Consequentialist Decision Theory and Utilitarian Ethics

Peter J. Hammond, Department of Economics
Stanford University, CA 94305–6072, U.S.A.

Original version prepared in May 1991 for presentation at the workshop of the International School of Economic Research on *Ethics and Economics* at the Certosa di Pontignano (Siena), July 1991. Parts of the paper are based on previous talks to the workshop on Operations Research/Microeconomics Interfaces at the European Institute for Advanced Studies in Management in Brussels, January 1990, and to the Economic Justice Seminar at the London School of Economics in November 1990.

Final version with minor revisions: January 1992; to appear in F. Farina, F. Hahn, and S. Vannucci (eds.) *Ethics, Rationality, and Economic Behaviour* (Oxford University Press).

ABSTRACT

Suppose that a social behaviour norm specifies ethical decisions at all decision nodes of every finite decision tree whose terminal nodes have consequences in a given domain. Suppose too that behaviour is both consistent in subtrees and continuous as probabilities vary. Suppose that the social consequence domain consists of profiles of individual consequences defined broadly enough so that only individuals' random consequences should matter, and not the structure of any decision tree. Finally, suppose that each individual has a "welfare behaviour norm" coinciding with the social norm for decision trees where only that individual's random consequences are affected by any decision. Then, after suitable normalizations, the social norm must maximize the expected value of a sum of individual welfare functions over the feasible set of random consequences. Moreover, individuals who never exist can be accorded a zero welfare level provided that any decision is acceptable on their behalf. These arguments lead to a social objective whose structural form is that of classical utilitarianism, even though individual welfare should probably be interpreted very differently from classical utility.

1. Introduction

Normative social choice theory seems to have started out as a discussion of how to design suitable political systems and voting schemes — as in the work of well-known writers like Borda (1781), Condorcet (1785), Dodgson (1884), Black (1948) and Arrow (1951, 1963). Yet, in its attempt to aggregate individual preferences or interests into some kind of collective choice criterion, it would appear equally suited to the general issue of how to make good decisions which affect several different individuals. This, of course, is the subject of ethics in general, rather than just of political philosophy. After all, the design of suitable political systems is just one particular kind of ethical issue. So is the design of economic systems, and even the adjustment of features like tax rates within an existing system.

This suggests that we should be most interested in a normative social choice theory that seems capable of handling practical ethical problems. My claim will be that a properly constructed form of utilitarianism has the best chance of passing this crucial test. Indeed, there are three main strands of normative social choice theory. The first is based on Arrow's original ideas, while the other two follow from succeeding major developments due to Harsanyi (1953, 1955, 1976, 1977, 1978) and Sen (1970a, b, 1977, 1982a) respectively. Of these three it is only the Harsanyi approach, when suitably and significantly modified, that appears not to create insuperable difficulties for a complete theory of ethical decision-making.

The limitations of Arrow's theory are fairly well understood, not least by Arrow himself. His crucial assumption was the avoidance of interpersonal comparisons — at least until the discussion of "extended sympathy" in the second edition of *Social Choice and Individual Values*, and a later (Arrow, 1977) article generously acknowledging the potential usefulness of work that d'Aspremont and Gevers (1977) and I (Hammond, 1976) had done in the 1970's, building on Sen's ideas (and those of Suppes, 1966). The four axioms of Arrow's impossibility theorem — namely, unrestricted domain, independence of irrelevant alternatives, the Pareto principle, and non-dictatorship — can all be satisfied if the definition of a "social welfare function" is generalized to allow interpersonal comparisons (Hammond, 1976, 1991a). It might have been better, however, if Arrow's "independence of irrelevant alternatives" axiom had been called "independence of irrelevant personal comparisons" instead, since this can then be weakened to "independence of irrelevant interpersonal comparisons" when interpersonal comparisons are allowed.

Sen's approach, by contrast, uses "social welfare functionals" that map profiles of interpersonally comparable utility functions into social orderings. These do allow interpersonal comparisons. The approach therefore does not automatically exclude rules such as Harsanyi's (or classical) utilitarianism and Rawlsian maximin. Indeed, as d'Aspremont and Gevers (1977), Roberts (1980), Blackorby, Donaldson and Weymark (1984), d'Aspremont (1985) and others have pointed out, there are many different possibilities. Actually, this indeterminacy of the social welfare functional could well be regarded as a serious weakness of Sen's approach.

Some other weaknesses of standard social choice theory, however, appear even more serious. For there are also several important questions for any ethical theory of this kind, based as it is on a social ordering, derived from unexplained interpersonal comparisons of personal utility, without making very clear what constitutes personal utility or what interpersonal comparisons are supposed to mean. These weaknesses were also present in much of my own earlier work on social choice theory.

Ultimately, in order to overcome all these defects, it would seem that a social-choice theoretic approach to ethical decision problems should be able to provide answers to the following important questions:

1) Why have a social ordering at all, instead of incomplete preferences such as those which lie behind the Pareto rule, or even some completely unpatterned social choice rule which obeys none of the usual axioms of rational choice?
2) Which individual preferences, and which individuals' preferences, should be reflected in the social choice rule? (The "which individuals" issue arises when we consider whether and how to include foreigners, animals and unborn generations in our social choice rule.)
3) What method of making interpersonal comparisons, if any, is the right one to use when arriving at the social choice rule, and what are these interpersonal comparisons meant to represent?
4) What should count in addition to individual preferences (or welfare)? Is it right that society should have preferences over issues like diet or religion which are usually regarded as purely personal? Do such personal issues force us to consider "non-welfarist" theories?

The theory I shall review in the following pages has grown over the years out of attempts to answer these and related questions. It is a utilitarian theory, but with "utility" defined in quite a different way from what almost all versions of utilitarianism seem to have used in the past. Indeed, an individual's utility — or rather, "welfare" as I shall often call it when I want to emphasize the distinction from these earlier concepts of utility — will be regarded as that function whose expected value ought to be maximized by decisions affecting only that particular individual. This implies that welfare acquires purely ethical significance. The relevance of personal tastes, preferences, desires, happiness to this ethical measure of an individual's welfare then becomes an ethical question, which is exactly what I believe it should be. Moreover, interpersonal comparisons will amount to preferences for different kinds of people — for rich over poor, for healthy over sick, for educated over ignorant, for talented over unskilled, etc. Note carefully that such preferences do not imply a disregard for those individuals who either are or will be unfortunate enough to experience poverty, sickness, ignorance or lack of skills. Rather, such preferences represent society's present and future gains from enriching the poor, healing the sick, educating the ignorant, and training the unskilled. In addition, there will be a zero level of utility which marks the threshold between the desirability and undesirability of adding an extra individual to the world's population. Finally, utility ratios will represent marginal rates of substitution between numbers of indifferent kinds of individual.

The key motivation for this revised utilitarian theory is its unique ability to treat properly multi-stage ethical decisions, represented as ethical decision trees, while at the same time recognizing that there is some concept of individual welfare which ought to determine ethical decisions. In particular, for decisions which affect only one individual, only that individual's welfare ought to matter. Of course, this excludes non-welfarist ethical theories by assumption. But I am going to claim that everything of ethical relevance to individuals can be included in our measures of the welfare of individuals, and that nothing else should matter anyway.

The first part of this paper is a review of the consequentialist approach to Bayesian decision theory. Section 2 explains why a new approach may be desirable, especially in connection with ethics and social choice theory. Section 3 considers ethical behaviour norms in finite decision trees under uncertainty and presents the two important axioms of unre-

stricted domain and consistency in continuation subtrees. Thereafter Section 4 explains the motivation for the "consequentialist" axiom, according to which only the consequences of behaviour in decision trees are relevant to proper decision making. The next few sections are inevitably rather more technical and cite results proved in Hammond (1988a). Section 5 discusses how the three axioms together imply the existence of an ethical preference ordering over uncertain consequences that satisfies the controversial independence axiom. Section 6 adds an extra continuity axiom which implies expected utility maximization.

The second part of the paper is much more directly concerned with ethics. Section 7 begins to apply the consequentialist decision theory of the first part of the paper to ethical decision problems concerning a society of individuals. To do so it introduces personal consequences, so that a social consequence is just a profile of personal consequences. Then Section 8 puts forward the hypothesis of individualistic consequentialism, according to which it is only the marginal distribution of each individual's personal consequences which is relevant to ethical decision making. In other words, it does not matter at all how different individuals' risky personal consequences are correlated — they can always be treated as if they were independently distributed.

While individualistic consequentialism captures one aspect of individualism, it does not give rise to any idea that social welfare arises from the individual welfares of different persons in society. This is remedied in Section 9, which introduces the concept of "individual welfarism." It is assumed that, just as society has its ethical behaviour norm for social decision trees, so there is an ethical behaviour norm for "individual decision trees." Such trees are particular social decision trees in which there is only one individual whose probability distribution of personal consequences can be affected by any decision that is taken. It is required that the social behaviour norm in any such individual decision tree should exactly match the ethical behaviour norm for the relevant individual. Under the consequentialist axioms of Sections 3–6, as well as the new conditions set out in Sections 7 and 8, individual welfarism implies the existence of a cardinal equivalence class of individual welfare functions for each individual, whose expected values are maximized by decisions corresponding to the individual norm.

Section 10 goes on to show, moreover, that the conditions of Harsanyi's (1955) utilitarian theorem are all satisfied. Thus there exists a cardinal equivalence class of social

welfare functions whose expected values are maximized by the social norm, and which can be expressed as the sum of suitably normalized individual welfare functions. So the social welfare functional linking individual and social welfare functions is simply additive, as in classical utilitarianism. As is pointed out in Section 10, however, the individual welfare functions which ought to be added have a very different interpretation from the classical concept of utility.

In Section 11 the vexed question of optimal population is taken up. It is assumed that individuals who do not exist already can be ignored whenever none of the decisions being contemplated could possibly lead to their coming into existence. Thus, in any individual decision tree where the only individual who could be affected never comes into existence anyway, it does not matter what decision is made. This assumption has the effect of determining a constant individual welfare level corresponding to non-existence. The individual welfare function can then be normalized so that this level is zero. The implication is that it is only necessary to sum the individual welfares of those individuals who do come into existence; all other individuals' welfare levels are zero and so their welfare can be ignored.

A crucial question raised above was how to make sense of the interpersonal comparisons which are needed in any satisfactory resolution of Arrow's impossibility theorem. This is the topic of Section 12 which, as promised, shows how the utilitarian objective being propounded here relates interpersonal comparisons to, logically enough, social preferences for different kinds of persons.

The final Section 13 contains a concluding assessment of what has and has not been achieved so far in this research project.

2. Bayesian Decision Theory

How does one make good ethical decisions? This is obviously the main question in any ethical theory which is going to arrive at specific recommendations for action. Moreover, this question is not so different from the general problem of how to make good decisions in general, which is the subject of decision theory. As in that theory, it will be helpful to consider what acts are possible, what consequences those acts lead to, and how those consequences should be evaluated. The only special features of ethical decision theory, in fact, are the kind of consequence which we shall admit as relevant, and the way we think about and evaluate those consequences.

In normative decision theory, a standard axiomatic approach was formulated during the 1940's and 1950's, based upon the major contributions of von Neumann and Morgenstern and Savage in particular. It involved a system of axioms whose implication was that agents should have subjective probabilities about uncertain events, and a (cardinal) utility function for evaluating consequences. Moreover, the best action was that which would maximize expected utility. This is the approach which, following Harsanyi (1978), I shall call *Bayesian decision theory*.

Even as a normative standard, this theory has come under heavy attack in recent years. Yet, with a few exceptions such as Machina (1989) and McClennen (1990), it seems to me that most of the critics have not really fully understood the theory. In particular, they have often failed to appreciate how adaptable it is, and how it can handle many of the familiar criticisms by a suitable extension of the concept of a relevant "consequence." In addition, it must be pointed out that the usual framework in which the axioms of decision theory are presented is very special. Following von Neumann and Morgenstern's recommended procedure, complicated intertemporal decision problems are generally collapsed into their "normal form," in which the decision maker makes a single choice of strategy or plan which is intended to cover all possible future contingencies. Yet real decision problems offer the chance to change one's mind in future, since decisions are not usually made as irrevocable commitments to a particular strategy. And, as I have pointed out before (Hammond, 1988c, 1989), the main alternatives to Bayesian decision theory, with its criterion of maximizing expected utility, create for the decision-maker the risk that *ex ante* plans will not be carried out but will get revised later, even though nothing unforeseen has happened in the

meantime. This is very like the inconsistency phenomenon in dynamic choice which Strotz (1956) was the first to explore formally; it is also related to "subgame imperfections" of the kind first considered by Selten (1965) in n-person game theory.

One of the most fundamental axioms of Bayesian decision theory is the existence of a preference ordering over the space of event-contingent consequences. With the notable exceptions of Levi (1974, 1980, 1986), Seidenfeld (1988a, b) and Bewley (1989), even most critics of the theory accept this axiom. Yet many ethical theorists do not. Some of these simply claim that nobody has any business constructing a "social preference ordering" over decisions or the consequences to which they lead. Others claim to find it objectionable that anything as subtle and complicated as ethics could be reduced to something as conceptually simple or crude as the maximization of a preference ordering. This, however, overlooks the obvious point that, though a preference ordering may seem like a simple concept, it could still range over an immensely complicated space of ethically relevant consequences. As an analogy, the original Zermelo (1913) theory of two-person games with complete information shows how each player has an optimal strategy in chess, specifying what move should be made in each possible position. It is inconceivable that the optimal strategy could even be found, however, because chess is far too complicated and subtle a game (and the Japanese game of Go, to which the same argument applies, is perhaps even more so).

In an attempt to meet all these cogent objections, I have therefore been developing a different justification for Bayesian decision theory. The standard axioms emerge as implications of what may seem less objectionable "consequentialist" axioms. Rather than assume that there is a preference ordering, the new theory *proves* that behaviour must reveal such an ordering. Under an additional minor but necessary continuity axiom, it also proves that there exists a unique cardinal equivalence class of utility functions whose expected value is maximized. Of course, the proofs of such results do rely on other axioms, but they may seem less unnatural or open to criticism than many have found the standard axioms to be.

3. Decision Trees and Ethical Theories

The approach I have adopted begins by recognizing that there are multi-stage decision problems which can be described by means of *decision trees*. The typical decision tree will be denoted by T. It has a set of *nodes* N. To avoid unnecessary technical complications, I shall work only with finite trees — i.e., trees for which N is finite. Among the nodes in N is a subset N^* of *decision nodes*, at which the decision maker is offered the choice of several different possible actions.

To represent uncertainty, there will also be a set N^0 of *chance nodes* at which "nature makes a decision" outside the decision-maker's control. Really we should now follow the argument presented in Hammond (1988a) and discuss decision theory in the absence of probabilities, seeing what assumptions are needed in order to ensure the existence of at least subjective probabilities. Rather than do so here, however, it will simply be assumed that at each chance node $n \in N^0$ there is always an associated probability distribution $\pi(n'|n)$ over the finite set $N_{+1}(n)$ of nodes n' which immediately succeed n. To avoid problems that arise in continuation subtrees which are only reached with probability zero, it will be assumed here that any node $n' \in N_{+1}(n)$ for which $\pi(n'|n) = 0$ gets "pruned" from the decision tree, along with the set $N(n')$ of all succeeding nodes. Then only nodes n' for which $\pi(n'|n)$ is positive will remain, and so we can indeed assume that $\pi(n'|n) > 0$ whenever $n \in N^0$ and $n' \in N_{+1}(n)$.

Any decision tree T starts at an *initial node* n_0, which could be either a decision node or a chance node. Since the tree is finite, it must also have a set X of *terminal nodes* at which everything has been resolved. Then N must be the union $N^* \cup N^0 \cup X$ of the three disjoint sets N^*, N^0 and X.

Nature's "choices" of events and the decision-maker's choices of acts will combine to determine a unique path through the decision tree, starting at the initial node n_0 and ending at some terminal node $x \in X$. In fact there is an obvious one-to-one correspondence between paths through the tree and terminal nodes. Along any such path, it is assumed that there is a history of ethically relevant consequences which can be summarized as just a *consequence* y in some domain of consequences Y. It does no harm then to think of there being a unique consequence attached to each terminal node of the tree — in other words, there is a function $\gamma : X \to Y$ mapping each terminal node x of the decision tree into

the consequence $\gamma(x)$ of following through the tree the unique path which ends at x. It is assumed that ethical decisions should depend only on their different consequences y in a fixed *consequence domain* Y; consequences outside Y are ethically irrelevant.

Each path through the decision tree also corresponds to a unique sequence of choices by nature, which then determines a history of events. Since the probabilities $\pi(n'|n)$ of these successive choices have been specified, there is a corresponding probability $\xi(x)$ of reaching any given terminal node $x \in X$ and then of getting the consequence $\gamma(x)$. In fact decisions will give rise to probability distributions of consequences, in a way to be explained in Section 5 below.

One last crucial ingredient is needed for an ethical theory. This is the concept of a behaviour norm. Behaviour at any decision node $n \in N^*$ can be regarded as selecting some non-empty set of nodes from $N_{+1}(n)$, the set of nodes immediately succeeding n. Note that multiple choices at n are allowed, just as economic theory allows consumers to be indifferent between two or more options. Moreover, there is no reason to think that only one decision at each decision node is ethically acceptable. Thus a *behaviour norm* is formally defined as a function β which specifies a non-empty *behaviour set* $\beta(T, n) \subset N_{+1}(n)$ of decisions which are ethically appropriate, or recommended, at each decision node n of each decision tree T in the tree domain $\mathcal{T}$.

An *ethical theory* will then consist of the following three items:

(i) a consequence domain Y of possible ethical consequences;

(ii) a tree domain $\mathcal{T}$ of finite decision trees whose terminal nodes $x \in X$ are mapped into consequences $\gamma(x) \in Y$;

(iii) a behaviour norm $\beta(T, n)$ defined for every decision node $n \in N^*$ of each decision tree T in the tree domain $\mathcal{T}$.

Two important axioms will now be imposed upon such an ethical theory. The first is that of an *unrestricted domain*: it is required that the tree domain $\mathcal{T}$ should consist of all logically possible finite decision trees, each with its own mapping $\gamma : X \to Y$ from terminal nodes to appropriate ethical consequences. Any theory which applies to only a restricted domain of decision trees will not be able to handle some ethical decision problems which might conceivably arise, or which a complete theory should be able to handle even if the

problem is entirely hypothetical. Thus, having an unrestricted domain seems necessary for a complete ethical theory.

The second axiom is *consistency in continuation subtrees* (or "consistency" for short). At any node n of a decision tree T, there is a corresponding *continuation subtree* $T(n)$ which is obtained by pruning T just before the node n, and retaining what gardeners would call a "cutting" consisting of both that node and all its successors in T. This subtree is, of course, a decision tree in its own right with an initial node n which is just after the cut and the set of nodes $N(n)$ which is a subset of N, the set of nodes in the original tree. The subtree's set of terminal nodes is $X(n)$, the set of those terminal nodes $x \in X$ of the original decision tree which succeed the initial node of the subtree — i.e., $X(n) = X \cap N(n)$. The terminal nodes $x \in X(n)$ in the subtree are still mapped by γ into consequences — that is, $\gamma(x)$ remains well defined for all $x \in X(n)$.

Because of the unrestricted domain assumption, any such continuation subtree $T(n)$ is in the tree domain $\mathcal{T}$. So the behaviour norm β is defined at each decision node $n \in N^*(n) = N^* \cap N(n)$ of the continuation subtree $T(n)$. Yet each such node is identical to a decision node $n \in N^*$ of the original tree. All that has happened in passing from tree T to tree $T(n)$ is that time has progressed, so that the set of possible courses of history has become narrowed. This is inevitable. So the description of behaviour $\beta(T, n')$ at each decision node n' of the continuation subtree should be the same, regardless of whether we think of n' as a node of the subtree $T(n)$ or as a node of the original tree T. Since the behaviour norm must describe possible behaviour, it is therefore required to specify the same set of decisions $\beta(T(n), n')$ at each decision node $n' \in N^*(n)$ of the continuation subtree as it does at the corresponding node of the full tree. In other words, $\beta(T(n), n') = \beta(T, n')$ whenever $n \in N$ and $n' \in N^*(n)$. This is (continuation) *consistency*, which the second axiom requires.

In a sense, this consistency condition is almost tautological. For, when a specific decision node $n \in N^*$ is reached, the decision maker is really faced only with the continuation subtree $T(n)$ starting at that node. What counts, therefore, is the behaviour set $\beta(T(n), n)$ which the norm prescribes for that decision node in the continuation tree. If this differs from $\beta(T, n)$, which was prescribed for the same decision node in the earlier and larger decision tree T, then this earlier recommendation really carries no force (unless it is recalled as an ethically relevant resolution to behave in a certain way, but then the history of consequences

should be expanded to include such resolutions and whether they become honoured or not). In which case we might as well define the behaviour set $\beta(T, n)$ at each decision node $n \in N^*$ of a decision tree T as the value of the behaviour set $\beta(T(n), n)$ at node n in the continuation tree $T(n)$ which starts at that node. The result will then be a behaviour norm which is automatically consistent because one will have $\beta(T(n), n') = \beta(T(n'), n') = \beta(T, n')$ whenever $n \in N$ and $n' \in N^* \cap N(n)$.

4. Consequentialism

A fundamental postulate of decision theory is that behaviour should be entirely explicable by its consequences. Indeed, this is so fundamental that standard decision theories such as that due to Savage (1954) have even *defined* an act as a mapping from states of the world into consequences. Obviously then, for Savage and other decision theorists, two acts which give rise to identical patterns of state contingent consequences are completely equivalent.

In ethics, the doctrine that an act should be judged by its consequences has been much more controversial. The idea can be traced back to Aristotle, who wrote:

> *If, then, there is some end of the things we do, which we desire for its own sake (everything else being desired for the sake of this), and if we do not choose everything for the sake of something else (for at that rate the process would go on to infinity; so that our desire would be empty and vain), clearly this must be the good and the chief good.*
>
> — Aristotle, Niomachean Ethics, 1094a 18.

Later, St. Thomas Aquinas sought to refute Aristotelian doctrine, and effectively defined consequentialism by defining its negation:

> *A consequence cannot make evil an action that was good nor good an action that was evil.*

More recently, Mill (see Warnock, 1962) and Moore (1912, p. 121) can be counted among those who thought that consequences are what matter about acts. The term "consequentialism" itself, however, seems rather recent — it was used by Anscombe (1958) to describe a doctrine she wished to criticize. The attacks have continued. Williams (1973) sought to rebut not only utilitarianism, but also consequentialism of which it is a special

case. In Williams (1985), he dismisses it in barely half a sentence as merely an elementary error. Sen and Williams (1982) had chosen "Beyond Utilitarianism" as the provisional title of the volume they edited until it was pointed out that some contributors were reluctant even to step beyond utilitarianism, let alone beyond the broader doctrine of consequentialism.

Sen (1987, pp. 74–78; and the articles cited there) has also remained a critic, even though his attacks may have become muted over the years. In fact, he has recognized the argument (which Williams had made earlier) that one could extend the domain of consequences until it incorporated everything relevant to the ethical merits of any act. What remains at issue, then, is included in the following passage from Sen (1987, pp. 75–76):

> *Consequentialism ... demands, in particular, that the rightness of actions be judged entirely by the goodness of consequences, and this is a demand not merely of taking consequences into account, but of ignoring everything else. Of course, the dichotomy can be reduced by seeing consequences in very broad terms, including the value of actions performed or the disvalue of violated rights. I have tried to argue elsewhere [Sen, 1982b, 1983]:*
>
> 1. *that such broadening is helpful, even essential; but*
> 2. *that nevertheless even after fully fledged broadening, there can remain a gap between consequentialist evaluation and consequence-sensitive deontological assessment.*

Moreover, Sen (1982b, 1983) points out that most of the concepts of "consequence" used in the past have been too narrow, because they pay too little attention to rights and to who performs what action. I agree too that how consequences are evaluated, and perhaps how they are even defined, can depend on who does the evaluation. After all, the world might be a better place if we all demanded higher ethical standards of ourselves than we do of others. I therefore prefer the wider concept of consequence which Sen prefers to call a "consequence-based evaluation".

In the end, however, I disagree with Sen (and Regan, 1983) because I do not see any "gap between consequentialist evaluation and consequence-sensitive deontological assessment." I claim that any apparent gap can be closed by expanding the domain of consequences even further, if necessary, in order to embrace all possible results of any "deontological assessment" which were not included in the original domain of consequences. As Williams recognized, this makes consequentialism become a tautology. In the past, the

tautology has sometimes been described as "meaningless." I am perfectly willing to admit that consequentialism does only acquire meaning with reference to some specific domain of consequences, in which case the tautology has been removed. But I would rather that future debates could be about what consequences really should be included in the domain because they are ethically relevant, instead of about the appropriateness of a doctrine which can be made into a tautology anyway. And if the term "consequence" remains anathema, perhaps we can change it to something else like "assessment." By doing so, however, we sever the convenient link to the standard terminology of decision theory. Accordingly, I prefer to stick to "consequentialism."

5. Consequentialist Choice

Let us proceed, then, to consider what it means for our ethical theory if actions are judged entirely by their consequences. Obviously, it must mean that our apparatus of decision trees whose terminal nodes have consequences is sufficient to describe the ethical decision problems which they represent. If two decision trees are identical, they represent the same decision problem. There is no need to concern ourselves with differences between the two problems when the consequences that are available in the decision tree and also in each continuation subtree are entirely equivalent. Behaviour should be equivalent at each (equivalent) division node of the two equivalent trees, and lead to an identical pattern of state-contingent consequences.

In fact the consequentialist hypothesis is stronger than this, but not too dissimilar in spirit. What it adds is the idea that not even the structure of the decision tree is important (unless it somehow affects the consequences which the decision maker has available). The hypothesis requires that the (choice set of) consequences of prescribed behaviour should be entirely explicable by the (feasible set of) consequences of possible behaviour. To explain this properly requires a somewhat careful construction of the feasible set and the associated choice set.

Take the feasible set $F(T)$ first. Its members are precisely those probability distributions $p(y)$ over the consequence domain Y that can result from some decision strategy which is available in the tree T — i.e., from some rule specifying a unique action $\alpha(n) \in N_{+1}(n)$ at every decision node $n \in N^*$ of T. Write $\Delta(Y)$ for the set of all probability distributions

over Y which attach positive probability to only a finite subset of Y (this finite subset is usually called the *support* of the distribution). Then every decision strategy results in a unique probability distribution $p(\cdot) \in \Delta(Y)$. Indeed, if $\xi_\alpha(x)$ $(x \in X)$ denotes the probability distribution over terminal nodes in X that is induced by the actions $\alpha(n)$ $(n \in N^*)$, then

$$p(y) = \sum\nolimits_{x \in \gamma^{-1}(y)} \xi_\alpha(x)$$

is the probability of consequence y, for each $y \in Y$. That is, the probability of y is the total probability of all the different terminal nodes x for which $\gamma(x) = y$.

Actually, the feasible set $F(T)$ can be constructed by backward recursion, starting at terminal nodes $x \in X$ where only a single determinate consequence $\gamma(x) \in Y$ is possible. There $F(T(x)) = \{\chi_{\gamma(x)}\}$ — i.e., the only member of $F(T(x))$ is the degenerate probability distribution $\chi_{\gamma(x)}$ which attaches probability one to the particular consequence $\gamma(x)$.

At any previous node n of the tree T, the feasible set $F(T(n))$ can be constructed from the collection of sets $F(T(n'))$ at the immediately succeeding nodes $n' \in N_{+1}(n)$ as follows. First, in the case when n is a decision node in N^*, the feasible set $F(T(n))$ is the union $\cup_{n' \in N_{+1}(n)} F(T(n'))$. This is because the decision maker's next move to $n' \in N_{+1}(n)$ determines which set $F(T(n'))$ of this union will still be possible after that move. In the case when n is a chance node in N^0, however, the feasible set satisfies

$$\begin{aligned} F(T(n)) &= \sum\nolimits_{n' \in N_{+1}(n)} \pi(n'|n)\, F(T(n')) \\ &= \{\, p \in \Delta(Y) \mid \exists p_{n'} \in F(T(n'))\ (n' \in N_{+1}(n)) : p = \sum\nolimits_{n' \in N_{+1}(n)} \pi(n'|n)\, p_{n'} \,\}. \end{aligned}$$

That is, $F(T(n))$ consists of all possible probability distributions which result from combining into a compound lottery in an appropriate way one member selected from each of the respective feasible sets $F(T(n'))$ $(n' \in N_{+1}(n))$. The explanation is that nature's next move will determine which term $F(T(n'))$ of the sum will be the next appropriate feasible set, and these different terms occur with probabilities $\pi(n'|n)$.

It can easily be shown by backward induction on n that $F(T(n)) \subset \Delta(Y)$ for each $n \in N$. Moreover, since the tree is finite, the backward recursion must eventually terminate at the initial node n_0 of the tree, and yield the appropriate *feasible set* of contingent consequences $F(T) = F(T(n_0))$ for the tree T as a whole.

The choice set $\Phi_\beta(T)$, on the other hand, will consist of those random consequences which can result from some *prescribed decision strategy* $\alpha(n)$ — i.e., a strategy which, at each decision node $n \in N^*$ of the given tree, selects a single member $\alpha(n)$ of the behaviour set $\beta(T,n) \subset N_{+1}(n)$ which the ethical behaviour norm prescribes for that node. This set can be constructed by backward recursion just as the feasible set was. The only difference is obvious: at each decision node $n \in N^*$, the choice set $\Phi_\beta(T(n))$ consists of the union $\cup_{n' \in \beta(T,n)} \Phi_\beta\left(T(n')\right)$ of the choice sets at only those immediately succeeding nodes $n' \in \beta(T,n)$ which could result from prescribed behaviour at node n; at each chance node $n \in N^0$, on the other hand, the choice set $\Phi_\beta(T(n))$ consists, as before, of the probability weighted sum $\sum_{n' \in N_{+1}(n)} \pi(n'|n)\, \Phi_\beta(T(n'))$ of all the choice sets $\Phi_\beta(T(n'))$ at the immediately succeeding nodes $n' \in N_{+1}(n)$. This backward recursion again terminates at the initial node, and yields the appropriate *choice set* $\Phi_\beta(T) = \Phi_\beta(T(n_0))$ of contingent consequences which could result from following the prescribed behaviour norm β throughout the whole decision tree T. Obviously, this choice set is a non-empty subset of the feasible set $F(T)$ — as is easily proved by considering each step of the backward recursion in turn and using mathematical induction. The choice set $\Phi_\beta(T)$ could consist of the whole feasible set, it should be remembered.

After these necessary preliminaries, the crucial hypothesis of *consequentialist behaviour* can be formally stated. It requires that, whenever two decision trees T and T' have identical feasible sets of contingent consequences $F(T) = F(T')$, the two choice sets $\Phi_\beta(T) = \Phi_\beta(T')$ must also be equal. If this hypothesis is true, the ethical theory is said to be *consequentialist*. In fact it obviously implies that there is a "revealed" *consequentialist choice function* C_β mapping each non-empty finite feasible set $F \subset \Delta(Y)$ of random consequences $p(y) \in \Delta(Y)$ into the choice set $C_\beta(F)$, which is some non-empty subset of the feasible set F. This revealed choice function must satisfy $\Phi_\beta(T) = C_\beta(F(T))$ for all finite decision trees $T \in \mathcal{T}$.

So far, then, the following three axioms have been formulated for normative behaviour in decision trees: (i) unrestricted domain; (ii) consistency in continuation subtrees; (iii) consequentialism. Following the arguments presented elsewhere (Hammond, 1988a), these three axioms imply that there exists a (complete and transitive) *revealed preference ordering* R_β on the set $\Delta(Y)$ with the property that

$$C_\beta(F) = \{\, p \in F \mid q \in F \Longrightarrow p\, R_\beta\, q \,\}$$

whenever $\emptyset \neq F \subset \Delta(Y)$ and F is finite.

Moreover, one other very important property also follows from these same three axioms. This is the controversial *independence* condition, according to which the revealed preference ordering R_β on $\Delta(Y)$ must satisfy

$$[\alpha\, p + (1-\alpha)\, \bar{p}]\, R_\beta\, [\alpha\, q + (1-\alpha)\, \bar{p}] \iff p\, R_\beta\, q$$

whenever $p, \bar{p}, q \in \Delta(Y)$ and $0 < \alpha \leq 1$.

As pointed out in Hammond (1988a), however, these are the only restrictions on behaviour which the three axioms imply. That is, given any preference ordering R satisfying the independence condition, behaviour whose set of random consequences always maximizes this preference ordering in each possible finite decision tree will certainly satisfy the three axioms.

Although the independence condition is both implied by expected utility maximization and is usually formulated as one of the axioms implying expected utility maximization, the three axioms enunciated here do not on their own imply expected utility maximization. The reason is that the revealed preference ordering R_β could still be discontinuous and not admit any utility representation at all. Indeed, consider the case when Y consists of three different members y_k $(k = 1, 2, 3)$. Then an ordering satisfying the independence axiom is given by

$$p\, R_\beta\, q \iff [p(y_1) > q(y_1)] \quad \text{or} \quad [p(y_1) = q(y_1) \quad \text{and} \quad p(y_2) \geq q(y_2)].$$

This is a lexicographic preference ordering, of course, giving priority first to increasing the probability of y_1 but then, if this probability can be increased no further, recognizing as desirable increases in the probability of y_2 (and so, since probabilities sum to one, decreases in the probability of y_3).

6. Continuity and Expected Utility

Such discontinuous preferences are easily excluded by imposing an additional axiom of continuity on behaviour norms in decision trees. Specifically, let T^m $(m = 1, 2, \ldots)$ be any infinite sequence of decision trees which all have the same sets of decision nodes N^*, chance nodes N^0, and terminal nodes X, the same sets $N_{+1}(n)$ of nodes immediately succeeding each node $n \in N$, and the same mapping $\gamma : X \to Y$ from terminal nodes to consequences. The only way in which the trees T^m differ is in the probability distributions $\pi^m(n'|n)$ $(n' \in N_{+1}(n)\ )$ attached to each chance node $n \in N^0$. Moreover, assume that $\pi^m(n'|n) \to \pi(n'|n)$ as $m \to \infty$, where $\pi(n'|n) > 0$ (all $n \in N^0$, $n' \in N_{+1}(n)$). Then the behaviour norm $\beta(T, n)$ is said to be *continuous* if, whenever $n \in N^*$ and $n' \in \beta(T^m, n)$ for all large m, then $n' \in \beta(T, n)$. Mathematical economists will recognize this as upper hemi-continuity of the behaviour correspondence as probabilities vary.

It is not difficult to prove that this additional continuity axiom implies that the revealed preference ordering R_β is continuous as well, in the sense that for all $\bar{p} \in \Delta(Y)$ the two preference sets

$$\{\, p \in \Delta(Y) \mid p\, R_\beta\, \bar{p}\,\}, \quad \{\, p \in \Delta(Y) \mid \bar{p}\, R_\beta\, p\,\}$$

are both closed sets of $\Delta(Y)$. Then the ordering R_β can certainly be represented by a utility function U defined on $\Delta(Y)$, in the sense that $p\, R_\beta\, q \iff U(p) \geq U(q)$ for all pairs $p, q \in \Delta(Y)$. Moreover the independence condition implies (Herstein and Milnor, 1953) that $U(p)$ can be chosen so that it takes the expected utility form

$$U(p) = \sum\nolimits_{y \in Y} p(y)\, v(y) = \mathbb{E}_p\, v(y)$$

for some unique cardinal equivalence class of *von Neumann-Morgenstern* utility functions (NMUF's) v defined on Y.

7. Social Norms and Personal Consequences

Having developed the basic decision theory, the next stage of the argument is much more directly concerned with ethical decisions whose consequences can affect many individuals simultaneously. To represent such consequences, the domain Y will now be given much more structure.

As in social choice theory, assume that there is some basic set A of possible *social states* $a \in A$. The *membership* M of a society is just the set of individuals i in that society. Given any $i \in M$, write A_i for a copy of the set A whose members a_i are *i's personalized social states*. As in the theory of public goods (Foley, 1970, p. 70; Milleron, 1972 etc.) it helps to imagine that we can choose different social states $a_i \neq a_j$ for individuals i and j whenever they are different members of M, even though this may well be impossible in practice.

In addition to social states in the conventional sense, it will be convenient to consider also for each $i \in M$ a space of *personal characteristics* $\theta_i \in \Theta_i$. Such characteristics determine *i's* preferences, interests, talents, and everything else (apart from the social state) which is ethically relevant in determining the welfare of individual i. In Section 11 below, θ_i will even indicate whether individual i ever comes into existence or not.

For each individual i, a *personal consequence* is a pair $z_i = (a_i, \theta_i)$ in the Cartesian product set $Z_i := A_i \times \Theta_i$ of personalized social states a_i and personal characteristics θ_i. Then, in a society whose membership M is fixed, a typical *social consequence* consists of a profile $z^M = (z_i)_{i \in M} \in Z^M := \prod_{i \in M} Z_i$ of such personal consequences — one for each individual member of society (both actual and potential). The consequence domain $Y = Z^M$ will then consist of all such social consequences, with typical member $y = z^M$.

The four consequentialist axioms given in Sections 3, 5 and 6 above can now be applied to a *social behaviour norm* $\beta(T, n)$ defined at all decision nodes n of all decision trees T in the domain of finite decision trees with consequences in Z^M. These axioms obviously imply the existence of a unique cardinal equivalence class of von Neumann-Morgenstern *social welfare functions* $w(y) \equiv w(z^M)$, defined on the space of social consequences, such that the social behaviour norm β always results in consequences that maximize the expected value $w(z^M)$ in every social decision tree. Thus, the only difference so far from Section 6 is that the consequence domain has become one of social consequences. What is most important, however, is the idea that each personal consequence $z_i \in Z_i$ captures everything of ethical

relevance to individual i — by definition, nothing else, including no other individual's personal consequences, can possibly be relevant to i's welfare.

8. Individualistic Consequentialism

A general random social consequence is some joint probability distribution $p \in \Delta(Z^M)$ over the product space Z^M of different individuals' personal consequences. Such personal consequences could be correlated between different individuals, or they could be independent. The extent of this correlation should be of no consequence to any individual, however. For, provided that everything relevant to individual $i \in M$ really has been incorporated in each personal consequence $z_i \in Z_i$, all that really matters to i is the distribution $p_i \in \Delta(Z_i)$ of these consequences. This leads to the *individualistic consequentialism* hypothesis that any two lotteries $p, q \in \Delta(Z^M)$ are to be regarded as equivalent random consequences whenever, for every individual $i \in M$, the marginal distributions $p_i = q_i \in \Delta(Z_i)$ of i's consequences are the same. This means in particular that if any such pair $p, q \in F(T)$ for any decision tree T in the domain, then

$$p \in \Phi_\beta(T) \iff q \in \Phi_\beta(T).$$

Considering the case when $F(T) = \{ p, q \}$ shows that, when individualistic consequentialism is combined with the other consequentialist axioms of Sections 3, 5 and 6, then

$$p_i = q_i \text{ (all } i \in M) \implies \mathbb{E}_p\, w(z^M) = \mathbb{E}_q\, w(z^M)$$

— i.e., p and q must be indifferent according to the relevant expected utility criterion whenever the personal marginal distributions are all equal.

Succinctly stated, individual consequentialism amounts to requiring that only each individual's distribution of personal consequences be relevant to any social distribution. There is no reason to take account of any possible correlation between different individuals' personal consequences.

9. Individual Welfarism

The second individualistic axiom which I shall use is that there is an individual welfare behaviour norm defined for all "individualistic" decision trees. The latter are trees for which there is only one individual whose distribution of personal consequences is affected by any decision within the tree. Thus, if $i \in M$ denotes the only individual affected in the tree T, then there must be a profile $\bar{p}_{-i} \in \prod_{h \in M \setminus \{i\}} \Delta(Z_h)$ of fixed lotteries $\bar{p}_h \in \Delta(Z_h)$ $(h \in M \setminus \{i\})$ for each individual h other than i, as well as a set $F_i(T) \subset \Delta(Z_i)$ of feasible lotteries over personal consequences for individual i, such that the set $F(T)$ of lottery profiles which are feasible in the tree T satisfies $F(T) = F_i(T) \times \{ \bar{p}_{-i} \}$. A decision tree with this property will be called an *individualistic decision tree*. If $i \in M$ is the only individual affected by decision in the tree T, then T can be called an *i-decision tree*. Let $\mathcal{T}_i$ denote the set of all such trees.

The crucial hypothesis to be introduced now is that there is an *individual welfare* behaviour norm $\beta_i(T, n)$ defined for every individual $i \in M$ and every decision node n of every i-decision tree $T \in \mathcal{T}_i$. It is this norm which, by definition, should represent ethical behaviour when only i is affected by whatever decision is taken. Moreover, it is natural to require β_i to satisfy the consequentialist axioms stated in Sections 3, 5 and 6 above, and even to do so in a way which is independent of the profile $\bar{p}_{-i}$ of fixed lotteries $\bar{p}_h$ for all unaffected individuals $h \in M \setminus \{i\}$. This last independence property is the key hypothesis here. The motivation is that, if only consequences to i are affected by any decision, the fixed consequences to all other individuals are ethically irrelevant — assuming, as I do, that everything relevant to ethical decision making is already included in the consequences, and that only (distributions over) personal consequences matter.

Given any individual $i \in M$ and i-decision tree $T \in \mathcal{T}_i$, the assumption of *individual welfarism* requires that the social norm β and the individual norm β_i should be identical at all decision nodes of T. Equivalently, in any i-decision tree $T \in \mathcal{T}_i$ with $\bar{p}_{-i}$ as a fixed profile of random consequences for individuals $h \neq i$, the two sets $\Phi_\beta(T)$ and $\Phi_{\beta_i}(T)$ of social consequences and of i's personal consequences which are revealed as chosen by β and β_i respectively should satisfy $\Phi_\beta(T) = \Phi_{\beta_i}(T) \times \{ \bar{p}_{-i} \}$. Thus, whenever there is "no choice" in the personal consequences of all other individuals, the social norm becomes identical to the only affected individual's welfare norm. Note especially that individual welfarism poses

no restrictions on what is allowed to count as part of a personal consequence and so to affect each individual's welfare. All it says is that, in "one person situations," social welfare is effectively identified with that one person's individual welfare.

In combination with the other consequentialist axioms, individual welfarism obviously implies the existence of a unique cardinal equivalence class of *individual welfare functions* $w_i(z_i)$ for each $i \in M$. These have the property that each individual i's welfare norm β_i will always yield in every i-decision tree T the set of random consequences $\Phi_{\beta_i}(T)$ that maximize with respect to $p_i \in \Delta(Z_i)$ the expected value $\mathbb{E}_{p_i} w_i(z_i)$ of w_i over the set $F_i(T) \subset \Delta(Z_i)$ of feasible probability distributions over i's personal consequences.

10. Utilitarianism

Individual welfarism has a much more powerful implication, however, when it is combined with individualistic consequentialism as defined in Section 8. For suppose that the two lotteries $p, q \in \Delta(Z^M)$ are such that $\mathbb{E}_{p_i} w_i(z_i) = \mathbb{E}_{q_i} w_i(z_i)$ for all $i \in M$, where $p_i, q_i \in \Delta(Z_i)$ denote the respective marginal distributions over just i's personal consequences. Now order the individuals $i \in M$ so that $M = \{ i_1, i_2, \ldots, i_r \}$ where r is the total number of individuals. We shall prove by induction on the integer s that, if w denotes a social welfare function as defined in Section 7, then the equation $E_s(p, q)$ expressed by

$$\mathbb{E}_{p_{i_1}, p_{i_2}, \ldots, p_{i_{s-1}}, p_{i_s}, q_{i_{s+1}}, \ldots, q_{i_r}} w(z^M) = \mathbb{E}_{p_{i_1}, p_{i_2}, \ldots, p_{i_{s-1}}, q_{i_s}, q_{i_{s+1}}, \ldots, q_{i_r}} w(z^M)$$

or by

$$\sum_{z_{i_1}} \cdots \sum_{z_{i_{s-1}}} \sum_{z_{i_s}} \sum_{z_{i_{s+1}}} \cdots \sum_{z_{i_r}} p_{i_1}(z_{i_1}) \ldots p_{i_{s-1}}(z_{i_{s-1}}) \, [p_{i_s}(z_{i_s}) - q_{i_s}(z_{i_s})]$$
$$q_{i_{s+1}}(z_{i_{s+1}}) \ldots q_{i_r}(z_{i_r}) \, w(z_{i_1}, \ldots z_{i_{s-1}}, z_{i_s}, z_{i_{s+1}}, \ldots z_{i_r}) = 0$$

must be true for $s = 1$ to r. Note that the expectations in this equation are taken with respect to two distributions having identical marginal distributions for all individuals except i_s.

Indeed, to show that the equation $E_s(p, q)$ is true, it suffices to consider an i_s-decision tree T in which

$$F(T) = \prod_{t=1}^{s-1} \{p_{i_t}\} \times \{ p_{i_s}, q_{i_s} \} \times \prod_{t=s+1}^{r} \{q_{i_t}\}.$$

Then, since $\mathbb{E}_{p_{i_s}} w_{i_s}(z_{i_s}) = \mathbb{E}_{q_{i_s}} w_{i_s}(z_{i_s})$ by hypothesis, it must be true that $\Phi_{\beta_{i_s}}(T) = \{p_{i_s}, q_{i_s}\}$. Hence individual welfarism implies that

$$\Phi_\beta(T) = \prod_{t=1}^{s-1} \{p_{i_t}\} \times \Phi_{\beta_{i_s}}(T) \times \prod_{t=s+1}^{r} \{q_{i_t}\} = F(T)$$

and so the equality above does indeed follow.

Stringing all the r equations $E_s(p, q)$ ($s = 1$ to r) together then implies that $\mathbb{E}_p\, w(z^M) = \mathbb{E}_q\, w(z^M)$. Thus it has finally been proved that

$$\mathbb{E}_{p_i}\, w_i(z_i) = \mathbb{E}_{q_i}\, w_i(z_i) \text{ (all } i \in M) \Longrightarrow \mathbb{E}_p\, w(z^M) = \mathbb{E}_q\, w(z^M).$$

Yet this is precisely the crucial "Pareto indifference" hypothesis of Harsanyi's (1955) utilitarian theorem. Indeed, given the Cartesian product structure of the social consequence domain $Z^M = \prod_{i \in M} Z_i$, there is not even any need here to amend Harsanyi's original proof. The implication of this theorem is that there must exist an additive constant α and a set of multiplicative constants δ_i ($i \in M$) such that

$$w(z^M) \equiv \alpha + \sum\nolimits_{i \in M} \delta_i\, w_i(z_i).$$

Moreover, individual welfarism implies that maximizing $\mathbb{E}_p\, w(z^M)$ must be equivalent to maximizing $\mathbb{E}_{p_i}\, w_i(z_i)$ in any i-decision tree, where p_i is the marginal distribution $\text{marg}_{Z_i}\, p$. So $w(z^M)$ and $w_i(z_i)$ must be cardinally equivalent when $w(z^M)$ is regarded as a function of z_i alone. This evidently implies that each constant δ_i ($i \in M$) is actually positive. Then, however, since the individual and social welfare functions are only unique up to a cardinal equivalence class, for each $i \in M$ we can replace the individual welfare function $w_i(z_i)$ by the cardinally equivalent function $\tilde{w}(z_i) := \delta_i\, w_i(z_i)$, and the social welfare function $w(z^M)$ by the cardinally equivalent function $\tilde{w}(z^M) := w(z^M) - \alpha$. The result is that

$$\tilde{w}(z^M) = w(z^M) - \alpha = \sum\nolimits_{i \in M} \delta_i\, w_i(z_i) = \sum\nolimits_{i \in M} \tilde{w}_i(z_i)$$

and so one is back to simple addition of individual "utilities," once these have all been suitably normalized. Because of this possible normalization, I shall assume in future that

$$w(z^M) \equiv \sum\nolimits_{i \in M} w_i(z_i).$$

Note, however, that these utilities are by no means the same as those in other more traditional versions of utilitarianism. They are merely representations of appropriate ethical decisions in individualistic decision trees, without any necessary relationship to classical or other concepts of utility such as happiness, pleasure, absence of pain, preference satisfaction, etc. This is a major difference from Harsanyi's (1955) utilitarian theory. On the other hand, as in that theory, the additive structure arises because independence as regards revealed preferences over lotteries is combined with independence of each individual's welfare norm from the (fixed) lotteries faced by all other individuals. The latter independence property is an implication of individual welfarism, as defined in Section 9.

11. Variable Population

So far the set of individuals M has been treated as though it were fixed. Yet many ethical issues surround decisions which affect the set of individuals who come into existence — both the number and the composition of the set M. Thus, it would seem that M itself should be treated as variable consequence along with z^M, as indeed it was in Hammond (1988b).

A simpler alternative, however, is to treat "non-existence" for any individual $i \in M$ as a particular personal characteristic $\theta_i^0 \in \Theta_i$ which i could have, and then to define M as the set of all potential rather than actual individuals. Thus M is divided into the two sets $M^* := \{\, i \in M \mid \theta_i \neq \theta_i^0 \,\}$ of actual individuals who do come into existence, and $M^0 := \{\, i \in M \mid \theta_i = \theta_i^0 \,\}$ of individuals whose potential existence never comes about in practice. Actually, not much is lost by doing this. Assuming that only a finite number of individuals can ever be born before the world comes to an end (as seems quite reasonable, despite economists' models of steady state growth etc.), one can regard each $i \in M$ as just an integer used to number each individual who comes into existence, more or less in order of date of birth. Everything that is really relevant about an individual i, including date of birth, can be included in i's individual characteristic θ_i. Thus every individual who is ever born certainly gets numbered. But we can also consider the (finite) maximum number N^* of individuals who could ever be born, and then let $M = \{\, 1, 2, \ldots, N^* \,\}$. Then, unless all N^* individuals do actually come into existence, there will be "unused" numbers which refer to potential rather than actual individuals.

For those individuals $i \in M^0$ who never come into existence, the concept of individual welfare hardly makes any sense. In decision-theoretic terms, this means that non-existent individuals are not affected by social decisions — all social decisions are the same to them (except for decisions causing them to come into existence, of course). Consider now any i-decision tree T with the property that $p \in F_i(T)$ only if $p\,(A_i \times \{\theta_i^0\}) = 1$ — i.e., the probability of i not existing is always 1, no matter what decision is taken in the tree T. Since all decisions in T are the same to this certainly non-existent individual, it follows that at all decision nodes n in this tree, individual i's welfare norm $\beta_i(T, n)$ can be allowed to make any decisions on i's behalf. Therefore $\beta_i(T, n) = N_{+1}(n)$ and so $\Phi_{\beta_i}(T) = F_i(T)$ whenever T is an i-decision tree like this. What this means is that for some constant w_i^0 the individual welfare function $w_i(z_i)$ should satisfy $w_i(a_i, \theta_i^0) = w_i^0$ for all $a_i \in A_i$. Thus w_i^0 can be regarded as the constant "utility of non-existence," which is entirely independent of the social state or any aspect of any social consequence in which i never exists.

Assuming this to be true, one more useful normalization of individuals' welfare functions is possible. We can replace each $w_i(z_i)$ by the cardinally equivalent function

$$\tilde{w}_i(z_i) := w_i(z_i) - w_i^0$$

because a constant is merely being subtracted. Then, of course, $\tilde{w}_i(a_i, \theta_i^0) = 0$ for all $a_i \in A_i$. This implies that $\tilde{w}_i(z_i) = 0$ whenever $i \in M^0$.

We can also replace $w(z^M) \equiv \sum_{i \in M} w_i(z_i)$ by

$$\tilde{w}(z^M) \equiv w(z^M) - \sum\nolimits_{i \in M} w_i^0$$

for exactly the same reason. Then, however,

$$\tilde{w}(z^M) \equiv \sum\nolimits_{i \in M} [w_i(z_i) - w_i^0] \equiv \sum\nolimits_{i \in M} \tilde{w}_i(z_i) \equiv \sum\nolimits_{i \in M^*} \tilde{w}_i(z_i)$$

where M^* is the set of individuals who come into existence. So only individuals in the set M^* need be considered when adding all individuals' welfare levels.

Once again, it will be assumed from now on that this normalization has been carried out. Therefore one has

$$w(a^M, \theta^M) = w(z^M) \equiv \sum\nolimits_{i \in M^*} w_i(z_i) = \sum\nolimits_{i \in M^*} w_i(a_i, \theta_i)$$

where $M^* = \{ i \in M \mid \theta_i \neq \theta_i^0 \}$.

Note that this is formally identical to classical utilitarianism. But, as pointed out in Section 10, the resemblance is only formal because the individual welfare functions $w_i(z_i)$ mean something quite different. In particular, the zero level of this function is, by its very construction, just that level of individual welfare at which it is ethically appropriate for the individual to come into existence. This does much to dilute the strength of Parfit's (1984) "repugnant conclusion," which is that classical utilitarianism recommends creating very many extra individuals who are barely able to live above a subsistence level set so low that anyone who was forced to live below it would prefer not to have been born at all. Here we can escape the repugnant conclusion because there is nothing to prevent the ethical values embodied in the function $w_i(z_i)$ from making w_i positive only if individual i would actually be quite well off if allowed to come into existence. The fact that the personal consequence z_i makes individual i glad to be alive is not enough by itself to make $w_i(z_i)$ positive.

Note too that having $w_i(z_i)$ positive would only be a sufficient condition on its own for wanting i to exist if i's existence could somehow be brought about without interfering with anybody else. Yet children cannot exist without having (or having had) parents. So the personal benefits (and costs) to i of coming into existence have to be weighed against any costs (or benefits) to other individuals, especially i's parents, etc. Some further discussion of such issues occurs in Hammond (1988b).

12. Interpersonal Comparisons

In the introduction I criticized Sen's social welfare functional approach to social choice theory on the grounds that it never made explicit the interpersonal comparisons on which it was based. It is now my duty to explain how the utilitarian theory expounded above remedies this defect.

In fact, as pointed out in Hammond (1991b), there are interpersonal comparisons embodied in the social welfare function $w(z^M) = \sum_{i \in M^*} w_i(z_i)$. These comparisons also meet Myerson's (1985) criticism that interpersonal comparisons have no decision-theoretic significance. For $w_h(z_h) > w_i(z_i)$ means that society is better off creating individual h with personal consequence z_h rather than individual i with personal consequence z_i. And $w_h(z_h) - w_h(z'_h) < w_i(z'_i) - w_i(z_i)$, which is of course equivalent to

$w_h(z_h) + w_i(z_i) < w_h(z'_h) + w_i(z'_i)$, really does mean that moving h from z_h to z'_h and i from z_i to z'_i produces a benefit to society (if nobody else is affected) because any loss to h is outweighed by the gain to i, or *vice versa*. Indeed, even welfare ratios acquire meaning. For $w_h(z_h)/w_i(z_i)$ can be regarded as the marginal rate of substitution between individuals like h facing personal consequence z_h and individuals like i facing personal consequence z_i. If this ratio is greater than 1, for instance, then society could gain by creating more individuals like h and fewer like i.

Thus we have a "cardinal ratio scale" measure of individual welfare, with "cardinal full comparability" of both welfare levels and differences, as well as a clearly defined zero level of welfare. Yet, of all the social welfare functionals considered by Roberts (1980) which have this property, only the simple sum is ethically appropriate — according to the theory expounded above. So the social welfare functional is no longer left indeterminate, as usually happens in this approach to social choice theory. Of course, this extra determinacy of the functional form comes at a price, since now all the indeterminacy has been displaced into the individual welfare function. In Sen's version of the theory, as well as in Harsanyi's version of utilitarianism, one could argue that the utility measure had some objective reality which was independent of any particular values; here the individual welfare measure has instead been defined so that it embodies certain relevant ethical values.

13. Concluding Assessment

In the introductory Section 1 there were four questions which, I suggested, any social choice theoretic approach to ethics should be able to answer. The reader is invited to review them once again, along with the succeeding paragraph which summarized the (admittedly still incomplete) answers that I promised to provide.

The theory has been based upon the idea of a social behaviour norm, defined for decision trees in which multi-stage decisions and compound lotteries combine to give random social consequences. The social consequences to be considered amount to interpersonal profiles of personal consequences, with each personal consequence summarizing everything which is ethically relevant to the corresponding individual. There should be a social norm for society as a whole, and also individual welfare norms specifying what decisions are ethically acceptable when there is only one individual whose personal consequences can be affected

by the decision. The relevant individual's welfare norm should coincide with the social norm in all such "one person situations."

Four axioms were imposed upon such behaviour norms. The first was unrestricted domain, requiring the norm to be defined for all finite decision trees in which only positive probabilities could occur at any chance node. The second was consistency in continuation subtrees, requiring the same behaviour to be acceptable at any decision node of a tree regardless of whether that node is regarded as belonging to the whole tree or to only a continuation subtree. The third "consequentialist" axiom was the crucial one — that knowledge of the feasible set of consequence lotteries arising from a decision tree would always suffice to determine the "revealed choice set" of all consequence lotteries which could possibly result from recommended behaviour. Later on this was further strengthened so that, in society, only the profile of marginal probability distributions over different individuals' personal consequences was relevant to recommended behaviour. A fourth technical axiom of continuity was finally imposed which had the effect of ruling out lexicographic preferences.

From these four axioms, applied to the social norm and so to the individual welfare norms which represent the social norm in one person situations, there follows a form of classical utilitarianism which is close in spirit to that of Harsanyi (1955), even if the interpretation of individual utilities is quite different. This gives a unique cardinal equivalence class of "von Neumann-Morgenstern social welfare functions" whose expected values society should maximize. Moreover, each such social welfare function can be expressed as the sum of suitably normalized individual welfare functions whose expected values it is right to maximize in one person situations where only one individual's personal consequence lottery is affected by any decision that could be made.

Finally, on the assumption that decisions affecting individuals who never come into existence regardless of what decision is made do not matter, and so can be made arbitrarily, it was shown that individuals' welfare levels could be normalized to zero for individuals who never exist. Then it is enough to sum the welfare levels of all individuals who do come into existence, and even to use this criterion to determine who should come into existence — exactly as with classical utilitarianism. The obvious defects of that ethical theory are avoided, however, by interpreting individual welfare as a purely ethical concept representing

what behaviour *ought* to maximize in decision trees affecting the personal consequences of only one individual.

In the end, the theory that has been sketched here determines completely the formal structure of the ethical decision criterion. Yet this formal structure remains an empty shell, to be completed by substantive ethical statements concerning what should count as an ethically relevant consequence, and even about what decisions really would be right in certain practical decision problems. The real work in ethics may only just be beginning. It promises to be much more interesting than the rather dry arguments about whether or not utilitarianism or consequentialism are appropriate. I have set out such arguments here only to explain why I personally regard the issue as virtually settled.

REFERENCES

Anscombe, G.E.M. (1958), "Modern Moral Philosophy," *Philosophy*, **33**: 1–19.

Arrow, K.J. (1951, 2nd edn., 1963), *Social Choice and Individual Values*. New York: John Wiley.

Arrow, K.J. (1977), "Extended Sympathy and the Possibility of Social Choice," *American Economic Review, Papers and Proceedings*, **67**: 219–25 reprinted as Chapter 11 of Arrow (1983).

Arrow, K.J. (1983), *Collected Papers of Kenneth J. Arrow, 1: Social Choice and Justice*. Cambridge, Mass.: The Belknap Press of Harvard University Press.

d'Aspremont, C. (1985), "Axioms for Social Welfare Orderings," in *Social Goals and Social Organization: Essays in Memory of Elisha Pazner* edited by L. Hurwicz, D. Schmeidler and H. Sonnenschein (Cambridge: Cambridge University Press), ch. 1, pp. 19–76.

d'Aspremont, C. and L. Gevers (1977), "Equity and the Informational Basis of Collective Choice," *Review of Economic Studies*, **44**: 199–209.

Bewley, T. (1989), *Knightian Uncertainty*. Evanston, IL: Northwestern University.

Black, D. (1948), "On the Rationality of Group Decision-Making," *Journal of Political Economy*, **56**: 23–34.

Blackorby, C., D. Donaldson and J.A. Weymark (1984), "Social Choice with Interpersonal Utility Comparisons: A Diagrammatic Introduction," *International Economic Review*, **25**: 327–56.

de Borda, J.-C. (1781), "Mémoire sur les élections au scrutin," *Mémoires de l'Académie Royale des Sciences,* pp. 657–65; translated by A. de Grazia (1953) in *Isis*, **44**: 42–52.

Condorcet, M.J.A.N.C., Marquis de (1785), *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.

Dodgson, C.L. (1884, 5), *The Principles of Parliamentary Representation*. London: Harrison.

Foley, D.K. (1970), "Lindahl's Solution and the Core of an Economy with Public Goods," *Econometrica*, **38**: 66–72.

Hammond, P.J. (1976), "Equity, Arrow's Conditions, and Rawls' Difference Principle," *Econometrica*, **44**: 793–804.

Hammond, P.J. (1988a), "Consequentialist Foundations for Expected Utility," *Theory and Decision*, **25**: 25–78.

Hammond, P.J. (1988b), "Consequentialist Demographic Norms and Parenting Rights," *Social Choice and Welfare*, **5**: 127–145.

Hammond, P.J. (1988c), "Orderly Decision Theory: A Comment on Professor Seidenfeld," *Economics and Philosophy*, **4**: 292–297.

Hammond, P.J. (1989), "Consistent Plans, Consequentialism, and Expected Utility," *Econometrica*, **57**: 1445-1449.

Hammond, P.J. (1991a), "Independence of Irrelevant Interpersonal Comparisons," *Social Choice and Welfare*, **8**: 1–19.

Hammond, P.J. (1991b), "Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made," in *Interpersonal Comparisons of Well-Being* edited by J. Elster and J.E. Roemer (Cambridge: Cambridge University Press), ch. 7, pp. 200–254.

Hammond, P.J. (1991c), "Harsanyi's Utilitarian Theorem: A Simpler Proof and Some Ethical Connotations," European University Institute, Working Paper ECO 91/32; to appear in R. Selten (ed.) *Essays in Honour of John Harsanyi* (Springer-Verlag).

Harsanyi, J.C. (1953), "Cardinal Utility in Welfare Economics and the Theory of Risk-Taking," *Journal of Political Economy*, **61**: 434–35.

Harsanyi J.C. (1955), "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, **63**: 309–21.

Harsanyi, J.C. (1976), *Essays in Ethics, Social Behaviour, and Scientific Explanation*. Dordrecht: Reidel.

Harsanyi, J.C. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Harsanyi, J.C. (1978), "Bayesian Decision Theory and Utilitarian Ethics," *American Economic Review, Papers and Proceedings*, **68**: 223–8.

Herstein, I.N. and J. Milnor (1953), "An Axiomatic Approach to Measurable Utility," *Econometrica*, **21**: 291–297.

Levi, I. (1974), "On Indeterminate Probabilities," *Journal of Philosophy*, **71**: 391–418.

Levi, I. (1980), *The Enterprise of Knowledge*. Cambridge, Mass.: M.I.T. Press.

Levi, I. (1986), *Hard Choices: Decision Making under Unresolved Conflict*. Cambridge: Cambridge University Press.

McClennen, E.F. (1990), *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

Machina, M.J. (1989), "Dynamic Consistency and Non-Expected Utility Models of Choice under Certainty," *Journal of Economic Literature*, **27**: 1622–68.

Milleron, J.C. (1972), "Theory of Value with Public Goods: A Survey Article," *Journal of Economic Theory*, **5**: 419–477.

Moore, G.E. (1912), *Ethics*. London: Home University Library, Williams and Norgate.

Myerson, R.B. (1985), "Bayesian Equilibrium and Incentive Compatibility: An Introduction," in *Social Goals and Social Organization* edited by L. Hurwicz, D. Schmeidler and H. Sonnenschein (Cambridge: Cambridge University Press), ch. 8, pp. 229–259.

Parfit, D. (1987), *Reasons and Persons*. Oxford: Oxford University Press.

Regan, D.H. (1983), "Against Evaluator Relativity: A Response to Sen," *Philosophy and Public Affairs*, **12**: 93–112.

Roberts K.W.S. (1980), "Possibility Theorems with Interpersonality Comparable Welfare Levels," *Review of Economic Studies*, **47**: 409–20.

Savage, L.J. (1954; 2nd revised edn., 1972), *The Foundations of Statistics*. New York: John Wiley and Dover.

Seidenfeld, T. (1988a), "Decision Theory without 'Independence' or without 'Ordering': What is the Difference," *Economics and Philosophy*, **4**: 267–290.

Seidenfeld, T. (1988b), "Rejoinder," *Economics and Philosophy*, **4**: 309–315.

Selten, R. (1965), "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit [Game Theoretical Treatment of an Oligopoly Model with Demand Inertia]," *Zeitschrift für die gesamte Staatswissenschaft*, **121**: 301–324 and 667–689.

Sen, A.K. (1970a), *Collective Choice and Social Welfare*. San Francisco: Holden Day.

Sen, A.K. (1970b), "Interpersonal Aggregration and Partial Comparability," *Econometrica*, **38**: 393–409; reprinted with correction in Sen (1982a).

Sen, A.K. (1977), "On Weights and Measures: Informational Constraints in Social Welfare Analysis," *Econometrica*, **45**: 1539–72.

Sen, A.K. (1982a), *Choice, Welfare and Measurement*. Oxford: Basil Blackwell and Cambridge, Mass.: M.I.T. Press.

Sen, A.K. (1982b), "Rights and Agency," *Philosophy and Public Affairs*, **11**: 3–39.

Sen, A.K. (1983), "Evaluator Relativity and Consequential Evaluation," *Philosophy and Public Affairs*, **12**: 113–132.

Sen, A.K. (1987), *On Ethics and Economics*. Oxford: Basil Blackwell.

Sen, A.K. and B.A.O. Williams (Eds.) (1982), *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Strotz, R.H. (1956), "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, **23**: 165-80.

Suppes, P. (1966), "Some Formal Models of Grading Principles," *Synthese*, **6**: 284–306.

Warnock, M. (Ed.) (1962), *Utilitarianism: Selections from the Writings of Jeremy Bentham, John Stuart Mill, and John Austin*. London: Collins, the Fontana Library in Philosophy.

Williams, B.A.O. (1973), "A Critique of Utilitarianism," in *Utilitarianism for and against* edited by J.J.C.Smart and B.A.O. Williams (Cambridge: Cambridge University Press).

Williams, B. (1985), *Ethics and the Limits of Philosophy*. London: Collins, the Fontana Library in Philosophy.

Zermelo, E. (1913), "Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels [On an Application of Set Theory to the Theory of the Game of Chess]," *Proceedings of the Fifth International Conference of Mathematicians*, **2**: 501–504.