Extending the Original Position:
Revisiting the Pattanaik Critique of Vickrey/Harsanyi Utilitarianism

Peter J. Hammond

No 1008

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF
WARWICK

Extending the Original Position: Revisiting the Pattanaik Critique of Vickrey/Harsanyi Utilitarianism

Peter J. Hammond*

Abstract

Harsanyi's original position treats personal identity, upon which each individual's utility depends, as risky. Pattanaik's critique is related to the problem of scaling "state-dependent" von Neumann–Morgenstern utility when determining subjective probabilities. But a unique social welfare functional, incorporating both level and unit interpersonal comparisons, emerges from contemplating an "extended" original position allowing the probability of becoming each person to be chosen. Moreover, the paper suggests the relevance of a "Harsanyi ethical type space", with types as both causes and objects of preference.

*Department of Economics, University of Warwick, Coventry CV4 7AL, UK. e-mail: p.j.hammond@warwick.ac.uk

"By nature, men are nearly alike; by practice, they get to be wide apart."
Confucius, *Analects* ch. 17 from `www.analects.org`.

1 Introduction

1.1 Vickrey, Rawls, and Harsanyi

Pattanaik (1968, pp. 1157–8) reminded economists and social choice theorists how the philosopher Hare (1961) defined ethics as being about universal prescriptive statements. An example is the Golden Rule of reciprocal ethics, which is usually stated as: "Do unto others as you would have them to do unto you." An improved version, however, would seem to be: "Do unto others as you should want them to do unto you." After all, one should recognize that a person's own wishes, even for themselves, may not always be for the best.

Be that as it may, an intuitively appealing way to introduce universalizability when discussing policy choices is through impersonality. That is, we seek normative statements about policy that remain valid no matter which is the individual in society whose position we think of occupying. This approach can be linked to contractarianism through what eventually became Rawls' (1951, 1958, 1971) device of an "original position", where each individual is put "behind the veil of ignorance". As Pattanaik points out, this adds to universalizability by requiring normative statements about policy to remain valid even when a person's identity remains unknown.[2]

Even before Rawls really elaborated his theory, Vickrey (1945, 1960) and Harsanyi (1953, 1955) had already independently formulated a different approach to the original position. In it, each individual faces an equal probability of becoming any person in the society. Then the standard hypotheses of Bayesian rationality imply that, in this original position, each individual should want to maximize the expected value of a von Neumann–Morgenstern utility function defined over lotteries whose outcomes combine social states with personal identities.

[2]See Mongin (2001) for a careful discussion of the ethics of impartiality.

1.2 Pattanaik, Kolm and Broome

This Vickrey/Harsanyi version of the original position is the subject of Pattanaik's (1968) important critique. Amongst other things, he carefully distinguished two different variations, namely an "objective" approach due to Vickrey, as opposed to a "subjective" approach due to Harsanyi. Both present their own difficulties, however. In particular, this original position fails to determine the social preference ordering uniquely. One difficulty is in knowing which von Neumann–Morgenstern utility function, within the relevant cardinal equivalence class, should be used to represent that person's welfare in the original position. More fundamentally, Pattanaik (1968) raised the issue of whether that utility function should represent that person's own attitudes to risk, as well as their preferences between different social states.

Here the fundamental difficulty is that the expected value of each relevant von Neumann–Morgenstern utility function represents only lotteries over social states, with personal identity fixed. In fact, for the original position argument to work, different von Neumann–Morgenstern utility functions for different personal identities need to be scaled appropriately so that their expected values are indeed defined over lotteries whose outcomes include variable personal identities as well as social states. This difficulty also lies at the heart of the debate between Broome (1993, 1994) and Kolm (1994a, b) over the difference between a cause of preference and an object of preference.

1.3 Extending the Original Position

This paper sets out to resolve many of these issues by considering a suitably extended original position. The extension consists in requiring that preferences apply not just in the Vickrey/Harsanyi original position where there is an equal chance of becoming any one person, nor just in arbitrarily biased original positions with unequal probabilities of becoming different individuals in society, but also in decision problems where these biased probabilities themselves become objects of choice. There is an obvious relationship here to the extension of the Anscombe and Aumann (1963) theory of subjective probability that allows state-independent utilities to be derived even for state-dependent consequence domains, as set out in Hammond (1998, 1999).

1.4 Ethical Types

As already suggested, another issue to be discussed here concerns the debate between Broome and Kolm on the validity of the alleged distinction between causes of preference on the one hand, and objects of preference on the other. Here it is argued that the construction of a suitable type space could allow those types to be simultaneously both causes and objects of preference.

1.5 Outline

Section 2 below introduces the notation and basic framework to be used throughout the subsequent sections of the paper, sometimes with suitable embellishment. A particular feature worth mentioning here is the inclusion of a special individual consequence representing non-existence. This permits the original position to include a variable number of individuals, thus allowing for the case when population is also affected by the social decision, as in Hammond (1988) and the works cited there, especially Sidgwick (1887), Meade (1955), Dasgupta (1969) and Parfit (1984). Specifically, one can convert "different numbers" problems in social choice to "same numbers" problems by adding as many non-existent individuals as necessary.

Thereafter the paper moves on in Section 3 to discuss Vickrey's original position where there is an even chance lottery of facing each individual's *objective* circumstances. This objective version of the original position obviously omits some relevant subjective circumstances, notably how the person would wish to make choices in the inherently risky original position.

In an attempt to address this failure, Harsanyi introduced "ethical" subjective preferences. These are the main subject of Pattanaik's critique, and of the latter part of this paper. As Section 4 reminds us, for the original position argument to work, different von Neumann–Morgenstern utility functions for different personal identities need to be scaled appropriately so that their expected values are indeed defined over lotteries whose outcomes include variable personal identities as well as social states.

Next, Section 5 explains in more detail how this scaling issue is really a special case of the difficulty that has to be confronted when contemplating decisions under uncertainty in case the consequence domain is state-dependent. After all, the domain of possible consequences that could result from being a particular person named Chris, say, are generally quite different from the possible consequences of being Pat, Robin, or Sam. Not least, each of them

could be either a man or a woman. The following Section 6 briefly discusses how the scaling issue is straightforward to resolve provided that one considers an extended original position where even the (biased) probabilities of becoming different individuals can be chosen.

Section 7 goes on to introduce the notion of an ethical type space, and briefly explores the relevance of some analogies with the kind of type space that Harsanyi (1967) used to develop the modern theory of games of incomplete information.

The concluding Section 8 contains a brief summary and suggestions for further work.

2 Notation and Basic Assumptions

2.1 Individuals and Their Personal Consequences

As in Hammond (1988) and Blackorby *et al.* (2005), for instance, we allow for a variable population that can be influenced by policy, so is effectively an object of choice. Accordingly, we consider a finite set M large enough to include all potential individuals. Let $m := \#M$ denote the size of this set. The actual population is assumed to be a subset $N \subseteq M$ of size $n \leq m$.

Let Y denote an arbitrary domain of possible **personal consequences**. The members y of the set Y are assumed to represent everything that could be relevant to a single individual when making an ethical policy decision that affects them.

One important personal consequence, naturally, is whether that person ever comes into existence. So we include a particular consequence $y_0 \in Y$ that represents non-existence. We assume that any other personal consequence $y \in Y \setminus \{y_0\}$ entails that particular individual existing as an ethically relevant person over a time interval whose specification can be included in the description of y, if those dates are also ethically relevant.

2.2 Social States as Personal Consequence Profiles

For each individual $i \in M$, let $Y_i \subseteq Y$ denote the subdomain of personal consequences that could be relevant to person i. We assume that $y_0 \in Y_i$ for each individual $i \in M$, meaning that social preferences are always defined even over what may be purely hypothetical social states in which i never comes into existence.

Next, we identify **social states** with **personal consequence profiles** $y^M = \langle y_i \rangle_{i \in M}$ which belong to the Cartesian product $Y^M := \prod_{i \in N} Y_i$ of the personal consequence subdomains. As in the theory of public goods, this product domain allows us to consider in principle varying just one individual's consequence at a time, even though feasibility may impose the constraint that some "public" components of individuals' consequences must all change together.

2.3 Simple Consequence Lotteries

Let $\Delta(Y)$ denote the space of **simple consequence lotteries** η, each of which attaches a positive probability $\eta(y)$ to each member y of a **finite support** of elements in Y. Obviously $\sum_{y \in Y} \eta(y) = 1$, where the sum is well-defined because only a finite set of terms are non-zero.

2.4 Expected Utility of Single-Person Situations

For each $i \in M$, consider **single-person situations** where any social decision that is taken affects only individual i, and results in a simple consequence lottery or random outcome $\eta \in \Delta(Y_i)$. Under standard axioms for normative decision-making with risky consequences, the appropriate ethical objective in such single-person situations is to maximize the expectation

$$\mathbb{E}_\eta[u_i(y)] \equiv \sum_{y \in Y} \eta(y) u_i(y) \tag{1}$$

of each von Neumann–Morgenstern utility function (NMUF) $u_i : Y_i \to \mathbb{R}$ in a **cardinal equivalence class**. This class is defined so that two functions u_i and $\tilde{u}_i$ are **cardinally equivalent** if and only if they coincide on Y_i up to a **positive affine transformation** of the form

$$\tilde{u}_i \equiv \alpha_i + \rho_i u_i, \quad \text{where } \rho_i > 0. \tag{2}$$

In particular, for each quadruple of personal consequences $a, b, c, d \in Y$, the ratio

$$\frac{u_i(a) - u_i(b)}{u_i(c) - u_i(d)} \tag{3}$$

of utility differences must equal the (constant) marginal rate of substitution $-\delta q/\delta p$ between small probability shifts δp from consequence b to a, and δq from consequence c to d, as one moves along the tangent to an indifference surface in the three-dimensional subset $\Delta(\{a, b, c, d)\})$ of lotteries in $\Delta(Y)$ whose support is the set $\{a, b, c, d)\}$. These ratios of utility differences, of course, are preserved by any positive affine transformation of the function $y \mapsto u_i(y)$.

2.5 Zero Utility of Non-Existence

A convenient normalization will be to put $u_i(y_0) = 0$, thus attaching a utility of zero to non-existence.[3] Then u_i and $\tilde{u}_i$ are **equivalent** if and only if they coincide on Y_i up to a **positive linear transformation** of the form $\tilde{u}_i \equiv \rho_i u_i$, where $\rho_i > 0$. In particular, following the above interpretation of (3), for each pair of personal consequences $y, y' \in Y$, the ratio $u_i(y)/u_i(y')$ of utilities equals the (constant) marginal rate of substitution $-\delta q/\delta p$ between small probability shifts δp from consequence y_0 to y, and δq from consequence y_0 to y'. These utility ratios are also preserved by any positive linear transformation of the utility function u_i taking the form

$$\tilde{u}_i \equiv \rho_i u_i, \quad \text{where } \rho_i > 0. \tag{4}$$

[3]By contrast, Blackorby, Bossert and Donaldson (2005) meet Parfit's (1984) "repugnant conclusion" by introducing a *critical level* of utility c_i for each individual $i \in M$, defined so that $u_i(y_0) = c_i$. This allows utility to be normalized in a different way so that $u_i(y) > 0$ if and only if the existing individual $i \in N$ regards consequence y as better than having never come into existence at all. One implication of their critical level approach is that, compared to the case when we ascribe a zero level of utility to non-existence, it becomes harder to interpret the critical and zero levels of utility "objectively" in the same way for different individuals. By contrast, in Hammond (1988) I argue that the "repugnant conclusion" is not so repugnant after all if one reckons the relevant utilities of parents, including their preferences for children, and perhaps even more relevantly, for reproductive freedom.

3 Vickrey's Objective Original Position

3.1 The Original Position Lottery

In Vickrey's version of the original position, each personal consequence in the space Y is identified with an observable "objective" circumstance such as income, educational qualifications, date of birth, marital status, etc. We are also extending Vickrey's version by including non-existence as a possible objective circumstance, represented by the specific consequence $y_0 \in \cap_{i \in M} Y_i \subseteq Y$. Accordingly, given any social state $y^M \in Y^M$, all individuals behind the veil of ignorance can be regarded as facing the same equal chance lottery of becoming any of the individuals $j \in M$, associated with the corresponding personal consequence $y_j \in Y$. That is, all individuals — whether actual or potential — are put in the position of facing the same lottery $\eta_{y^M} \in \Delta(Y)$ over personal consequences defined by

$$\eta_{y^M}(E) := \frac{1}{m}\#\{j \in M \mid y_j \in E\} \quad \text{for each finite } E \subseteq Y. \tag{5}$$

Suppose now we pay special attention to the set M of all individuals who ever come into existence. Then $\eta_{y^M}(Y \setminus \{y_0\}) = n/m$ and, for each finite $E \subseteq Y \setminus \{y_0\}$, one has

$$\eta_{y^M}(E) := \frac{1}{m}\#\{j \in N \mid y_j \in E\} = \frac{1}{n}\#\{j \in N \mid y_j \in E\}\ \eta_{y^M}(Y \setminus \{y_0\}). \tag{6}$$

3.2 Individuals' Expected Utilities

Though each social state $y^M \in Y^M$ is supposedly objective, in principle different individuals may have different preferences over the lotteries η_{y^M}. Indeed, Vickrey (1960) reinterprets the original position as one confronted by a potential immigrant to a society, who has an equal chance of acquring the personal consequence y_j of each individual $j \in M$. Thus, each existing individual $i \in N$ is still free to have preferences represented by the expected value of his or her own NMUF u_i. The only new requirement is that u_i should be defined over the whole domain Y of possible personal consequences, instead of just over the restricted domain Y_i of possible consequences that are specific to person i.

3.3 Vickrey's Social Welfare Function

Taking the expectation of the different individuals' utility functions w.r.t. η_{y^N} leads to the Vickrey social welfare function $w_i^m : Y^M \to \mathbb{R}$ for a society of m individuals. This can be written as the average utilitarian objective

$$w_i^m(y^M) := \frac{1}{m} \sum_{j \in M} u_i(y_j). \tag{7}$$

Because the population of potential individuals is fixed at size m, the function w_i^m is cardinally equivalent to the total utilitarian objective

$$w_i(y^M) = \sum_{j \in M} u_i(y_j) = \sum_{j \in N} u_i(y_j), \tag{8}$$

where the last equality holds because of the normalization $u_i(y_0) = 0$, which holds for all $i \in M \setminus N$.

3.4 Risky Social States

Finally, there is an straightforward extension to the entire space $\Delta(Y^M)$ of all possible risky social states, each of which is described by an objective joint probability distribution λ^M of the personal consequences of all the m different potential individuals in the set M, including the state y_0 of non-existence. Indeed, in this extension, each individual $j \in M$ faces his or her own personal lottery $\lambda_j \in \Delta(Y_j)$ that must be equal to the marginal distribution on Y_j induced by λ^M on Y^M. Working back to the original position, each individual $i \in N$ then faces the identical compound lottery $\eta_{\lambda^M} \in \Delta(Y)$ over personal consequences defined by

$$\eta_{\lambda^M}(E) := \frac{1}{m} \sum_{j \in M} \lambda_j(E) \quad \text{for each finite } E \subseteq Y. \tag{9}$$

That is, for each finite $E \subseteq Y$, one has

$$\lambda_j(E) = \lambda^M \left(E \times \prod\nolimits_{i \in M \setminus \{j\}} Y_i \right). \tag{10}$$

So, extending to the entire space $\Delta(Y^M)$ of all possible risky social states the social welfare function given by (7) or, equivalently, by (8), we obtain a

(complete and transitive) preference ordering $\succsim$ on $\Delta(Y^M)$ represented by the expectation in the original position compound lottery of (7) or, equivalently, of (8). Specifically, i's version of the social welfare function becomes

$$W_i(\lambda^M) := \frac{1}{m} \sum_{j \in M} \mathbb{E}_{\lambda_j} u_i \tag{11}$$

4 Harsanyi's Subjective Original Position

4.1 Harsanyi's Ethical Types

Vickrey (1945, 1960) regarded the consequence domain Y as if it consisted only of objective circumstances. Yet, as Harsanyi (1955) certainly recognized, subjective features of individuals such as their tastes or preferences are also relevant in ranking these objective circumstances. After all, consider the following radical adaptation of an example discussed by Sen (1970). There are two individuals of whom one (labelled C) is a devoted cricketer enthusiast, whereas the other (labelled B) is a fanatic follower of baseball. Decisions in the original position, such as whether these two people should accompany each other to a cricket match or to a baseball game, then require going behind the veil of ignorance where one would be forced to choose between: (i) becoming C and being dragged reluctantly to watch a baseball game; (ii) or becoming B and being forced to face up to the unfathomable mysteries of cricket. Furthermore, as Pattanaik (1968) points out, if ethical preferences are to determine decisions in the original position, they must also allow for individuals' own subjective attitudes to the risk they face in the equal chance lotteries of the original position.

Despite these evident difficulties, Harsanyi (1955) postulated "ethical" preferences. One way of modelling these is through a space H of **Harsanyi ethical types** h which parametrize the family $y \mapsto u(y; h)$ of NMUFs whose expected values represent ethical preferences over $\Delta(Y)$. In the original position, for each $j \in M$ the utility $u_j(y_j)$ of j's objective personal consequence becomes replaced by the utility $u(y_j; h_j)$ of j's subjective personal consequence $(y_j, h_j) \in Y \times H$.

4.2 Spurious Observer Independence

One key difference from Vickrey's original position is Harsanyi's claim that the new objective function can be expressed as

$$w(y^M; h^M) = \sum_{j \in M} u(y_j; h_j) = \sum_{j \in N} u(y_j; h_j), \tag{12}$$

independent of i. That is, the objective is the same no matter which individual i is taken as the observer. However, this is not really the case, and in fact we have so far failed to determine uniquely the original position preferences for even one observer.

This is because, as Pattanaik (1968) makes clear, the parametric family of utility functions $y \mapsto u(y; h)$ is not determined uniquely. Indeed, after taking into account the normalized utility $u(y_0; h) = 0$ of non-existence, each ethical type h determines only an equivalence class of such utility functions, defined up to separate positive linear transformations of the form

$$\tilde{u}(y; h) \equiv \rho(h) u(y; h) \quad \text{for all } y \in Y, \tag{13}$$

where $\rho(h)$ is a positive constant for each $h \in H$. So far, therefore, our theory is unable to distinguish between the function $w(y^M; h^M)$ given by (12) and any alternative function $\tilde{w}(y^M; h^M)$ given by

$$\tilde{w}(y^M; h^M) = \sum_{j \in M} \rho(h_j) u(y_j; h_j) = \sum_{j \in N} \rho(h_j) u(y_j; h_j), \tag{14}$$

where $\rho(h_j)$ $(j \in M)$ is an arbitrary family of positive constants. In particular, given any finite subset $X \subseteq Y^M$ and any fixed profile $\bar{h}^M \in H^M$ of Harsanyi ethical types, one make any individual $d \in M$ a dictator over the set X of all degenerate lotteries in $\Delta(X)$ by setting all the ratios $\rho(h_j)/\rho(h_d)$ $(j \in M \setminus \{d\})$ sufficiently close to zero.

5 Consequence Domains

5.1 Personal Consequences

The key difficulty with Vickrey's version of the original position described in Section 3 is that the objective consequences $y \in Y$ do not describe completely what is relevant to ethical decisions. As explained in Section 4, the same objective consequence $y \in Y$ has a different ethical significance depending on the Harsanyi ethical type $h \in H$ of the individual who is going to experience y.

What we would really like is for each potential individual $j \in M$ to face the same extended common domain $Z := Y \times H$ whose members $z = (y, h)$ are pairs of *personal consequences* that combine objective circumstances with personal ethical types. In effect, therefore, we simply replace Y in Section 3 by the new space Z, and the assessment by each existing individual $i \in N$ of the individual welfare function $u_i : Y \to \mathbb{R}$ to use in the original position by the new function $u_i^* : Z \to \mathbb{R}$. After retaining the obvious normalization $u_i^*(y_0, h) = 0$ for all $h \in H$, we arrive at the modified form

$$w_i^*(z^M) = \sum_{j \in M} u_i^*(z_j) = \sum_{j \in N} u_i^*(z_j), \tag{15}$$

of equation (8).

Indeed, the lotteries we have considered so far are limited to the particular subdomain of $\Delta(Y \times H)$ whose members ζ_{z^M}, in an obvious adaptation of equation (5) from Y to $Z = Y \times H$, take the form

$$\zeta_{z^M}(F) := \frac{1}{m} \#\{j \in M \mid z_j \in F\} \quad \text{for each finite } F \subseteq Z. \tag{16}$$

In particular, the probability of any finite $F \subseteq Z$ must be some integer multiple of $1/m$.

Nevertheless, there is an obvious extension to the space $\Delta(Z^M)$ of risky social states, each described by an objective joint distribution λ^M of the personal consequences of the m different potential individuals in the set M. Then each $j \in M$ faces a personal lottery $\lambda_j \in \Delta(Z_j)$, equal to the marginal distribution on Z_j induced by λ^M on Z^M. That is, following (10), for each finite $F \subseteq Z$ one has

$$\lambda_j(F) = \lambda^M \left(F \times \prod\nolimits_{i \in M \setminus \{j\}} Z_i \right). \tag{17}$$

Working back to the original position, all individuals $i \in N$ face the identical compound lottery $\zeta_{\lambda^M} \in \Delta(Z)$ over personal consequences defined by

$$\eta_{\lambda^M}(F) := \frac{1}{m} \sum_{j \in M} \lambda_j(F) \quad \text{for each finite } F \subseteq Z. \tag{18}$$

Each such lottery, however, is restricted to the domain

$$\frac{1}{m} \sum_{j \in M} \Delta(Z_j) := \left\{ \zeta \in \Delta(Z) \mid \exists \lambda_j \in \Delta(Z) \ (j \in M) : \zeta = \frac{1}{m} \sum_{j \in M} \lambda_j \right\}. \tag{19}$$

5.2 State-Dependent Consequence Domains

It makes little sense, however, to assume that the extended consequence domain Z is really the same for each individual. After all, in our earlier example, can we really imagine a common consequence domain which includes not only the cricket enthusiast's experience when watching baseball, but also the baseball fan's experience at a cricket match? Thus, we really need to allow different consequence domains Z_h for each ethical type $h \in H$. Indeed, one might argue that $Z_h \subseteq Y \times \{h\}$ for each $h \in H$, in which case the two domains $Z_h, Z_{h'}$ are not only different when $h \neq h'$, but pairwise disjoint.

Now, in the theories of subjective probability developed by Savage (1954) and by Anscombe and Aumann (1963), a key assumption is that, given any fixed consequence $\bar{z} \in Z$, there exists a "constant act" that could produce $\bar{z}$ as the consequence in all possible uncertain states of the world. If we liken ethical types in the original position to uncertain states of nature, this would imply that the consequence domains satisfy $Z_h = Z$ for all $h \in H$, and so are state-independent. But in the setting of this paper, state-dependent consequence domains Z_h must clearly be allowed.

Because the consequence domains Y_h may be entirely disjoint for different $h \in H$, the problem we face here is formally identical to the subjective expected utility model with state-dependent consequence domains that was set out in Hammond (1999), as well as Section 6 of Hammond (1998). Actually, following the key pioneering ideas of Drèze in the 1960s, most of the previous literature on decision theory with state-dependent consequence domains has considered state-dependent utility — see, for example, the work that is collected, discussed, and surveyed in Drèze (1987), Karni (1985, 2008), Drèze and Rustichini (2004), etc. These works have typically sought to extend Savage's theory of subjective probability; in the original position we are

considering, however, with an equal probability of becoming any potential or existing individual, there are clearly objective probabilities only.

Still, one could try applying the kind of extended preference ordering considered in Karni, Schmeidler and Vind (1983) to the present context. This makes it seem natural to consider "biased" original positions with unequal probabilities of acquiring different Harsanyi ethical types. Or, following Hammond (1999), even an **extended original position** where these biased probabilities can be chosen. This is our next topic.

6 An Extended Original Position

6.1 Biased Original Positions

So far, in each original position we have considered, there has been an equal probability $1/m$ of becoming any potential individual $j \in M$. Now we consider and even compare *biased original positions* with generally different probabilities μ_j of becoming each $j \in M$, where $\mu \in \Delta(M)$. Or more exactly, we consider the domain whose members are profiles $(y^M, h^M) \in Y^M \times H^M$ of consequence–type pairs $(y, h) \in Y \times H$, together with lotteries $\mu \in \Delta(M)$. Then each triple $(\mu, y^M, h^M) \in \Delta(M) \times Y^M \times H^M$ together determines the (simple) probability measure on consequence–type pairs $(y, h) \in Y \times H$ defined by

$$\pi_{(\mu, y^M, h^M)} := \mu\left(\{j \in M \mid (y_j, h_j) \in E\}\right) \quad \text{for each finite } E \subseteq Y \times H. \tag{20}$$

6.2 Comparing Biased Original Positions

In the extended original position we consider, we postulate that each individual $i \in N$ has preferences over biased original positions represented by triples $(\mu, y^M, h^M) \in \Delta(M) \times Y^M \times H^M$. Specifically, we assume that these preferences correspond to preferences over the induced lotteries $\pi_{(\mu, y^M, h^M)} \in \Delta(Y \times H)$. Moreover, we postulate that these preferences are represented by the expected value $\mathbb{E}[v_i(y, h)]$ of an **extended** NMUF $v_i : Y \times H \to \mathbb{R}$ defined over consequence–type pairs.

To be explicit, the expected utility to person i of the biased original position lottery determined by the triple (μ, y^M, h^M) is

$$W_i(\mu, y^M, h^M) := \mathbb{E}_{\pi_{(\mu, y^M, h^M)}} v_i(y, h) = \sum_{j \in M} \mu(\{j\}) v_i(y_j, h_j). \tag{21}$$

This implies that v_i is the von Neumann–Morgenstern utility function defined on $Y \times H$ whose expected value determines individual i's preferences, not only in the Vickrey/Harsanyi unbiased original position, but also in every biased original position, and also in comparing different biased original position.

6.3 Personal Interpersonal Utility Functions

Using information about comparisons between such biased original positions allows the simultaneous derivation of an interpersonal utility function (unique up to a cardinal ratio scale) defined on person-dependent consequence domains. Like the "fundamental" interpersonal notion of utility considered by Tinbergen (1957) and Kolm (1972, 1994), and later endorsed by Rawls (1982), this function incorporates both level interpersonal comparisons, representing who one would prefer to be with certainty, as well as unit interpersonal comparisons, representing marginal rates of substitution between the probabilities of becoming different kinds of person. Also, unlike the general "state-dependent" utilities that have appeared in the literature cited above, the ideas in Hammond (1999) imply that this fundamental utility must be equal for indistinguishable individuals.

Nevertheless, one must emphasise a key difference from Tinbergen and Kolm, as well as Rawls. Nowhere do we claim that our notion of "fundamental" interpersonal utility must be equal for distinguishable individuals when facing an extended original position. This point may be the key to resolving the difference between Broome and Kolm regarding whether there is a valid distinction between the causes and objects of preference.

7 Causes versus Objects of Preference

7.1 Broome versus Kolm on Fundamental Preferences

Broome's (1993) critique of Kolm (1972) denied that a cause of preference could be an object of preference. Actually, Kolm's claim regarding fundamental preferences can usefully be split into two parts. The first claim is that a cause of preference *can* be an object of preference. We will investigate how constructing a suitable type space could help justify this claim. But a second important claim that Kolm makes, acknowledging inspiration from Tinbergen (1957), is that, for a suitably constructed type space, *all* individuals will

share identical "fundamental preferences" over (random) consequence–type pairs. This second much stronger claim remains seems to lack a formal justification, rather like the common prior assumption in Harsanyi's theory of games of incomplete information.

7.2 Ethical Type Spaces

In the approach set out in Sections 4–6, we would like each individual j's ethical type $h_j \in H$ to be both:

1. a *cause* of preference, represented by the parameter h_j of the NMUF $u_i(y; h_j)$ that individual i in any biased original position applies to lotteries faced by individual j;

2. an *object* of preference, such that the expected value $\mathbb{E}u_i(y; h_j)$ of the NMUF $u_i(y; h_j)$ represents i's preferences over lotteries in $\Delta(Y \times H)$ faced by individual j.

Recall that in mathematical analysis a *Polish space* Z is metric, complete, and separable. Given any Polish space Z, let $\mathcal{U}(Z)$ denote the (Polish space) of *continuous* bounded normalized NMUFs $u : Z \to [-1, 1]$ equipped with the sup metric defined for all $u, v \in \mathcal{U}(Z)$ by

$$d(u, v) := \sup_{z \in Z} |u(z) - v(z)|. \tag{22}$$

Note that the normalization has been chosen to allow the existence of personal consequences $z \in Z$ for which $u(z) < 0$, indicating that they are so bad as to be worse than non-existence.

Ideally, then, our aim should be to construct an *ethical type* space H of "causes of ethical preference" $h \in H$. Moreover, this space should have the defining property that H and $\mathcal{U}(Y \times H)$ are homeomorphic as metric spaces. That is, there should exist a continuous bijective mapping $\psi : H \to \mathcal{U}(Y \times H)$ whose inverse $\psi^{-1} : \mathcal{U}(Y \times H) \to H$ is also continuous. This mapping identifies each ethical type $h \in H$ with a unique utility function $\psi(h)$ with the property that $y \mapsto \psi(h)(y) = u(y; h)$ can also be regarded as a function $(y, h) \mapsto u(y; h)$ in the space $\mathcal{U}(Y \times H)$ of utility functions defined on the domain $Y \times H$. Conversely, each utility function $u \in \mathcal{U}(Y \times H)$ is identified with a uniaue ethical type $h = \psi^{-1}(u) \in H$.

It remains to be seen if this is possible, perhaps following the ideas set out first by Armbruster and Boğe (1979), and Boğe and Eisele (1979), followed by Mertens and Zamir (1985) and then by Brandenburger and Dekel (1993), amongst many others, for constructing a complete Polish type space in a game of incomplete information. In this game-theoretic setting, types are belief hierarchies about belief hierarchies, where beliefs are σ-additive probability measures.

For the time being, however, it may be enough to consider an *incomplete* ethical type space H which is homeomorphic only to some proper subset of $\mathcal{U}(Y \times H)$. That is, the mapping $\psi : H \to \mathcal{U}(Y \times H)$ is an injection rather than a bijection. For example, H could even be some (large) finite dimensional parameter space, implying that the range $\psi(H)$ of the mapping ψ is a finite parameter family of utility functions $y \mapsto u(y; h)$. Then, viewed as a parameter vector, the ethical type h can still be regarded as a cause of preference; on the other hand, when h is regarded as an argument of any function $(y, h) \mapsto u(y, h)$ in the range $\psi(H) \subset \mathcal{U}(Y \times H)$, then it becomes as an object of preference. We are assuming that there is a homeomorphism between causes of preference and a subspace of the space of utility functions that treat those causes as objects of preference; the distinction between causes and objects of preference therefore loses its significance.

7.3 Broome versus Kolm Revisited

This approach, however, leaves each individual $i \in N$ with not only their own ethical type $h_i \in H$, but also their own corresponding utility function $u_i \in \mathcal{U}(Y \times H)$. We have definitely *not* shown the existence of a truly "fundamental" utility function which is independent of i, which is what Tinbergen and Kolm presume. It remains to be seen whether it is actually possible to construct a single ethical type space such that all differences in individual's utility functions can be ascribed to differences in their ethical types. Unless, that is, one treats each individual $i \in N$ as having his or her own idiosyncratic ethical type space H_i, disjoint from H_j for each $j \in N \setminus \{i\}$, which is moreover homeomorphic to a subset of the space of personal utility functions $u_i \in \mathcal{U}(Y \times H^*)$, where $H^* := \cup_{i \in N} H_i$.

8 Concluding Remarks

8.1 Summary

Pattanaik (1968) criticized Harsanyi (1953, 1955) for failing to give a proper account of how to determine the interpersonal comparisons which determine preferences in his version of the original position, as well as Vickrey's (1945, 1960) related version. The main contribution of this paper is to show how one can circumvent Pattanaik's critique by considering preferences that represent rational choice in an extended original position that allows the individual to choose even the probabilities of becoming different individuals in various biased original positions. Of course, Hare's case for making universalizable prescriptive statements by considering only an unbiased original position retains its normative appeal. Nevertheless, considering the extended original position serves to uncover an interpersonal utility function that, given the normalization requiring the utility of non-existence to be zero, is unique up to a cardinal ratio scale.

8.2 Unsolved Problems

One of Harsanyi's claims is that all individuals would want to make the same decision in a properly defined original position. This requires all individuals to agree on each others' utility functions, given their "ethical types" that determine these functions. But these ethical types have to be not only determinants of preferences, but also objects of the preferences involved in interpersonal comparisons. This is the subject of interchange between Broome and Kolm.

Resolving the conflicting views expressed in this interchange seems possible if one can construct suitable ethical type spaces. Doing this properly, however, remains an unsolved problem. This paper approaches the issue rather like Harsanyi himself did when considering players' types in a game of incomplete information. It does not essay the kind of construction made famous in the work of Mertens and Zamir.

A somewhat related problem occurs when different individuals can have inconsistent probabilistic beliefs. This leads to issues of the kind raised in the literature on the contrast between the *ex ante* and *ex post* approaches to welfare economics that Harsanyi himself had hinted at, and which was initiated by a footnote in Diamond's (1967) renowned paper on the stock

market, followed by Starr (1973), Harris (1978), Harris and Olewiler (1979), Hammond (1981, 1982), and Nielsen (2009) — amongst others. To analyse these issues along the lines of this paper, the ethical types considered here will need supplementing by belief types. Again, however, this is an issue for future work.

Acknowledgements

This is a very welcome opportunity to express in print my deep gratitude to Prasanta Pattanaik for many years of friendship and intellectual support. In particular, his lectures at Nuffield College, Oxford, during 1969–70 did much to awaken my interest and broaden my education in social choice theory, as well as its links to ethics and political philosophy. So did the opportunity to read the typescript of what became his book Pattanaik (1971), in which, whenever a paragraph raised a question in my mind, in the next paragraph I found it answered. Furthermore, while I was visiting the University of California at Riverside in May 2007, Prasanta himself raised a question that encouraged me to revisit this topic.

Generous support from a Marie Curie Chair funded by the European Commission under contract number MEXC-CT-2006-041121 is also gratefully acknowledged. This supported the preparation of earlier versions that were presented to the conference honouring Prasanta at the University of California at Riverside in Autumn 2008, as well as to a workshop on Heterogeneous Beliefs which Mordecai Kurz organized at Stanford University in August 2009.

This 2013 revision owes much to the warm hospitality provided by the Hitotsubashi Institute for Economic Research, through its Global Centre of Excellence for the Ethical Foundations of Social Choice Theory, especially to Reiko Gotoh and Naoki Yoshihara. Many thanks also to Takashi Kunimoto for a timely reminder of the relevance of an ethical type space similar to those used in Harsanyi's theory of games of incomplete information.

References

Anscombe, F.J. and R.J. Aumann (1963) "A Definition of Subjective Probability," *Annals of Mathematical Statistics*, 34: 199–205.

Armbruster, W. and W. Böge (1979) "Bayesian Game Theory," in O. Moeschlin and D. Pallaschke (eds.) *Game Theory and Related Topics* (Amsterdam: North-Holland), pp. 17–28.

Blackorby, C., Bossert, W. and D.J. Donaldson (2005) *Population Issues in Social Choice Theory, Welfare Economics, and Ethics* (Cambridge: Cambridge University Press).

Böge, W. and T. Eisele (1979) "On Solutions of Bayesian Games," *International Journal of Game Theory*, 8: 193–215.

Brandenburger, A., and E. Dekel (1993) "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory*, 59: 189–198.

Broome, J. (1993) "A Cause of Preference Is Not an Object of Preference" *Social Choice and Welfare* 10: 57–68.

Broome, J. (1994) "Reply to Kolm" *Social Choice and Welfare* 11: 199–201.

Dasgupta, P. (1969) "On the Concept of Optimum Population" *Review of Economic Studies* 36: 295–318.

Diamond, P.A. (1967) "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty" *American Economic Review* 57: 759–776.

Drèze, J.H. (1987) *Essays on Economic Decisions under Uncertainty* Cambridge University Press: Cambridge.

Drèze, J.H. and A. Rustichini (2004) "State-Dependent Utility and Decision Theory" in S. Barberà, P.J. Hammond, and C. Seidl (eds.) *Handbook of Utility Theory, vol. 2: Extensions* (Dordrecht: Kluwer Academic) ch. 16, pp. 839–892.

Grant, S., Kajii, A., Polak, B. and Z. Safra (2010) "Generalized Utilitarianism and Harsanyi's Impartial Observer Theorem" *Econometrica* 78: 1939–1971.

Hammond, P.J. (1981) "*Ex-Ante* and *Ex-Post* Welfare Optimality under Uncertainty" *Economica* 48: 235–250.

Hammond, P.J. (1982) "Utilitarianism, Uncertainty and Information" in A.K. Sen and B. Williams (eds.) *Utilitarianism and Beyond* (Cambridge University Press, 1982), ch. 4, pp. 85–102.

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

Hammond, P.J. (1998) "Subjectively Expected Utility" in S. Barberà, P.J. Hammond, and C. Seidl (eds.) *Handbook of Utility Theory, vol. 1: Principles* (Dordrecht: Kluwer Academic) ch. 6, pp. 213-271.

Hammond, P.J. (1999) "Subjectively Expected State-Independent Utility on State-Dependent Consequence Domains" in M.J. Machina and B. Munier (eds.) *Beliefs, Interactions, and Preferences in Decision Making* (Dordrecht: Kluwer Academic), pp. 7–21.

Hare, R.M. (1961) *The Language of Morals* (London: Oxford University Press).

Harris, R. G. (1978) "Ex-post Efficiency and Resource Allocation under Uncertainty" *Review of Economic Studies* 45: 427–436.

Harris, R. G. and N. Olewiler (1979) "The Welfare Economics of Ex-Post Optimality". *Economica* 46: 137-147.

Harsanyi, J.C. (1953) "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking," *Journal of Political Economy* **61**, 434–435.

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

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

Karni, E. (1985) *Decision Making under Uncertainty: The Case of State-Dependent Preferences*. Harvard University Press: Cambridge MA.

Karni, E. (2008) "State-Dependent Preferences" in L. Blume and S. Durlauf (eds.) *The New Palgrave Dictionary of Economics, 2nd edn.* (Palgrave Macmillan).

Karni, E. and P. Mongin (2000) "On the Determination of Subjective Probability by Choices" *Management Science* 46: 233–248.

Karni, E., Schmeidler, D. and K. Vind (1983) "On State Dependent Preferences and Subjective Probabilities," *Econometrica* **51**, 1021–1031.

Kolm, S.-C. (1972; 1998) *Justice et équité* (Paris: Editions du CNRS); translated as *Justice and Equity* (Cambridge, MA: MIT Press).

Kolm, S.-C. (1994a) "The Meaning of Fundamental Preferences" *Social Choice and Welfare*, **11**, 193–198.

Kolm, S.-C. (1994b) "Rejoinder to John Broome" *Social Choice and Welfare*, **11**, 203–204.

Meade, J.E. (1955) *Trade and Welfare* (Oxford: Oxford University Press).

Mertens, J.-F. and S. Zamir (1985) "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14: 1–29.

Mongin, P. (2001) "The Impartial Observer Theorem of Social Ethics" *Economics and Philosophy* 17: 147–179.

Nielsen, C.K. (2009) "Rational Overconfidence and Social Security" available at `http://econ.korea.ac.kr/~ri/WorkingPapers/w0916.pdf`

Parfit, D. (1984) *Reasons and Persons* (Oxford: Oxford University Press).

Pattanaik, P.K. (1968) "Risk, Impersonality, and the Social Welfare Function," *Journal of Political Economy* **76**, 1152–1169.

Pattanaik, P.K. (1971) *Voting and Collective Choice* (Cambridge: Cambridge University Press).

Perlman, M. (ed.) (1974) *The Economics of Health and Medical Care*: Proceedings of a Conference Held by the International Economic Association at Tokyo.

Rawls, J. (1951) "Outline of a Decision Procedure for Ethics" *Philosophical Review* 60: 177–197.

Rawls, J. (1958) "Justice as Fairness" *Philosophical Review* 67: 164–194.

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

Rawls, J. (1982) "Social Unity and Primary Goods" in A.K. Sen and B. Williams (eds.) *Utilitarianism and Beyond* (Cambridge: Cambridge University Press).

Roberts, K.W.S. (1995) "Valued Opinions or Opinionated Values: The Double Aggregation Problem" in K. Basu, P. Pattanaik and K. Suzumura (eds.), *Choice, Welfare and Development: A Festschrift for Amartya Sen* (Oxford: Oxford University Press).

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

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

Sidgwick, H. (1887) *The Elements of Politics* (London: Macmillan).

Starr, R. (1973) "Optimal Production and Allocation Under Uncertainty" *Quarterly Journal of Economics*, 87: 81–95.

Tinbergen, J. (1957) "Welfare Economics and Income Distribution" *American Economic Review (Papers and Proceedings)* 47: 490–503

Vickrey, W.S. (1945) "Measuring Marginal Utility by the Reactions to Risk," *Econometrica* **13**: 319–333.

Vickrey, W.S. (1960) "Utility, Strategy and Social Decision Rules" *Quarterly Journal of Economics* **74**: 507–535.