisons is imaginative empathy. We imagine ourselves to be in the shoes of another person, and ask ourselves the question, 'If I were now really in *his* position, and had *his* taste, *his* education, *his* social background, *his* cultural values, and *his* psychological make-up, then what would now be *my* preferences between various alternatives, and how much satisfaction or dissatisfaction would *I* derive from any given alternative?' (An 'alternative' here stands for a given bundle of economic commodities plus a given position with respect to various noneconomic variables, such as health, social status, job situation, family situation, etc.)

In other words, any interpersonal utility comparison is based on what I will call the *similarity postulate*, to be defined as the assumption that, once proper allowances have been made for the empirically given differences in taste, education, etc., between me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given alternative will be otherwise much the same. Of course, it is only too easy to misapply this similarity postulate. For instance, I may fail to make proper allowances for differences in our tastes, and may try to judge the satisfaction that a devoted fish eater derives from eating fish in terms of my own intense dislike for any kind of sea food. Of course, sensible people will seldom make such an obvious mistake. But they may sometimes make much subtler mistakes of the same fundamental type.

In general, if we have enough information about a given person, and make a real effort to attain an imaginative empathy with him, we can probably make reasonably good estimates of the utilities and disutilities he would obtain from various alternatives. But if we have little information about him, our estimates may be quite wrong.

In any case, utilitarian theory does not involve the assumption that people are very good at making interpersonal utility comparisons. It involves only the assumption that, in many cases, people simply *have* to make such comparisons in order to make certain moral decisions – however badly they may make them. If I am trying to decide which member of my family is in greatest need of food, I may sometimes badly misjudge the situation. But I simply *have* to make *some* decision. I cannot let *all* members of my family go hungry because I have philosophical scruples about interpersonal comparisons and cannot make up my mind.

Nevertheless, interpersonal utility comparisons do pose important philosophical problems. In particular, they pose the problem that they require us to use what I have called the similarity postulate. Yet this

postulate, by its very nature, is not open to any direct empirical test. I may very well assume that different people will have similar psychological feelings about any given situation, once differences in their tastes, educations, etc. have been allowed for. But I can never verify this assumption by direct observation since I have no direct access to their inner feelings.

Therefore, the similarity postulate must be classified as a nonempirical a priori postulate. But, of course, interpersonal utility comparisons are by no means unique among empirical hypotheses in their dependence on such nonempirical postulates. In actual fact, whenever we choose among alternative empirical hypotheses, we are always dependent on some nonempirical choice criteria. This is so because the empirical facts are always consistent with infinitely many alternative hypotheses, and the only way we can choose among them is by using a priori nonempirical choice criteria, such as simplicity, parsimony, preference for the 'least arbitrary' hypothesis, etc.

Our similarity postulate is a nonempirical postulate of the same general type. Its intuitive justification is that, if two individuals show exactly identical behaviour – or, if they show different behaviour but these differences in their observable behaviour have been properly allowed for – then it will be a completely arbitrary and unwarranted assumption to postulate some further hidden and unobservable differences in their psychological feelings.

We use this similarity postulate not only in making interpersonal utility comparisons but also in assigning other people human feelings and conscious experiences at all. From a purely empirical point of view, a world in which I would be the only person with real conscious experiences while all other people were mindless robots would be completely indistinguishable from our actual world where all individuals with human bodies are conscious human beings. (Indeed, even a world in which I alone would exist, and all other people as well as the whole physical universe would be merely my own dream – solipsism – would be empirically indistinguishable from the world we actually live in.) When we choose the assumption that we actually live in a world populated by millions of other human beings, just as real and just as conscious as we are ourselves, then we are relying on the same similarity postulate. We are essentially saying that, given the great basic similarity among different human beings, it would be absurd to postulate fundamental hidden differences between them by making one person a conscious human being while making the others mere robots, or by making one person real while making the others mere dream figures. (Strictly speaking, we cannot exclude the possibility that somebody who looks human will turn out to be an unfeeling robot; but we

have no scientific or moral justification to treat him like a robot before the evidence for his being a robot becomes overwhelming.)

There is no logical justification for using the similarity postulate to reject the hypothesis that other people are mere robots (or mere dream figures) yet to resist interpersonal utility comparisons based on the very same similarity postulate. It is simply illogical to admit that other people do have feelings and, therefore, do derive *some* satisfaction from a good meal in the same way we do; yet to resist the quantitative hypothesis that the *amount* of satisfaction they actually obtain from a good dinner – that is, the personal importance they attach to a good dinner – must be much the same as it is in our own case, after proper allowances have been made for differences in our tastes, in the food requirements of our bodies, in our state of health, etc. A willingness to make interpersonal comparisons is no more than an admission that other people are just as real as we are, that they share a common humanity with us, and that they have the same basic capacity for satisfaction and for dissatisfaction, in spite of the undeniable individual differences that exist between us in specific detail.

The long-standing opposition by many philosophers and social scientists to interpersonal utility comparisons goes back to the early days of logical positivism, when the role of nonempirical a priori principles, like the similarity postulate, in a choice among alternative empirical hypotheses was very poorly understood. We owe an immense intellectual debt to the logical positivists for their persistent efforts to put philosophy on truly scientific foundations by combining strict empiricism with the strict mathematical rigour of modern logic. But there is no denying that many of their specific philosophical views were badly mistaken, and that they had little appreciation in their early period for the importance of a priori principles and, more generally, for the importance of theoretical ideas in empirical science.

One would think that after so many years the time had come to escape the narrow confines of a long-obsolete logical-positivist orthodoxy and to have a fresh look at the problem of interpersonal utility comparisons.

6 The use of von Neumann–Morgenstern utility functions

The utilitarian theory I have proposed makes an essential use of von Neumann–Morgenstern (= vNM) utility functions. Many critics have argued that any use of vNM utility functions is inappropriate, because they merely express people's attitudes toward gambling, and these attitudes have no moral significance.[20] This objection is based on a rather common misinterpretation of vNM utility functions. These utility functions do express people's attitudes to risk taking (in gambling, buying insurance, investing and other similar activities). But they do not merely register these attitudes; rather, they try to explain them in terms of the relative importance (relative utility) people attach to possible gains and to possible losses of money or of other economic or noneconomic assets.

[20] See, for example, Rawls 1971, pp. 172, 323.

For example, suppose that Mr *X* is willing to pay \$5 for a lottery ticket that gives him a 1/1,000 chance of winning \$1,000. Then we can explain his willingness to gamble at such very unfavourable odds as follows. He must have an unusually high utility for winning \$1,000, as compared with his disutility for losing \$5. In fact, even though the ratio of these two money amounts is only 1,000 : 5 = 200 : 1, the ratio of the corresponding utility and disutility must be at least 1,000 : 1. (If we know Mr *X*'s personal circumstances, then we will often be able to carry this explanation one step – or several steps – further. For instance, we may know that his strong desire for winning \$1,000 arises from the fact that he needs the money for a deposit on a badly needed car, or for some other very important large and indivisible expenditure; while his relative unconcern about losing \$5 is due to the fact that such a loss would not seriously endanger his ability to pay for his basic necessities – food, lodging, etc.[21])

In other words, even though a person's vNM utility function is always estimated in terms of his behaviour under risk and uncertainty, the real purpose of this estimation procedure is to obtain cardinal-utility measures for the relative personal importance he assigns to various economic (and noneconomic) alternatives.

No doubt, since social utility is defined in terms of people's vNM utility functions, our utilitarian theory will tend to assign higher social priorities to those individual desires for which people are willing to take considerable risks in order to satisfy them. But this is surely as it should be. Other things being equal, we *should* give higher social priorities to intensely felt human desires; and one indication that somebody feels strongly about a particular desired objective is his willingness to take sizable risks to attain it. For example, if a person is known to have risked his life in order to obtain a university education (e.g., by escaping from a despotic government which had tried to exclude him from all higher education), then we can take this as a reasonably sure sign of his attaching very high personal importance (very high utility) to such an education; and I cannot see anything wrong with our assigning high social priority to helping him to such an education on the basis of this kind of evidence.

[21] Fundamentally, any explanation of why a given person's vNM utility function has any specific shape and, in particular, why its convex and concave segments are distributed in the way they are, will be typically in terms of substitution and complementarity relations among the commodities consumed by him. Mathematically, indivisible commodities are a special case of complementarity.

7 Preference utilitarianism, hedonism, ideal utilitarianism, and the question of irrational preferences

The utilitarian theory I have proposed defines social utility in terms of individual utilities, and defines each person's utility function in terms of his personal preferences. Thus, in the end, social utility is defined in terms of people's personal preferences. This approach may be called *preference utilitarianism*. It is not the same approach that was used by the nineteenth-century utilitarians. They were *hedonists* (hedonistic utilitarians), and defined both social utility and individual utility functions in terms of feelings of pleasure and pain. A third approach, called *ideal utilitarianism*, was proposed by the Cambridge philosopher Moore, who defined both social utility and individual utilities in terms of amounts of 'mental states of intrinsic worth', such as the mental states involved in philosophy, science, aesthetic appreciation of works of art, experiences of personal friendship, etc.[22]

Both hedonistic and ideal utilitarianism are open to serious objections. The former presupposes a now completely outdated hedonistic psychology. It is by no means obvious that all we do we do only in order to attain pleasure and to avoid pain. It is at least arguable that in many cases we are more interested in achieving some objective state of affairs than we are interested in our own subjective feelings of pleasure and pain that may result from achieving it. It seems that when I give a friend a present my main purpose is to give *him* pleasure rather than to give pleasure to myself (though this may very well be a secondary objective). Even if I want to accomplish something for myself, it is by no means self-evident that my main purpose is to produce some feelings of pleasure in myself, and it is not the actual accomplishment of some objective condition, such as having a good job, solving a problem, or winning a game, etc. In any case, there is no reason whatever why any theory of morality should try to prejudge the issue whether people are always after pleasure or whether they also have other objectives.

As to ideal utilitarianism, it is certainly not true as an empirical observation that people's only purpose in life is to have 'mental states of intrinsic worth'. But if this is not in fact the case, then it is hard to see how we could prove that, even though they may not in fact act in this way, this is how they *should* act. Moreover, the criteria by which 'mental states of intrinsic worth' can be distinguished from other kinds of mental states are extremely unclear. (Moore's own theory that they differ from other mental states in having some special 'nonnatural qualities' is a very unconvincing

[22] Moore 1903.

old-fashioned metaphysical assumption lacking any kind of supporting evidence.)

More fundamentally, preference utilitarianism is the only form of utilitarianism consistent with the important philosophical principle of *preference autonomy*. By this I mean the principle that, in deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own wants and his own preferences. To be sure, as I will myself argue below, a person may irrationally want something which is very 'bad for him'. But, it seems to me, the only way we can make sense of such a statement is to interpret it as a claim to the effect that, in some appropriate sense, his own preferences at some deeper level are inconsistent with what he is now trying to achieve.

Any sensible ethical theory must make a distinction between rational wants and irrational wants, or between rational preferences and irrational preferences. It would be absurd to assert that we have the same moral obligation to help other people in satisfying their utterly unreasonable wants as we have to help them in satisfying their very reasonable desires. Hedonistic utilitarianism and ideal utilitarianism have no difficulty in maintaining this distinction. They can define rational wants simply as ones directed toward objects having a real ability to produce pleasure, or a real ability to produce 'mental states of intrinsic worth'; and they can define irrational wants as ones directed toward objects lacking this ability. But it may appear that this distinction is lost as soon as hedonistic and ideal utilitarianism are replaced by preference utilitarianism.

In actual fact, there is no difficulty in maintaining this distinction even without an appeal to any other standard than an individual's own personal preferences. All we have to do is to distinguish between a person's manifest preferences and his true preferences. His manifest preferences are his actual preferences as manifested by his observed behaviour, including preferences possibly based on erroneous factual beliefs, or on careless logical analysis, or on strong emotions that at the moment greatly hinder rational choice. In contrast, a person's true preferences are the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice. Given this distinction, a person's rational wants are those consistent with his true preferences and, therefore, consistent with all the relevant factual information and with the best possible logical analysis of this information, whereas irrational wants are those that fail this test.

In my opinion, social utility must be defined in terms of people's true preferences rather than in terms of their manifest preferences. But, while it is only natural to appeal from a person's irrational preferences to his

underlying 'true' preferences, we must always use his own preferences in some suitable way as our final criterion in judging what his real interests are and what is really good for him.

8 Exclusion of antisocial preferences

I have argued that, in defining the concept of social utility, people's irrational preferences must be replaced by what I have called their true preferences. But I think we have to go even further than this: some preferences, which may very well be their 'true' preferences under my definition, must be altogether excluded from our social-utility function. In particular, we must exclude all clearly antisocial preferences, such as sadism, envy, resentment, and malice.[23]

According to utilitarian theory, the fundamental basis of all our moral commitments to other people is a general goodwill and human sympathy. But no amount of goodwill to individual *X* can impose the moral obligation on me to help him in hurting a third person, individual *Y*, out of sheer sadism, ill will, or malice. Utilitarian ethics makes all of us members of the same moral community. A person displaying ill will toward others does remain a remember of this community, but not with his whole personality. That part of his personality that harbours these hostile antisocial feelings must be excluded from membership, and has no claim for a hearing when it comes to defining our concept of social utility.[24]

9 Rule utilitarianism vs. act utilitarianism

Just as in making other moral decisions, in choosing between rule utilitarianism and act utilitarianism the basic question we have to ask is this: Which version of utilitarianism will maximise social utility? Will society be better off under one or the other? This test very clearly gives the advantage to rule utilitarianism.

In an earlier paper[25] I proposed the following decision-theoretical model for studying the moral implications of the two utilitarian theories. The

[23] For a contrary view, see Smart 1961, pp. 16–17.

[24] The German neo-Kantian utilitarian philosopher Leonard Nelson proposed a distinction between legitimate and illegitimate personal interests (Nelson 1917–32). He argued that the only interests we are morally obliged to respect are legitimate interests. Thus, under his theory, exclusion of antisocial preferences from our concept of social utility is merely a special case of the general principle of disregarding all illegitimate interests. Unfortunately, Nelson did not offer any clear formal criterion for defining legitimate and illegitimate interests. But it seems to me that a really satisfactory theory of legitimate and illegitimate interests would be a major step forward in utilitarian moral philosophy. Yet discussion of this problem must be left for another occasion. (The reference to Nelson's work I owe to Reinhard Selten.)

[25] Harsanyi 1977.

problem we want to consider is that of making moral decisions, that is, the problem of deciding what the morally right action is in a given situation or in a given class of situations. In actual fact, analytically it is preferable to redefine this problem as one of choosing a morally right strategy. Here the term 'strategy' has its usual decision-theoretical and game-theoretical meaning. Thus a strategy is a mathematical function assigning a specific action to any possible situation, subject to the requirement that, if the agent has insufficient information to distinguish one situation from another, then any strategy of his must assign the same specific action to both situations. (In technical language, all choice points belonging to the same information set must have the same specific action assigned to them.)

The two utilitarian theories use different decision rules in solving this moral decision problem. For both theories, a moral decision problem is a maximisation problem involving maximisation of the same quantity, namely, social utility. But the two theories impose very different mathematical constraints on this maximisation problem. An act-utilitarian moral agent assumes that the strategies of all other moral agents (including those of all other utilitarian agents) are given and that his task is merely to choose his own strategy so as to maximise social utility when all other strategies are kept constant. In contrast, a rule-utilitarian moral agent will regard not only his own strategy but also the strategies of all other rule-utilitarian agents as variables to be determined during the maximisation process so as to maximise social utility. For him this maximisation process is subject to two mathematical constraints: one is that the strategies to be chosen for all rule-utilitarian agents must be identical (since, by the definition of rule utilitarianism, all rule-utilitarian agents are always required to follow the same general moral rules); the other is that the strategies of all nonutilitarian agents must be regarded as given. (On this last point both utilitarian theories agree: people known not to believe in a utilitarian philosophy cannot be expected to choose their strategies by trying to maximise social utility. They may follow traditional morality, or some other nonutilitarian morality, or may simply follow self-interest, etc. But, in any case, for the purposes of a utilitarian decision problem, their strategies must be regarded as being given from outside of the system.)

These differences in the decision rules used by the two utilitarian theories, and in particular the different ways they define the constraints for the utilitarian maximisation problem, have important practical implications. One implication is that rule utilitarianism is in a much better position to organise cooperation and strategy coordination among different people (coordination effect).

For example, consider the problem of voting when there is an important measure on the ballot but when voting involves some minor incon-

venience. Suppose there are 1,000 voters strongly favouring the measure, but it can be predicted with reasonable certainty that there will also be 800 negative votes. The measure will pass if it obtains a simple majority of all votes cast. How will the two utilitarian theories handle this problem?

First suppose that all 1,000 voters favouring the measure are act utilitarians. Then each of them will take the trouble to vote only if he thinks that his own vote will be decisive in securing passage of the measure, that is, if he expects *exactly* 800 other people favouring the measure to vote (since in this case his own vote will be needed to provide the 801 votes required for majority). But of course each voter will know that it is extremely unlikely that his own vote will be decisive in this sense. Therefore, most act-utilitarian voters will not bother to vote, and the measure will fail, possibly with disastrous consequences for their society.

In contrast, if the 1,000 voters favouring the measure are rule utilitarians, then all of them will vote (if mixed strategies are not permitted). This is so because the rule-utilitarian decision rule will allow them a choice only between two admissible strategies: one requiring everybody to vote and the other requiring nobody to vote. Since the former will yield a higher social utility, the strategy chosen by the rule-utilitarian criterion will be for everybody to vote. As this example shows, by following the rule-utilitarian decision rule people can achieve successful spontaneous cooperation in situations where this could not be done by adherence to the act-utilitarian decision rule (or at least where this could not be done without explicit agreement on coordinated action, and perhaps without an expensive organisation effort).

Though in some situations this coordination effect may be quite important, it seems to me that the main advantage of rule utilitarianism over act utilitarianism really lies in a different direction, namely, in its ability to take proper account of the implications that alternative systems of possible moral rules would have for people's expectations and incentives (expectation and incentive effects).

For example, consider the problem of keeping promises. Traditional morality says that promises should be kept, with a possible exception of cases where keeping a promise would impose excessive hardship on the promise maker (or perhaps on third persons). In contrast, act utilitarianism would make the breaking of a promise morally permissible whenever this would yield a slightly higher social utility – perhaps because of unexpected changes in the circumstances – than keeping of the promise would yield. But this would greatly reduce the social benefits associated with the making of promises as an institution. It would make it rather uncertain in most cases whether any given promise would be kept. People would be less able to form definite expectations about each other's future

behaviour and would have a general feeling of insecurity about the future. Moreover, this uncertainty would greatly reduce their incentives to engage in various socially very useful activities on the expectation that promises given to them would be kept. (For instance, they would become much less willing to perform useful services for other people for promised future rewards.)

As compared with act utilitarianism, rule utilitarianism will be much closer to traditional morality in maintaining that promises should be kept, subject only to rather rare exceptions. An act utilitarian always asks the question, 'Would this one act of possible promise breaking increase or decrease social utility?' In contrast, a rule utilitarian has to ask, 'What particular moral rule for promise keeping would maximise social utility?'

As a result, an act utilitarian can consider the socially unfavourable effects of promise breaking only to the extent that these have the nature of causal consequences of individual acts of promise breaking. No doubt, one act of promise breaking already will somewhat reduce people's trust in promises, but normally this effect will be quite small. In contrast, a rule utilitarian can also consider the causal consequences of a general practice of repeated promise breaking. But, more importantly, he can also consider the noncausal logical implications of adopting a moral rule permitting many easy exceptions to promise keeping.

More particularly, he will always have to ask the question, 'What would be the social implications of adopting a moral rule permitting that promises should be broken under conditions *A*, *B*, *C*, etc. – assuming that all members of the society would know[26] that promise breaking would be permitted under these conditions?' Thus he will always have to balance the possible direct benefits of promise breaking in some specific situations against the unfavourable expectation and incentive effects that would arise if people knew in advance that in these situations promises would not be kept. In other words, rule utilitarianism not only enables us to make a rational choice among alternative possible general rules for defining morally desirable behaviour. Rather, it also provides a rational test for determining the exceptions to be permitted from these rules.

Limitations of space do not allow me to discuss the moral implications of rule utilitarianism here in any greater detail.[27] It can be shown, however, that rule utilitarianism comes fairly close to traditional morality in recognising the importance of social institutions which establish a network of

[26] In trying to evaluate a possible moral rule from a rule-utilitarian point of view, we must always assume that everybody would know the content of this moral rule. This is so because in principle everybody can always find out by direct computation what particular set of moral rules (i.e., what particular moral strategy) is optimal in terms of the rule-utilitarian criterion.

[27] But see Harsanyi 1977.

moral rights and of moral obligations among different people in society; and in maintaining that these rights and obligations must not be infringed upon on grounds of immediate social utility, with the possible exception of some very rare and very special cases. (The main social advantages of such stable rights and stable obligations, once more, lie in their beneficial expectation and incentive effects.) But of course we cannot expect that the rule-utilitarian criterion would confirm traditional views on these matters in all particulars.

10 The utility of free personal choice

As Rawls has rightly pointed out,[28] traditional utilitarianism tries to impose unreasonably strict moral standards on us, because it requires us to choose every individual action of ours so as to maximise social utility. Thus, if I feel like reading a book for entertainment, I must always ask myself whether my time could not be more usefully devoted to looking after the poor, or to converting some as yet unconverted colleagues to utilitarianism, or to taking part in some other socially beneficial project, etc. The only ways I could possibly justify my taking out time for reading this book would be to argue that reading it would give me exceptionally high direct utility (so as to exceed any social utility I could possibly produce by alternative activities), or that my reading the book would have great instrumental utility – for example, by restoring my temporarily depleted mental and physical energy for future socially very beneficial activities.

There is obviously something wrong with this moral choice criterion. It is not hard to see where the problem lies. Any reasonable utilitarian theory must recognise that people assign a nonnegligible positive utility to free personal choice, to freedom from unduly burdensome moral standards trying to regulate even the smallest details of their behaviour. Suppose we could choose between a society with the highest possible moral standards, regulating every minute of our waking lives in full detail, and a society with somewhat more relaxed moral standards, leaving us a reasonable amount of free choice in planning our everyday activities. It is very possible (though it is by no means certain) that, by imposing much stricter standards, the former society would attain higher levels of economic and cultural achievement than the latter would. Nevertheless, many of us might very well prefer to live in the latter society – at least if the differences in the economic and cultural standards between the two societies were not unduly large.

[28] Rawls 1971, p. 117.

What this means analytically is that, apart from the social utility W we assign to the outcome of any given activity, we must also assign some procedural utility $V > 0$ to our having a free personal choice among alternative activities. Suppose we have to choose between two alternative strategies S^* and S^{**} likely to yield the outcome utilities W^* and W^{**}, with $W^* > W^{**}$. Then classical utilitarianism would select strategy S^* as the *only* morally permissible strategy. But, it seems to me, we must recognise S^{**} as being an equally permissible strategy, provided that $W^{**} + V \geqq W^*$.

11 Conclusion

I have tried to show that there is a unique rational answer to the philosophical question, 'What is morality?' I have argued that, by answering this question, we obtain a very specific decision rule for choosing between alternative possible moral codes.

Even if this conclusion is accepted, this will not mean that practical moral problems from now on will become simply matters of solving some well-defined mathematical maximisation problems. Solving such problems will always involve extremely important questions of personal judgement because we have to use our own best judgement whenever we lack completely reliable factual information about some of the relevant variables. We will often lack reliable information about other people's manifest preferences and, even more so, about their true preferences. Our interpersonal utility comparisons may also be based on insufficient information, etc.

But the most fundamental source of uncertainty in our moral decisions will always lie in our uncertainty about the future, including our uncertainty about the future effects of our present policies, both in the short run and in the long run. It seems to me that careful analysis will almost invariably show that the most important source of moral and political disagreements among people of goodwill lies in divergent judgements about future developments and about the future consequences of alternative policies.

I have tried to show that an updated version of classical utilitarianism is the only ethical theory consistent with both the modern theory of rational behaviour and a full commitment to an impartially sympathetic humanitarian morality.

On the other hand, neither the concept of rationality alone, nor a commitment to a humanitarian morality alone, could yield a useful ethical theory. Rather, we need a combination of both. As I have argued in discussing Rawls's theory, even the best intuitive insight into the nature of

morality will yield a highly unsatisfactory ethical theory if these insights are conjoined with a highly irrational decision rule like the maximin principle. Conversely, even the most careful analysis of the concept of rationality cannot show that rationality entails a commitment to a humanitarian morality.

Kant believed that morality is based on a categorical imperative so that anybody who is willing to listen to the voice of reason must obey the commands of morality. But I do not think he was right. All we can prove by rational arguments is that anybody who wants to serve our common human interests in a rational manner must obey these commands. In other words, all we can prove are hypothetical imperatives of the form: 'If you want to act in a way that an impartially sympathetic observer would approve of, then do such and such', or: 'If you want your behaviour to satisfy the axioms . . . then do such and such.'[29] But I do not think that this negative conclusion is a real setback for moral philosophy, or has any important practical implications at all. As a practical matter, all of us have always known that rational discussion about moral issues is possible only between people who share some basic moral commitments, such as a common interest in a truly humanitarian morality.

Let me end with a disclaimer. I think the utilitarian theory I have described in principle covers all interpersonal aspects of morality. But I do not think it covers *all* morality. There are some very important moral obligations it fails to cover because they are matters of individual morality and of individual rationality. Perhaps the most important such obligation is that of intellectual honesty, that is, the duty to seek the truth and to accept the truth as far as it can be established – regardless of any possible positive or negative social utility this truth may have. (Telling the truth to others may be constrained by tact, respect for other people's feelings, or commitments to secrecy, etc. But admitting the truth to ourselves is not.)

Intellectual honesty requires us to accept even very unpleasant truths rather than withdraw into a dream world or a fool's paradise based on self-deception. It also requires us to accept wholeheartedly the truth that we are not alone in this world but rather share a common human nature with many millions of others. Acceptance of this particular truth is, of course, not merely a matter of theoretical rationality; rather, it is also the intellectual basis of all social morality.

[29] Harsanyi 1958.

* The author wants to thank the National Science Foundation for supporting this research through Grant Soc 77–06394 to the Center for Research in Management Science, University of California, Berkeley.

3 The economic uses of utilitarianism[1]

J. A. MIRRLEES

Some economists, when evaluating alternative economic policies, are utilitarians. At any rate they look at something they call the total utility of the outcome. This paper is intended to argue in favour of this procedure.[2] It may be as well first to exemplify it.

An interesting question is how much income ought to be redistributed from those with high wages and salaries to those with low wages. To answer it, one can set up a model in which each individual's utility is a numerical function of his net income, after taxes and subsidies, and of the quantity of labour he supplies. Each individual, supposedly knowing how his income depends on the labour he supplies, decides how much to supply by computing what will maximise his utility. All these labour supply decisions taken together determine the output of the economy. A redistributive system, consisting of taxes and subsidies, is feasible provided that the output of the economy is sufficient to provide for public and private expenditures, private expenditures being determined by private net incomes. The object of the exercise is to find which feasible redistributive system yields the greatest total utility.

This is not the place to defend the simplifications of such an economic analysis, far less to discuss how it might be improved. Even within the model outlined, assumptions as to the kinds of taxes and subsidies that are possible have a substantial effect on the results. I shall want to return to

[1] A public lecture with the same title was given at University College, London, in February 1977. The main arguments were the same, but it is doubtful whether there are any common sentences. Nevertheless I am grateful for that invitation and the opportunity it provided to attempt to articulate an economist's defence of utilitarian methods, as used in much contemporary welfare economics. I should like to acknowledge valuable discussions on these questions with J. R. Broome, P. A. Diamond, and A. K. Sen, and their comments on the first draft of this paper. Comments by P. S. Dasgupta and Q. R. D. Skinner were also useful.

[2] There have been so many papers presenting versions of utilitarianism, or defending it against criticism (many of which I have read only cursorily or not at all), that it is hard to defend writing another. But there are differences of emphasis from the major statement by Vickrey (1960), and more substantial differences from Harsanyi (1953, 1955, and later books), both of whom discuss these matters from the point of view of economic problems. Taking that point of view, I found that I wanted to deal with a number of matters not discussed by Hare (1976) and Smart (1973) in their statements.