Harsanyi's Utilitarian Theorem: A Simpler Proof and Some Ethical Connotations

PETER J. HAMMOND, Department of Economics
European University Institute, Badia Fiesolana
50016 S. Domenico di Fiesole (FI), Italy;
and Stanford University, CA 94305–6072, U.S.A.

ABSTRACT

Harsanyi's utilitarian theorem states that the social welfare function is the weighted sum of individuals' utility functions if: (i) society maximizes expected social welfare; (ii) individuals maximize expected utility; (iii) society is indifferent between two probability distributions over social states whenever all individuals are. After giving a simpler proof, an alternative axiomatic foundation for Vickrey-Harsanyi utilitarianism is provided. By making using an extended version of Harsanyi's concept of a player's "type" in the theory of games with incomplete information, the problem of forming social objectives when there is incomplete information can also be resolved, at least in principle.

1. Introduction

Gabriel Cramer (1728) and then Daniel Bernouilli (1738, 1954) first proposed as a decision criterion the maximization of expected utility rather than of expected wealth. Much later, in an appendix to their classic work, von Neumann and Morgenstern (1943) set out for the first time an axiomatic justification for this criterion. They were also the first to appreciate fully how their theory provided a cardinal concept of utility — i.e., one that is unique up to linear (or affine) transformations of the utility function. This was in contrast to the ordinal concept of utility which is unique up to general increasing transformations.

Very shortly thereafter, Lerner (1944) showed that expected social welfare could be maximized, under certain symmetry conditions, by equalizing income — see Sen (1969, 1973) for futher discussion of this result. Perhaps more important, however, was Vickrey's (1945) realization that this von Neumann-Morgenstern cardinalization could be used to measure marginal utility in a way that relates to statements about what redistributions of income would be desirable. Of course, for any one individual's cardinal utility function, it is true that measures of that individual's marginal utility for different levels of income are all uniquely determined up to a single multiplicative constant. This use of the cardinal utility function was contested by Friedman and Savage (1952), which led in turn to Harsanyi's (1953) comment on their paper. It was in this comment that Harsanyi first enunciated his idea of "impersonality," according to which ethical decisions should be based upon the interests of persons who have had all personal biases removed by being put in a situation of complete uncertainty about their true identity.

Also in the early 1950's, a paper by Fleming (1952) appeared which advocated that the social welfare function should be additively separable over individuals, over time periods, and also over uncertain states of the world. This prompted Harsanyi (1955) to expand his idea of impersonality into a complete new theory. In part, Harsanyi adapted Lerner's idea significantly and considered, several years before Rawls (1959, 1971),

an "original position" in which all individuals are supposed to choose the social state they prefer without knowing which members of the society they will become upon emerging from behind what Rawls so aptly called the "veil of ignorance." Unlike Rawls, however, Harsanyi has always stuck to the orthodox theory of choice under uncertainty — namely, the existence of subjective probabilities, and then the maximization of expected utility. Of course, in the original position, symmetry was postulated, so that there was an equal probability of becoming each individual in the society. Later, Harsanyi (1975a, 1975b) emphasized this crucial difference from what Rawls called the "difference principle." These articles by Harsanyi appeared after Rawls' theory had acquired its great popularity — which it fully deserved, though mostly for reasons having little to do with the difference principle *per se*. Indeed, the version of Harsanyi (1975a) which is reprinted in Harsanyi (1976) contains an additional section responding to Rawls' (1974) more thorough discussion of his reasons for using the difference principle rather than an expected social welfare criterion.

A key step in Harsanyi's (1955) argument was the claim that expected social welfare would be the weighted sum of expected individual utility functions, assuming that whenever all individuals are indifferent between any two probability distributions over social states, then so is society. Strictly speaking, Harsanyi's justification for this claim relied on some implicit assumptions concerning possible variations in individuals' expected utility levels — assumptions similar to those which were also made in Hammond (1983). This was first pointed out by Domator (1979), it seems, who, along with a number of other authors more recently, have given rigorous proofs without such additional assumptions — see especially Border (1985), Coulhon and Mongin (1989), and also Broome (1990). Section 2 below will present what I believe to be an equally rigorous, but rather simple proof. For the case of a finite number of social states, this proof uses an elementary result in linear algebra which can be found, for instance, in Gale (1960). The idea of using this kind of result is due to Border (1981), which was a privately circulated precursor to Border (1985). Very similar proofs for this special case can also be found in Selinger (1986) and Weymark (1990). For the general case of an infinite number of social states, the proof presented here relies only on the finite intersection property of compact sets.

For too long a time Harsanyi's approach was not very widely appreciated, and even today remains controversial. Fleming (1957), Diamond (1967), and Pattanaik (1968) made relatively early criticisms. Diamond's criticism, which Sen (1970) also expressed, and to which Harsanyi (1975b) contains a response, was that maximizing expected social welfare could produce unacceptable inequalities of utility. Yet it is not clear what these inequalities really signify until we give "utility" some concrete meaning; once we do, the criticism essentially loses its force, as Broome (1989) in particular has pointed out.

Pattanaik's concern was more with Harsanyi's original position argument, and the claim that a better understanding of individual psychology was likely to bring us closer to a social welfare function that all could agree to. In fact, despite Harsanyi's serious attempts to argue otherwise, it seems all too likely that different individuals, even with a perfect understanding of psychology, and even behind an apparently common veil of ignorance, would still retain their different views about what other individuals' attitudes to risk are likely to be, and about how to weight the von Neumann-Morgenstern utility functions of different individuals which represent these attitudes to risk. Oddly enough, a similar debate surrounds the assumption of Harsanyi (1967–8) and Aumann (1987) regarding the existence of common prior beliefs in game theory.

Given these and other problems with original position arguments, Section 3 suggests a procedure for side-stepping the issue entirely. The argument is actually no more than a summary, and perhaps a clearer presentation, of ideas discussed more extensively in Hammond (1987). Indeed, those ideas build on or relate to Hammond (1983, 1986, 1988a, b, c) and the realization that a new "consequentialist" framework, based on analysing behaviour in decision trees, could also help to justify the axioms behind conventional expected utility theory. This is really the reason why I find Harsanyi's fundamental work so relevant to ethical decision making.

Finally, Section 4 discusses a natural extension of the previous formulation to societies in which there is incomplete information about individuals' true utilities and other features relevant to a proper ethical decision. Adapting Harsanyi's (1967–8) key insight regarding games of incomplete information, it becomes clear that one needs to consider not just social states in the usual sense, but *contingent* social states which depend on different individuals' types. These are closely related to the "game forms" which Sugden (1985, 1986) has argued, in opposition to Sen, are the right way of modelling individual rights — see also Gaertner, Pattanaik and Suzumura (1988) and Riley (1989, 1990).

2. Proof of Harsanyi's Theorem

Let X be the space of social states, which is assumed to be a (Borel) measurable set with σ-algebra $\mathcal{X}$. Let $\mathcal{M}(X)$ be the set of probability measures on X with this σ-algebra. Suppose that each individual i in the membership M (a finite set) has a welfare ordering $\succsim_i$ on $\mathcal{M}(X)$ represented by the expected value $\mathbb{E}_\mu v_i(x)$ with respect to $\mu \in \mathcal{M}(X)$ of the von Neumann-Morgenstern utility function (NMUF) $v_i : X \to \Re$. Suppose too that there is a social ordering $\succsim$ on $\mathcal{M}(X)$ which is represented by the expected value $\mathbb{E}_\mu w(x)$ of a "von Neumann-Morgenstern Bergson social welfare function" $w : X \to \Re$. Finally, suppose that Pareto indifference is satisfied — i.e., that

$$[\forall i \in M : \mu_i \sim_i \nu_i] \Longrightarrow \mu \sim \nu,$$

or equivalently, that

$$[\forall i \in M : \mathbb{E}_\mu v_i = \mathbb{E}_\nu v_i] \Longrightarrow \mathbb{E}_\mu w = \mathbb{E}_\nu w$$

for all pairs $\mu, \nu \in \mathcal{M}(X)$. Then *Harsanyi's theorem* is the implication that there exist *welfare weights* ω_i ($i \in M$) and an additive constant α such that

$$w(x) \equiv \alpha + \sum\nolimits_{i \in M} \omega_i \, v_i(x) \tag{1}$$

on X. This theorem was proved by Harsanyi (1955). As pointed out in the introduction, however, there were a number of unnecessary implicit assumptions concerning how possible variations in the social state x could lead to entirely independent variations in the value of each individual's utility $v_i(x)$.

The following proof considers first the case when $X = A$, a finite set consisting of $\#A$ members. Then the argument uses ideas similar to those in Border (1981). The (new) proof for a general measurable space $(X, \mathcal{X})$ follows later.

Proof (when $X = A$, a finite set).

Let $\eta(x)$ ($x \in A$) be any set of $\#A$ real numbers satisfying

$$\sum\nolimits_{x\in A} \eta(x) = 0 \quad \text{and} \quad \sum\nolimits_{x\in A} \eta(x)\, v_i(x) = 0 \quad (\text{all } i \in M). \tag{2}$$

For all $x \in A$, define

$$\mu(x) := (1/\#A) + \lambda\, \eta(x); \quad \nu(x) := (1/\#A) - \lambda\, \eta(x) \tag{3}$$

where $\lambda > 0$ is small enough to ensure that $\mu(x), \nu(x) \geq 0$ for all $x \in A$. Then $\mu, \nu \in \mathcal{M}(A)$ and also

$$\mu(x) - \nu(x) = 2\lambda\, \eta(x) \quad (\text{all } x \in A). \tag{4}$$

It follows that

$$\mathbb{E}_\mu\, v_i(x) - \mathbb{E}_\nu\, v_i(x) = \sum\nolimits_{x\in A} [\mu(x) - \nu(x)]\, v_i(x) = 2\lambda \sum\nolimits_{x\in A} \eta(x)\, v_i(x) = 0 \tag{5}$$

for every $i \in M$. So, by the assumption of Pareto indifference,

$$0 = \mathbb{E}_\mu\, w(x) - \mathbb{E}_\nu\, w(x) = \sum\nolimits_{x\in A} [\mu(x) - \nu(x)] w(x) = 2\lambda \sum\nolimits_{x\in A} \eta(x) w(x). \tag{6}$$

Thus any $\#A$-vector $\eta(x)$ ($x \in A$) satisfying (2) above also satisfies

$$\sum\nolimits_{x\in A} \eta(x)\, w(x) = 0. \tag{7}$$

Consider now any row vector η with elements $\eta(x)$ ($x \in A$). Suppose that η lies in the null space of the $\#A \times (\#M + 1)$ matrix V whose elements in the first column are all ones and whose elements in the last $\#M$ columns are $v_i(x)$ ($x \in A, i \in M$). Then $\eta\, V = 0$, which implies that η satisfies (2) above, and so we have just shown that it must also satisfy (7). From the *Solvability Theorem* due to Gale (1960, p. 41), it follows that the column vector w with components $w(x)$ ($x \in A$) is spanned by the columns of V matrix. So there must exist constants α and ω_i ($i \in M$) for which $w = V \binom{\alpha}{\omega^M}$ or

$$w(x) \equiv \alpha + \sum\nolimits_{i\in M} \omega_i\, v_i(x) \quad (\text{all } x \in A) \tag{1$'$}$$

as required. ∎

The above result can now be used to prove Harsanyi's theorem for a general (possibly infinite) measurable set X as follows.

Proof (when X is an infinite set).

First, reduce M if necessary to a subset M^* with the linear independence property that the only solution to the equation

$$0 \equiv \alpha + \sum\nolimits_{i \in M^*} \omega_i \, v_i(x) \quad (\text{all } x \in X) \tag{8}$$

in the unknown constants α and ω_i ($i \in M^*$) is the trivial solution with $\alpha = \omega_i = 0$ (all $i \in M^*$). This is possible because, if the identity

$$0 \equiv \alpha + \sum\nolimits_{i \in M} \omega_i \, v_i(x) \quad (\text{all } x \in X) \tag{9}$$

has a non-trivial solution in α and ω_i ($i \in M$), then there must be at least one $j \in M$ for which $\omega_j \neq 0$. But this implies that the corresponding function $v_j(x)$ can be expressed as the linear combination

$$v_j(x) \equiv \left(-\frac{\alpha}{\omega_j}\right) + \sum\nolimits_{i \in M \setminus \{j\}} \left(-\frac{\omega_i}{\omega_j}\right) v_i(x) \quad (\text{all } x \in X) \tag{10}$$

of a constant term and of all the other functions $v_i(x)$ ($i \neq j$). If the linear independence property (8) is still not satisfied even by the reduced set $M^* := M \setminus \{j\}$, then M^* can be reduced still further in this way as many times as necessary, until M^* does finally satisfy the linear independence property. After this reduction has been completed, for every $j \in M \setminus M^*$ it will be true that the function $v_j(x)$ can be expressed as a linear combination

$$v_j(x) \equiv \alpha'_j + \sum\nolimits_{i \in M^*} \omega'_{ji} \, v_i(x) \quad (\text{all } x \in X) \tag{11}$$

for some constants α'_j and ω'_{ji} ($i \in M^*$).

For any finite $A \subset X$, define the set

$$\begin{aligned} \Omega(A) := \{\, (\lambda, \alpha, \omega^{M^*}) \in \Re^{\#M^*+2} \mid \forall x \in A : \lambda \, w(x) \equiv \alpha + \sum\nolimits_{i \in M^*} \omega_i \, v_i(x) \\ \text{and} \quad \lambda^2 + \alpha^2 + \sum\nolimits_{i \in M^*} \omega_i^2 = 1 \,\}. \end{aligned} \tag{12}$$

Since Harsanyi's theorem has just been proved for the case when A is finite, there certainly exist both an $\#M$-dimensional vector $\tilde{\omega}^M \in \Re^M$ and a constant $\tilde{\alpha} \in \Re$ such that

$$w(x) \equiv \tilde{\alpha} + \sum\nolimits_{i \in M} \tilde{\omega}_i \, v_i(x) \quad (\text{all } x \in A). \tag{13}$$

But then (11) implies that in fact there must exist a different $\#M^*$-dimensional vector $\omega^{M^*} \in \Re^{M^*}$ and a constant $\alpha \in \Re$ such that

$$w(x) \equiv \alpha + \sum\nolimits_{i \in M^*} \omega_i \, v_i(x) \quad (\text{all } x \in A). \tag{14}$$

Now just multiply each side of (14) by $\lambda := \left(1 + \alpha^2 + \sum_{i \in M^*} \omega_i^2\right)^{-\frac{1}{2}}$ in order to ensure that the new normalized coefficients lie on the surface of the unit sphere: the result makes it evident that the set $\Omega(A)$ is non-empty for every finite $A \subset X$.

Notice how, for $A \subset X$, the set $\Omega(A) = \cap_{x \in A} \, \Omega(\{x\})$. Now, for each $x \in X$, the set $\Omega(\{x\})$ is evidently compact because it is a closed subset of the surface of the unit sphere in $\Re^{\#M^*+2}$. But when A is finite, it has just been shown that the intersection $\Omega(A) = \cap_{x \in A} \, \Omega(\{x\})$ must be non-empty. By the finite

intersection property for arbitrary families of compact sets, it follows that the intersection $\cap_{x \in X} \Omega(\{x\})$ is non-empty. So there exists some combination

$$(\lambda, \alpha, \omega^{M^*}) \in \cap_{x \in X} \Omega(\{x\}) \subset \Re^{\#M^*+2}. \tag{15}$$

By definition of $\Omega(\{x\})$, this combination must satisfy

$$\lambda\, w(x) \equiv \alpha + \sum\nolimits_{i \in M^*} \omega_i\, v_i(x) \quad (\text{all } x \in X) \tag{16}$$

and also

$$\lambda^2 + \alpha^2 + \sum\nolimits_{i \in M^*} \omega_i^2 = 1. \tag{17}$$

Now, if it were true that $\lambda = 0$, then (8) would be satisfied for some constants satisfying $\alpha^2 + \sum_{i \in M^*} \omega_i^2 = 1$, which therefore cannot all be zero. This would contradict the construction of the reduced set M^* which has to satisfy the linear independence property that (8) has only a trivial solution. So $\lambda \neq 0$. One can therefore divide each side of (16) by λ in order to obtain

$$w(x) \equiv \bar{\alpha} + \sum\nolimits_{i \in M^*} \bar{\omega}_i\, v_i(x) \equiv \bar{\alpha} + \sum\nolimits_{i \in M} \bar{\omega}_i\, v_i(x) \quad (\text{all } x \in X) \tag{18}$$

for the newly defined constants

$$\bar{\alpha} := \alpha/\lambda \quad \text{and} \quad \bar{\omega}_i := \begin{cases} \omega_i/\lambda & \text{if } i \in M^*; \\ 0 & \text{if } i \in M \setminus M^*. \end{cases} \tag{19}$$

This completes the proof. ∎

3. Social Welfare, Personal Issues, and Individual Welfare

This section will provide an alternative motivation to that of Harsanyi's original position for his form of utilitarianism. Indeed, it will be claimed that his form of utilitiarianism is a logical implication of the standard "Bayesian" approach to decision-making under uncertainty, when this is combined with a plausible formalization of the key idea in individualistic ethics — namely, that actions should be judged by whether they produce desirable consequences for individuals.

The argument will rely upon the fiction that a different personalized version x_i of the social state $x \in X$ is possible for each different individual $i \in M$. This idea is very similar to the notion of personalized public goods, as exploited by Milleron (1972) and many successors. So the extended Cartesian product domain $X^M := \prod_{i \in M} X_i$ of possible *personalized social states* will be considered, in which each component space X_i ($i \in M$) is a copy of the space X. Of course, the constraint that $x_i = x$ for all $i \in M$ and for some single social state $x \in X$ will usually have to be observed in practice, just as all individuals are generally required to have the same bundle of public goods. Nevertheless, in thinking about society's objectives, it will be useful to contemplate what would be possible in the absence of this constraint.

Ethical decisions whose consequences lie in this space of personalized social states are assumed to satisfy the standard *Bayesian rationality* postulate. Thus it will be assumed that:

ASSUMPTION 1. *There exists some* **social welfare ordering** $\succsim$ *on the space* $\mathcal{M}(X^M)$ *which represents the relative ethical desirabilities of different uncertain social consequences. Moreover, there is a* **von Neumann-Morgenstern social welfare function** $W : X^M \to \Re$ *with the property that, for every pair of probability distributions* $\mu, \nu \in \mathcal{M}(X^M)$*, one has* $\mu \succsim \nu$ *if and only if the expected levels of welfare satisfy* $\mathbb{E}_\mu W \geq \mathbb{E}_\nu W$*.*

So far, $\mathbb{E}_\mu W$ could be an expected welfare function representing a purely collectivist ethic, paying no attention to individuals whatsoever. To capture the idea that it is only each individual's welfare which matters, two other assumptions will be made. Before they can be stated, however, some additional notation is needed. For any given joint probability distribution $\mu \in \mathcal{M}(X^M)$ over all the different individuals' personalized social states $x^M = \langle x_i \rangle_{i \in M}$, and any given individual $i \in M$, let $\mu_i \in \mathcal{M}(X_i)$ denote the marginal distribution $\text{marg}_{X_i} \mu$ of i's personalized social state x_i.

It will now be assumed that any stochastic dependence between the personalized consequences of different individuals is irrelevant to the welfare ordering, so that the social ordering depends only on the marginal distribution over the personalized consequences of each individual. Formally:

ASSUMPTION 2. *For all* $\mu, \nu \in \mathcal{M}(X^M)$*, assume that if* $\mu_i = \nu_i$ *for all* $i \in M$*, then* $\mu \sim \nu$*.*

Let $\mu^M := \prod_{i \in M} \mu_i$ denote the joint probability distribution of the vector x^M having the property that the distributions of the different individuals' personalized social states x_i ($i \in M$) are all independent, and equal to the appropriate marginal distributions μ_i. As an implication of Assumption 2 notice that $\mu \sim \mu^M$ for all $\mu \in \mathcal{M}(X^M)$. Indeed, it is always sufficient to consider the collection μ_i ($i \in M$) of marginal distributions, without worrying at all about any interdependence within the joint distribution μ of different personalized social states x_i ($i \in M$).

Finally, it will also be assumed that any ethically relevant external effects which are of consequence to individual i have already been included within the personalized social state x_i. In particular, it will be assumed that the ethically proper choice of the marginal distribution $\mu_i \in \mathcal{M}(X_i)$ of i's personalized social state x_i is completely unaffected by the joint distribution $\mu_{-i} \in \mathcal{M}(X^{M \setminus \{i\}})$ of the profile $x_{-i} \in X^{M \setminus \{i\}} := \prod_{j \in M \setminus \{i\}} X_j$ consisting of all the other individuals' personalized social states. Formally:

ASSUMPTION 3. *For each individual* $i \in M$*, there exists an ordering* $\succsim_i$*, called* **individual** i*'s* **ethical welfare ordering***, such that for every fixed distribution* $\bar{\mu}_{-i} \in \mathcal{M}(X^{M \setminus \{i\}})$ *and all pairs* $\mu_i, \nu_i \in \mathcal{M}(X_i)$*, one has*

$$\mu_i \succsim_i \nu_i \iff (\mu_i, \bar{\mu}_{-i}) \succsim (\nu_i, \bar{\mu}_{-i}).$$

Note how each individual's welfare ordering has essentially been derived from the social ordering over personal issues, rather than the social ordering being derived from all the different individuals' welfare orderings. Moreover, it is not explicitly assumed that each individual's welfare ordering can be represented by the expected value of some cardinal individual welfare function, even though this will turn out to be an implication of all the three assumptions together. This indirect approach to the definition of an individual's welfare ordering is similar to that of Broome (1987). At first it may seem excessively paternalistic. Yet

if each individual's "ethical welfare" preferences or "interests" get properly respected in constructing the social ordering $\succsim$ and so each individual i's welfare ordering $\succsim_i$, then this is entirely consistent with what was called "ethical liberalism" in Hammond (1987).

Because the social welfare ordering $\succsim$ on $\mathcal{M}(X^M)$ is represented by the expected value of the social welfare function $W(x^M)$, it follows that, for each fixed $\bar{x}_{-i} \in X^{M\setminus\{i\}}$, the individual welfare ordering $\succsim_i$ on $\mathcal{M}(X_i)$ is represented by the expected value of $W(x_i, \bar{x}_{-i})$. So, for each $i \in M$, there exists a unique cardinal equivalence class of individual welfare functions $w_i(x_i)$ with the property that

$$w_i(x_i) \equiv \alpha_i(\bar{x}_{-i}) + \beta_i(\bar{x}_{-i})\, W(x_i, \bar{x}_{-i})$$

for all $x_i \in X_i$ and all $\bar{x}_{-i} \in X^{M\setminus\{i\}}$, where $\alpha_i(\bar{x}_{-i})$ is an arbitrary real-valued function of $\bar{x}_{-i}$, and $\beta_i(\bar{x}_{-i})$ is a positive real-valued function of $\bar{x}_{-i}$. In future $w_i(x_i)$ will be used to denote any particular member of this equivalence class.

Assumptions 2 and 3 above now imply that, for all pairs $\mu, \nu \in \mathcal{M}(X^M)$ satisfying $\mu_i \sim_i \nu_i$ for all $i \in M$, it must be true that

$$(\mu^J, \nu^{M\setminus J}) = (\mu^{J\setminus\{j\}}, \mu_j, \nu^{M\setminus J}) \sim (\mu^{J\setminus\{j\}}, \nu_j, \nu^{M\setminus J}) = (\mu^{J\setminus\{j\}}, \nu^{M\setminus[J\setminus\{j\}]})$$

whenever $j \in J \subset M$. From this it follows easily by induction on the number of members in the set J that $\mu_i \sim_i \nu_i$ (all $i \in M$) implies $\mu \sim \nu$. We have therefore confirmed that

$$[\forall i \in M : \mathbb{E}_\mu w_i = \mathbb{E}_\nu w_i] \Longrightarrow \mathbb{E}_\mu W = \mathbb{E}_\nu W.$$

So, by Harsanyi's theorem which was proved in the last section, there must exist welfare weights ω_i $(i \in M)$ and an additive constant α such that

$$W(x^M) \equiv \alpha + \sum\nolimits_{i\in M} \omega_i\, w_i(x_i).$$

Indeed, for the special case being considered in this section, the device of personalized social states allows sufficient independent variations in different individuals' utilities for Harsanyi's (1955) original proof to be used.

Moreover, Assumption 3 in particular allows a stronger conclusion in this case. This is because, for every fixed $\bar{\mu}_{-i} \in \prod_{j\in M\setminus\{i\}} \mathcal{M}(X_j)$ and for all pairs $\mu_i, \nu_i \in \mathcal{M}(X_i)$, the weak preference $(\nu_i, \bar{\mu}_{-i}) \succsim (\mu_i, \bar{\mu}_{-i})$ would imply that $\nu_i \succsim_i \mu_i$ because of Assumption 3. Therefore, for all $i \in M$ it must actually be true that

$$\mu_i \succ_i \nu_i \Longrightarrow (\mu_i, \bar{\mu}_{-i}) \succ (\nu_i, \bar{\mu}_{-i}),$$

implying that all the welfare weights ω_i are strictly positive. Now, however, the cardinal individual welfare functions $w_i(x_i)$ can be re-normalized so that they become $\tilde{w}_i(x_i) := \omega_i\, w_i(x_i)$, and the social welfare function $W(x^M)$ can be replaced by $\tilde{W}(x^M) := W(x^M) - \alpha$. Accordingly, one has

$$\tilde{W}(x^M) \equiv \sum\nolimits_{i\in M} \tilde{w}_i(x_i).$$

This has therefore become a version of classical utilitarianism, but with a much more general interpretation of individual utility or welfare. Each individual i's function $w_i(x_i)$ represents the ethical value or "goodness" of i's personalized consequence x_i. For Benthamites, goodness corresponded to pleasure minus pain. But much wider and less naïve interpretations of individual utility or welfare functions are certainly allowed. The possibilities are rich enough, in fact, to embrace almost any individualistic ethical theory.

Of course, all the constructions presented here rely on interpersonal comparisons of utility. These have been controversial among economists. Nevertheless, they can be interpreted as ethical preferences for different kinds of people — an idea which is expounded at some length in Hammond (1991b), so I shall not repeat the discussion here. Nor will I repeat here the possible reformulations of Arrow's independence of irrelevant alternatives condition so that it becomes consistent with Harsanyi's form of utilitarianism — on this topic, see Hammond (1991a).

4. Societies with Incomplete Information and Personal Rights

The previous sections have considered only societies in which each individual $i \in M$ was assumed to have a known welfare ordering. Yet real ethical decisions often have to be taken which affect individuals whose interests may be known only very imperfectly. In addition, the ethical decision maker is usually not the only person taking decisions. The individuals in the set M may also be making their own choices. These may affect each other both directly, and also indirectly through their effect on what possibilities remain open to the ethical decision maker.

Such complicated interactions are, of course, the subject of game theory. It is as though some ethical "principal" were confronting a set of individual "agents," each having their own personal objective. Moreover, it is natural to think of the society having to be described by a game of incomplete information, in the sense which Harsanyi (1967–8) was the first to set out formally. After all, both the agents and the ethical principal are players who are likely to be imperfectly informed about one another's objectives, beliefs, etc. In addition, as I have tried to argue in Hammond (1990b), the (ethical) principal will not be able to make Bayesian rational decisons unless they are analysed within some complete game model of the society in which they are all living. It is then necessary as well to form appropriate prior beliefs about what strategies the various agents will choose in that model. This is entirely in accord with Bernheim (1984, 1986) and Pearce's (1984) work on rationalizable strategies, as well as that of Aumann (1987) and others on correlated equilibrium.

A framework is needed to describe the ethical principal's incomplete information, including incomplete information about the incomplete information of other individuals, and about their incomplete information regarding the incomplete information of others, etc. Such a framework is provided by Harsanyi's (1967–8) notion of the "type" of a player in the game. It is actually a far from trivial issue whether a big enough space of possible types to accommodate this infinite regress of incomplete information could ever be constructed. Yet an affirmative answer, at least in principle, has now been provided by, for example, Mertens and Zamir (1985) or Tan and Werlang (1988, pp. 373–5). They use ideas that were, however, pioneered earlier by Böge and his associates — see Armbruster and Böge (1979), Böge and Eisele (1979), and the earlier unpublished work cited therein.

In game theory, a player's type should include everything relevant for determining that player's: (i) payoff function; (ii) beliefs about all the other players' types; (iii) rule for selecting a particular strategy that maximizes expected utility, whenever there is more than one strategy that does so. In ethics, it is also necessary to include: (iv) everything relevant to determining that individual's ethical welfare ordering (which is generally different from the individual's own payoff function).

A society with incomplete information can now be formulated. It will consist of a set of individuals whose "names" or labels lie in the finite set N. For each $j \in N$ there is a set T_j of potential types for person j. Each possible type $t_j \in T_j$ of each individual $j \in N$ will be regarded as a separate "contingent individual," described by the pair (j, t_j). The set of contingent individuals who actually exist is effectively random, since it is unknown to the ethical decision maker. Thus society must be thought of as consisting of *all* possible contingent individuals, since all have potential interests which can be affected by the ethical decision. The membership of the society is therefore given by the set

$$M := \bigcup\nolimits_{j \in N} (\{j\} \times T_j)$$

of all possible pairs (j, t_j) satisfying $j \in N$ and $t_j \in T_j$. In the special case when $T_j = T$ for all $j \in N$, then M can be expressed more simply as the product space $N \times T$. A serious complication, of course, is that if any of the type spaces T_j is infinite then the set M will also be infinite — even though N itself is only finite. This gives rise to analytical complexities such as the need to replace sums by integrals and to discuss measurability issues. In order to avoid these, it will simply be assumed here that, as in Harsanyi's original formulation of games of incomplete information, each player j has a finite type space T_j.

As in Section 2, the ethical decision maker is assumed to be concerned about the social states or consequences in some domain X. Only now it has to be recognized that, even if full control over the social state really were posible, it might well not be ethically desirable. Instead, the social state should probably respond to changes in individuals' types, because part of each individual i's type describes i's own ethical welfare ordering. Changes in this ordering should often give rise to changes in the social state. In economics, for example, where the social state is the entire allocation of resources within an economy, maintaining efficiency requires responding appropriately to changes in each individual's wants and needs. Moreover, such changes may be inevitable because the ethical decision maker has limited ability to prevent individuals from choosing certain aspects of the social state as they please.

To represent this dependence of the social state on individuals' types, an extended notion of *type-contingent social state* becomes necessary. First, let $t^N \in T^N := \prod_{j \in N} T_j$ denote a typical *type profile*, with one type for each named individual $j \in N$. Responsiveness to types obviously requires that the social state x should be a function $t^N \mapsto x = \xi(t^N)$ of the variable type profile, where $\xi : T^N \to X$. The space of all possible type-contingent social states is then

$$\Xi := X^{T^N} := \prod\nolimits_{t^N \in T^N} X(t^N) = \{\, \xi : T^N \to X \,\},$$

where $X(t^N)$ denotes the set of social states that can occur when the type profile is t^N. Such type-contingent social states are effectively the same as the "game forms" used by Sugden (1985, 1986) and others in their

discussion of rights. Individuals' types are equivalent to strategies in such a game form, and the outcome of the game is what I am calling a social state.

All the previous arguments of Harsanyi and of this paper could now be applied to the society with membership M and space of type-contingent social states Ξ. They suggest that an appropriate ethical objective is the maximization of the expected value of some cardinal social welfare function having the additive form $W(\xi) \equiv \sum_{i \in M} W_i(\xi)$ for suitable cardinal individual welfare functions W_i ($i \in M$). Now, however, there is much more structure. Really, the only reason why the ethical decision maker needs to consider all different possible type profiles $t^N \in T^N$ is because of uncertainty about which is the right one. Along with Harsanyi, I continue to impose full Bayesian rationality, and so claim that this uncertainty should be described by some subjective probability distribution $\pi(t^N)$ ($t^N \in T^N$). Then the ethical decision maker should be maximizing the expected value with respect to π of some welfare function $W(x; t^N)$ which would be the appropriate one if the type profile were known to be t^N. Moreover, the previous arguments can be used yet again to claim that, for each $t^N \in T^N$, the social welfare function $W(x; t^N)$ should have the additive form $W(x; t^N) \equiv \sum_{j \in N} w_j(x; t_j)$ for suitable type-dependent cardinal individual welfare functions $w_j(\cdot; t_j)$ ($j \in N$; $t_j \in T_j$). Note how it is being assumed that j's welfare function $w_j(\cdot; t_j)$ depends only on j's own type t_j; this is natural on the understanding that all kinds of external effects should be included within the definition of any social state or consequence $x \in X$.

When the type-contingent social state is $\xi : T^N \to X$, putting these different functions $W(x; t^N)$ together gives

$$W(\xi) \equiv \sum\nolimits_{t^N \in T^N} \pi(t^N) \sum\nolimits_{j \in N} w_j(\xi(t^N); t_j)$$

as the appropriate measure of expected social welfare, after allowing for uncertainty about the type profile $t^N \in T^N$. Generally this uncertainty forces the ethical decision maker to trade off decisions leading to good consequences for different type profiles, and to do so in a way that is sensitive to the subjective probability assessments $\pi(t^N)$ ($t^N \in T^N$). Only by discovering the true type profile can this uncertainty be avoided. There are, however, serious obstacles in the way of doing so; this is the final topic to be discussed in this paper.

Indeed, the question of how to elicit private information has been the topic of much work since the early 1970's. There is no space here to discuss properly what has become an enormous literature. Nor is there any need, since the basic point can be made very briefly. It is that, in order to be able to have the social state adapt to changes in individuals' private information, it is necessary that the right incentives be created. I have already suggested that one can usefully think of a society as composed of individuals who interact within some enormously complicated game of incomplete information — see also Hammond (1990a). The issue becomes one of how the ethical decision maker should behave in such a game. The ethical objectives have already been discussed. They are the topic of social choice theory, and all the previous part of this paper has been devoted to giving reasons why the most suitable objective is likely to be some version of Harsanyi's form of utilitarianism. The need to provide incentives, however, arises because the ethical decision maker does not have full control over the process which determines the social state — other individuals' actions are also important, and will usually be much more so. Thus incentives are concerned with constraints on what the ethical decision maker can achieve, rather than with desirable social objectives. For this reason,

they really lie beyond the scope of this paper.

Nevertheless, it is already possible to see that the issue of individual rights arises much more naturally when society is modelled as having incomplete information. Dasgupta (1982) was probably the first to point this out with any kind of formal argument. An increased respect for rights also emerges if we take into account the limited control that the ethical decision maker has over individual actions which affect the social state. Both incomplete information and moral hazard give individuals powers to affect their own destinies in ways which would meet with the approval of libertarians. An ethical decision maker can only control individuals' actions if some enforcement mechanism is in place. Most of the time, such enforcement mechanisms are inevitably both intrusive and costly in other ways. Then, however, a proper ethical utilitarian calculation of the costs and benefits of having an enforcement mechanism is quite likely to decide that effective enforcement is not worthwhile. The same is true of incomplete information about something which, if known, could improve the quality of some important ethical decision. In order that something which is known only privately can be discovered and the information used, individuals have to be provided with incentives to reveal what they alone know. These can be normal incentives, such as those a shop-keeper provides to encourage customers to reveal what they want to buy and what is an upper bound on the price which they are willing to pay. Alternatively, the ethical decision maker may be able to coerce the information out of the individual in some unpleasant way. In the latter case, however, the coercion itself imposes enormous costs which are very rarely going to be outweighed by any ethical benefits which the knowledge might yield. The point is that a full description of the type-contingent social state has to specify what means will be used to enforce certain kinds of behaviour and to encourage certain kinds of private information to be revealed. When this is done properly, many attempts to infringe what people see as their rights, even if such attempts would otherwise be ethically valid, are likely to be evaluated as ethically unacceptable when all kinds of enforcement cost are taken into account.

REFERENCES

W. ARMBRUSTER AND W. BÖGE (1979), "Bayesian Game Theory," in *Game Theory and Related Topics* edited by O. Moeschlin and D. Pallaschke (Amsterdam: North-Holland), pp. 17–28.

R.J. AUMANN (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, **55**: 1–18.

B.D. BERNHEIM (1984), "Rationalizable Strategic Behavior," *Econometrica*, **52**: 1007–1028.

B.D. BERNHEIM (1986), "Axiomatic Characterizations of Rational Choice in Strategic Environments," *Scandinavian Journal of Economics*, **88**: 473–488.

D. BERNOUILLI (1738; translated by L. Sommer, 1954) "Specimen theoriae novae de mensura sortis" [title of translation: 'Exposition of a New Theory on the Measurement of Risk'], *Econometrica*, **22**: 23–36.

W. BÖGE AND T. EISELE (1979), "On Solutions of Bayesian Games," *International Journal of Game Theory*, **8**: 193–215.

K.C. BORDER (1981), "Notes on von Neumann-Morgenstern Social Welfare Functions," unpublished preprint, California Institute of Technology.

K.C. BORDER (1985), "More on Harsanyi's Utilititarian Cardinal Welfare Theorem," *Social Choice and Welfare*, **1**: 279–281.

J. BROOME (1987), "Utilitarianism and Expected Utility," *Journal of Philosophy*, **84**: 405–422.

J. BROOME (1989) "Should Social Preferences Be Consistent?" *Economics and Philosophy*, **5**: 7–17.

J. BROOME (1990), "Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism," *Review of Economic Studies*, **57**: 477–502.

T. COULHON AND P. MONGIN (1989), "Social Choice Theory in the Case of Von Neumann-Morgenstern Utilities," *Social Choice and Welfare*, **6**: 175–187.

G. CRAMER (1728), Letter to Nicolas Bernouilli; extracts printed in Bernouilli (1738) and in Sommer's (1954) translation.

P.S. DASGUPTA (1982), "Utility, Information, and Rights," in *Utilitarianism and Beyond* edited by A.K. Sen and B. Williams (Cambridge: Cambridge University Press), pp. 199–218.

P.A. DIAMOND (1967), "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility: Comment," *Journal of Political Economy*, **75**: 765–766.

Z. DOMOTOR (1979), "Ordered Sum and Tensor Product of Linear Utility Structures," *Theory and Decision*, **11**: 375–399.

M. FLEMING (1952), "A Cardinal Concept of Welfare," *Quarterly Journal of Economics*, **66**: 366–384.

M. FLEMING (1957), "Cardinal Welfare and Individualistic Ethics: A Comment," *Journal of Political Economy*, **65**: 355–357.

M. FRIEDMAN AND L.J. SAVAGE (1952), "The Expected Utility Hypothesis and Measurement of Utility," *Journal of Political Economy*, **60**: 463–474.

W. GAERTNER AND P.K. PATTANAIK (EDS.) (1988), *Distributive Justice and Inequality*. Berlin: Springer Verlag.

W. GAERTNER, P.K. PATTANAIK, AND K. SUZUMURA (1988), "Individual Rights Revisited," Fachbereich Wirtschaftswissenschaft, Universität Osnabrück; to appear in *Economica.*

D. GALE (1960), *Theory of Linear Economic Models*. New York: McGraw-Hill.

P.J. HAMMOND (1983), "Ex-Post Optimality as a Dynamically Consistent Objective for Collective Choice under Uncertainty," in *Social Choice and Welfare* edited by P.K. Pattanaik and M. Salles (Amsterdam: North Holland), ch. 10, pp. 175–205.

P.J. HAMMOND (1986), "Consequentialist Social Norms for Public Decisions," in *Social Choice and Public Decision Making: Essays in Honor of Kenneth J. Arrow, Vol. 1* edited by W. P. Heller, R. M. Starr and D. A. Starrett (Cambridge: Cambridge University Press), ch. 1, pp. 3–27.

P.J. HAMMOND (1987), "On Reconciling Arrow's Theory of Social Choice with Harsanyi's Fundamental Utilitarianism," in *Arrow and the Foundations of the Theory of Economic Policy* edited by G.R. Feiwel (London and New York: Macmillan and New York University Press), ch. 4, pp. 179–222.

P.J. HAMMOND (1988a), "Consequentialism and the Independence Axiom," in *Risk, Decision and Rationality* edited by B.R. Munier (Dordrecht: D. Reidel), pp. 503–516.

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

P.J. HAMMOND (1988c), "Consequentialist Demographic Norms and Parenting Rights," *Social Choice and Welfare*, **5**: 127–145; reprinted in Gaertner and Pattanaik.

P.J. HAMMOND (1990a), "Incentives and Allocation Mechanisms," in *Advanced Lectures in Quantitative Economics* edited by R. van der Ploeg (New York: Academic Press), ch. 6, pp. 213–248.

P.J. HAMMOND (1990b), "A Revelation Principle for (Boundedly) Bayesian Rationalizable Strategies," European University Institute, Working Paper ECO 90/4; to appear in R.P. Gilles and P.H.M. Ruys (eds.), *Economic Behaviour in an Imperfect Environment* (Amsterdam: North-Holland).

P.J. HAMMOND (1991a), "Independence of Irrelevant Interpersonal Comparisons," European University Institute, Working Paper ECO 90/5; *Social Choice and Welfare* (in press).

P.J. HAMMOND (1991b), "Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made," European University Institute, Working Paper ECO 90/3; to appear in J. Elster and J. Roemer (eds.), *Interpersonal Comparisons of Well-Being* (New York: Cambridge University Press).

J.C. HARSANYI (1953), "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking," *Journal of Political Economy*, **61**: 434–5; reprinted in Harsanyi (1976).

J.C. HARSANYI (1955), "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, **63**: 309–321; reprinted in Phelps (1973) and in Harsanyi (1976).

J.C. HARSANYI (1967–8), "Games with Incomplete Information Played by 'Bayesian' Players, I–III," *Management Science*, **14**: 159–182, 320–334, 486–502.

J.C. HARSANYI (1975a), "Nonlinear Social Welfare Functions: Do Welfare Economists Have a Special Exemption from Bayesian Rationality?," *Theory and Decision*, **6**: 311–32; reprinted in Harsanyi (1976).

J.C. HARSANYI (1975b), "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *American Political Science Review*, **69**: 594–606; reprinted in Harsanyi (1976).

J.C. HARSANYI (1976), *Essays on Ethics, Social Behavior, and Scientific Explanation*. Dordrecht: D. Reidel.

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

A.P. LERNER (1944), *The Economics of Control*. London: Macmillan.

J.-F. MERTENS AND S. ZAMIR (1985), "Formalization of Bayesian Analysis of Games with Incomplete Information," *International Journal of Game Theory*, **14**: 1–29.

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

J. VON NEUMANN AND O. MORGENSTERN (1943; 3rd edn., 1953), *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

P.K. PATTANAIK (1968), "Risk, Impersonality, and the Social Welfare Function," *Journal of Political Economy*, **76**: 1152–1169; reprinted in Phelps (1973).

D. PEARCE (1984), "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, **52**: 1029–1050.

E.S. PHELPS (ED.) (1973), *Economic Justice*. Harmondsworth: Penguin Books.

J. RAWLS (1959), "Justice as Fairness," *Philosophical Review*, **67**: 164–94.

J. RAWLS (1971), *A Theory of Justice*. Cambridge, Mass: Harvard University Press.

J. RAWLS (1974), "Some Reasons for the Maximin Criterion," *American Economic Review (Papers and Proceedings)*, **64**: 141–6.

J. RILEY (1989, 1990), "Rights to Liberty in Purely Private Matters: Parts I and II," *Economics and Philosophy*, **5**: 121–166 and **6**: 27–64.

S. SELINGER (1986), "Harsanyi's Aggregation Theorem without Selfish Prefences," *Theory and Decision*, **20**: 53–62.

A.K. SEN (1969), "Planner's Preferences: Optimality, Distribution and Social Welfare," in *Public Economics* edited by J. Margolis and H. Guitton (London: Macmillan), ch. 8, pp. 201–221.

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

A.K. SEN (1973), "On Ignorance and Equal Distribution," *American Economic Review*, **63**: 1022–4; reprinted in Sen (1982).

A.K. SEN (1982), *Choice, Welfare and Measurement*. Oxford: Basil Blackwell, and Cambridge, Mass.: MIT Press).

R. SUGDEN (1985), "Liberty, Preference and Choice," *Economics and Philosophy*, **1**: 213–229.

R. SUGDEN (1986), *The Economics of Rights, Cooperation and Welfare*. New York: Basil Blackwell.

T.C.-C. TAN AND S.R. DA C. WERLANG (1988), "The Bayesian Foundations of Solution Concepts of Games," *Journal of Economic Theory*, **45**: 370–391.

W.S. VICKREY (1945), "Measuring Marginal Utility by Reactions to Risk," *Econometrica*, **13**: 319–33.

J.A. WEYMARK (1990), "Harsanyi's Social Aggregation Theorem with Alternative Pareto Principles," University of British Columbia, Department of Economics, Discussion Paper No.: 90-28; presented to the Fifth Karlsruhe Seminar on *Models and Measurement of Welfare and Inequality*.