14 October 2002

Accidents of Birth, Life Chances and the Impartial Observer.*

Abstract

We confront two common objections to Harsanyi's impartial observer theorem; one to do with 'fairness', and the other to do with different individuals' having different attitudes toward risk. Both these objections can be accommodated if we drop the reduction axiom; in particular, if we distinguish between 'accidents of birth' and real-world 'life chances'. We do not drop the independence axiom that has often been seen as the source of these problems. Just as Harsanyi's theorem yields a utilitarian form of social welfare function, so our approach axiomatizes a *generalized* utilitarian form. If in addition, welfare is cardinally measurable, then we can interpret the shape of our generalized utilitarian functional form in terms of attitudes toward risk and toward *interim* inequality.

Keywords: impartial observer, generalized utilitarian social welfare functions, reduction axiom, inequality of what?

JEL Classification: D63, D71

Simon Grant
Department of Economics
Rice University

Atsushi Kajii
Institute of Social and Economic Research
Osaka University

Ben Polak
Department of Economics & School of Management
Yale University

*Various versions of this paper have been presented at workshops at the University of Tasmania, Tilburg University, the University of Birmingham, Rice University, the University of British Columbia, the University of Heidelberg the University of Sydney, Yale University and Vanderbilt University. We thank participants at these workshops and also in particular, Jurgen Eichberger, Edi Karni, David Pearce, John Quiggin, Zvi Safra, John Roemer and John Weymark for many useful comments which have helped in the development of this paper.

1 Introduction

Almost fifty years ago, John Harsanyi (1953,55) proposed a theory of social choice under uncertainty leading to a social welfare function that is the sum of individual utilities. In his theory, a social policy ℓ is a lottery that yields the social outcome x with probability $\ell(x)$. Harsanyi argued that the I individuals in the society should imagine themselves in a hypothetical position as an *impartial observer* who is uncertain about the actual identity he will assume in this society. In evaluating the social policy ℓ, this impartial observer should view it as an *extended* lottery over both his identity and the social state in which there is a probability $1/I \times \ell(x)$ that he will assume the identity of person i and that the social outcome x will obtain. If all individuals are expected utility maximizers, Harsanyi argued, then it is natural that the impartial observer should also exhibit such 'Bayesian rationality'. This yields a representation for his preferences over extended-lotteries that takes the 'utilitarian' form $\sum_i \sum_x (1/I)\,\ell(x)\,u_i(x) = (1/I)\sum_i U_i(\ell)$, where $U_i(\ell)$ is the individual i's expected utility for the social policy ℓ.[1]

More than a decade later, Peter Diamond (1967) suggested that the imputation of such Bayesian rationality for the impartial observer was not consistent with some social norms of justice or 'fairness'. We can illustrate Diamond's point with an example adapted from Myerson (1981). Consider the parents of two brothers who want to set them up in their careers. Given the family's limited financial resources, they can only afford to provide one brother with higher education to become a doctor while the other brother will have to become a clerk. Let us identify the brothers as 1 and 2, where 1 is the elder. There are two relevant social states x_1, the state in which the elder brother becomes a doctor while the younger becomes a clerk, and x_2, the state in which the younger brother becomes a doctor and the elder a clerk. Suppose the parents believe the children are equally capable (ex ante) to take on either profession and equally capable of enjoying any income generated from their profession. The expected utility functions for the two children may be summarized in tabular form as follows:

[1] As Sen(1970, 1977), Weymark (1991) and others have observed, this is not welfare utilitarianism, but is just a representation of the impartial observer's preferences. We will use the term 'utilitarian form' to distinguish this from welfare utilitarianism.

	x_1	x_2
$u_1(x_j)$	1	0
$u_2(x_j)$	0	1

Table 1: Values of $u_1(.)$ and $u_2(.)$

Suppose the impartial observer believes herself equally likely to be born as either brother. In this case, *any* social policy ℓ (that is, any lottery over the two social states) yields an ex ante expected utility for the impartial observer of 1/2, since $\frac{1}{2}U_1(\ell)+\frac{1}{2}U_2(\ell) = \frac{1}{2}\ell(x_1)\,u_1(x_1) + \frac{1}{2}\ell(x_2)\,u_2(x_2) = \frac{1}{2}(\ell(x_1)+\ell(x_2)) = \frac{1}{2}$. In particular, the social policy that awards the education opportunities outright to the first-born is equally desirable to one which awards the education opportunities by flipping a 'fair' coin. Diamond argued, however, the coin-flip should be strictly preferred since it gives each child a 'fair shake' while allocation by primogeniture does not.

Several authors have tried to address this issue by relaxing the independence axiom of expected utility theory, and hence dropping utilitarian social welfare functions.[2] In our view, however, expected utility is not the source of the problem. Indeed we will show that it is possible to accommodate Diamond's notions of justice or fairness while allowing preferences to respect independence. Rather, the source of the problem is that Harsanyi's impartial observer implicitly treats as equivalent two randomizations that (we would argue) Diamond's social norm requires to be treated as distinct.

In a reply to Diamond, Harsanyi (1975) argued that, even if randomizations were of value for promoting 'fairness' (which he doubted), any randomization after birth is superfluous since 'the great lottery of (pre-)life' may be viewed as having given each child an equal chance of being the first-born. That is, in his view, the randomizations associated with 'accidents of birth' (i.e., those resolved before birth) are equivalent to randomizations like the coin-flip or other 'life chances' (i.e., those resolved after birth). His impartial observer simply multiplies through the probabilities of the accidents of birth $(1/I)$ and the life chances $(\ell(x))$ to construct a reduced single-stage lottery over identity/social-state pairs.

[2] For a discussion, see section 7.

We do not see any compelling reason, however, why the impartial observer should view these two sources of randomization as equivalent. The realm in which 'accident of births' resolve is an analytic artifact associated with imaging oneself to be in the hypothetical setting of the impartial observer. The resolution of an individual's life-chances are actually experienced by the individual. If we view these as equivalent risks then the fiction of accident of births could be used as an ex post justification for discrimination. But we suspect that it is scant comfort to an individual born into a discriminated-against group to know that her imaginary pre-life self had a equal chance of being born into privilege.

To expand this point further, imagine two societies. The first is a 'caste' society in which each individual's final position is completely determined by her accident of birth (Hollywood's cliche image of 1920's Britain perhaps). The other society is still a 'class' society in that it has the exactly the same ex post inequality as the first. In this society, however, nothing is determined by birth: each individual faces the same uncertain life chances (Hollywood's cliche image of 1920's America perhaps). Harsanyi's framework implicitly forces his impartial observer to be indifferent between these two societies; in fact, to regard them as equivalent. There may, however, be impartial observers who are 'Bayesian rational' but nevertheless have strict preferences between the two.

In this paper, therefore, we allow the impartial observer to make a distinction between (fictional) accidents of birth and (real, post-birth) life chances. We can think of this as relaxing a reduction of compound lottery axiom that is implicit in Harsanyi's framework. Once we drop reduction, we can accommodate Diamond's preference for 'fairness' while still maintaining the expected utility axioms.

There is a second (more mundane) reason for dropping reduction. In Harsanyi's thought experiment, the impartial observer is asked to imagine herself in the shoes of different individuals facing the consequences of various social policies. Harsanyi allows that different individuals will have different preferences over final social states, and assumes that the impartial observer adopts individual i's preferences when she imagines herself in individual i's shoes.[3] Each social policy, however, is associated not just with a final social state x, but with a life chance ℓ; that is, a

[3] Like Harsanyi, we assume the impartial observer faces no (additional) uncertainty over what are individual i's preferences. To see how this could be extended for the case of a Bayesian social planner, see Pearce (1995).

randomization over social states. Just as different individuals may have different preferences over final social states, so they may have different attitudes to the risks involved in these randomizations. It would seem natural for the impartial observer, when assessing the life chance ℓ from the point of view of individual i, to apply individual i's preferences about risk. In Harsanyi's analysis, however, since the impartial observer is forced to compound accidents of birth with life chances, she is forced to adopt the same attitude toward randomizations that take place before she has any identity and those that take place once she has an identity. This, in turn, implies that Harsanyi's impartial observer must have the same attitude toward risk across all her possible identities.[4]

To see how this has bite, return to the example of granting education only to one of two brothers. Suppose the impartial observer regards as equally good being the first born in state x_1 in which only the first born gets an education, or being the second born in the state x_2 in which only the second born gets the education. Suppose the impartial observer also regards as equally good being the first born in state x_2 or being the second born in state x_1. In this case, Harsanyi's axioms force the impartial observer to be indifferent between being the first born and having a fifty-fifty chance of x_1 and x_2, or being the second born and having the same fifty-fifty chance. But suppose the confident first born is risk loving while the timid second born is extremely risk averse. In this case, it seems entirely rational for an impartial observer, if she knows that society is going to use the fifty-fifty allocation rule, to prefer being the first born.

Once we drop reduction, the problem of different risk attitudes disappears. The impartial observer can assess each individual's life-chance risk using the risk attitudes of that individual, and she can assess the randomization in accidents of birth using her own attitudes toward that (possibly distinct form of) risk.

Section 2 develops our framework with a distinction between accidents of birth and (post-birth) life chances. Section 3 provides representations for the impartial observer's preferences under different assumptions. We show that if we impose the reduction of compound lotteries axiom (which is implicit in Harsanyi's framework), then we re-obtain his *utilitarian form* of social welfare function, $(1/I)\sum_i U_i(\ell)$. If we relax reduction but retain the other assumptions (including

[4] We think that this point was first made, as an aside, by Pattanaik (1968).

the independence axiom of 'expected utility'), then we obtain a *generalized utilitarian form* of social welfare function, $(1/I)\sum_i \phi_i(U_i(\ell))$, where, again, each $U_i(\ell)$ is individual i's expected utility for the social policy ℓ. This is a form of social welfare function much discussed in welfare economics. Harsanyi's utilitarian form corresponds to the special case where each ϕ_i-transformation is affine. In section 4, we show that Diamond's social preferences correspond to a preference for randomization to take place post birth — that is, for life chances over accidents of birth — and that this preference forces the ϕ_i-transformations to be concave. We also consider an impartial observer who, although she rejects Diamond's fairness critique of Harsanyi, nevertheless wishes to respect different individuals' risk attitudes. Allowing for different risk attitudes also entails non-linear ϕ_i-transformations. Section 5 discusses our results in relation to 'welfarism'. We introduce an explicit notion of comparable welfare and use it to interpret our functional forms and how they relates to fairness, risk aversion and inequality. Section 6 provides a more general representation result. Section 7 discusses some of the related literature.

2 The Setting

Society consists of a finite set of individuals $\mathcal{I} = \{1, \ldots, I\}$ with generic elements given by i and j. The set of final consequences or outcomes that individuals in this society ultimately care about is denoted by $\mathbf{X}$ with generic element x. The set $\mathbf{X}$ is assumed to be a compact metrizable space and associated with it is the set of events $\mathcal{E}$, which is taken to be the Borel sigma-algebra of $\mathbf{X}$. Let $\triangle(\mathbf{X})$ (with generic element ℓ) denote the set of life chances; that is the set of probability measures on $(\mathbf{X}, \mathcal{E})$ endowed with the weak convergence topology. For each social-state x in $\mathbf{X}$, δ^x is the degenerate lottery that assigns probability weight 1 to social state x.

Each individual i in $\mathcal{I}$, is endowed with a preference relation $\succsim_i$ defined over the set of life-chances $\triangle(\mathbf{X})$. We assume throughout that for each i in $\mathcal{I}$, $\succsim_i$ is a complete, transitive and continuous binary relation on $\triangle(\mathbf{X})$, and that its asymmetric part, $\succ_i$, is non-empty. Hence for each $\succsim_i$ there exists a non-constant function $V_i : \triangle(\mathbf{X}) \to \mathbb{R}$, satisfying for any ℓ and ℓ' in $\triangle(\mathbf{X})$, $V_i(\ell) \geq V_i(\ell')$ if and only if $\ell \succsim_i \ell'$. In summary, a society may be characterized by the tuple $\langle \mathbf{X}, \mathcal{E}, \mathcal{I}, \{\succsim_i\}_{i\in\mathcal{I}} \rangle$.

By hypothesis, the impartial observer for this society is cloaked by a veil of ignorance and is uncertain about which identity he will assume. Harsanyi and others considered the set of probability measures defined over identity/outcome pairs $(i, x) \in \mathcal{I} \times \mathbf{X}$. Let $\triangle(\mathcal{I} \times \mathbf{X})$ denote this space of 'extended lotteries'. We want, however, a framework in which a distinction can be drawn between lotteries that resolve at or before birth and lotteries that resolve after birth. The first is a 'hypothetical' stage in which the impartial observer is unaware which identity he will assume or which life chance will obtain. The second is a 'real' stage in which the individual now knows her identity but still faces further uncertainty about which social state will ultimately obtain. That is, immediately after her birth, an individual faces a pair $(i, \ell) \in \mathcal{I} \times \triangle(\mathbf{X})$ in which the first element is her identity and the second element is her remaining life chances. To model the hypothetical uncertainty facing the impartial observer, let $\triangle(\mathcal{I})$ denote the set of lotteries on I with typical element $z = (z_1, \ldots, z_I) \in [0,1]^I$ $(\sum_{i \in \mathcal{I}} z_i = 1)$. We interpret these as lotteries that resolve at birth over identity. Let $\triangle_0(\triangle(\mathbf{X}))$ denote the set of simple lotteries (i.e., measures with finite support) on $\triangle(\mathbf{X})$, with typical element P. We interpret these as lotteries that resolve at birth over life chances. That is, each P is a two-stage lottery over social states in which the first stage is resolved at birth and the second stage (the life chance) resolves after birth. We assume that the impartial observer views the birth lottery over identity and the birth lottery over future life chances as independent. Thus, the impartial observer is faced with a pair $(z, P) \in \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$.

Each particular society or social policy is associated with a particular life chance ℓ. Therefore, we will later have interest in the subset of (z, P) in which P is degenerate in the first (pre-birth) stage. Let δ^ℓ denote the degenerate birth-lottery on the life-chance ℓ; that is, the element of $\triangle_0(\triangle(\mathbf{X}))$ that assigns probability one to the life-chance ℓ. Let $\triangle_0^D(\triangle(\mathbf{X}))$ denote the set of birth lotteries that are degenerate in the first stage. A typical element $\left(z, \delta^\ell\right) \in \triangle(\mathcal{I}) \times \triangle_0^D(\triangle(\mathbf{X}))$ is then a two-stage randomization in which pre-birth uncertainty is only about identity and all uncertainty over which social state x will obtain is resolved post birth.

The space $\triangle_0(\triangle(\mathbf{X}))$ has a natural linear structure. That is, if P and Q are elements of $\triangle_0(\triangle(\mathbf{X}))$ then for any α in $[0,1]$, $\alpha P + (1-\alpha)Q$ is the element of $\triangle_0(\triangle(\mathbf{X}))$ that assigns to

each life chance ℓ in $\triangle(\mathbf{X})$ the pre-birth probability $\alpha P(\ell)+(1-\alpha)Q(\ell)$. Using this rule, for any finite list $(\ell_1, \ldots, \ell_N)$ where each ℓ_n is in $\triangle(\mathbf{X})$, the convex combination $\sum_{n=1}^{N} q_n \delta^{\ell_n}$ is the element of $\triangle_0(\triangle(\mathbf{X}))$ that assigns to each life-chance ℓ in $\triangle(\mathbf{X})$, the probability $\sum_{\{n:\ell_n=\ell\}} q_n$. With slight abuse of notation, we shall write $P = \left[(\ell_n; q_n)_{n=1}^{N}\right]$ to denote the birth-lottery over life-chances $\sum_{n=1}^{N} q_n \delta^{\ell_n}$, even though the ℓ_n's need not all be distinct.

Each birth-life lottery $(z, \left[(\ell_n; q_n)_{n=1}^{N}\right])$ in $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$, could be 'reduced' to an extended-lottery of the type considered by Harsanyi in $\triangle(\mathcal{I} \times \mathbf{X})$ by evaluating the probability of each identity/social-state pair (i, x) by multiplying through the probabilities. The reduced extended-lottery is a product measure with marginal distribution over $\triangle(\mathcal{I})$ equal to z and a marginal distribution over $\triangle(\mathbf{X})$ equal to $\sum_{n=1}^{N} q_n \ell_n$. Such a reduction, however, subsumes the explicit temporal structure and conflates the two distinct stages of resolutions of uncertainty, one at birth and the other post-birth, into a single 'atemporal' stage.

An example may help clarify the structure of the model. Consider the lotteries in the Diamond example in the Introduction. There are two individuals and four identity/social-state pairs. Assume that the impartial observer thinks herself equally likely to be the older or the younger brother, so $z = (1/2, 1/2)$. The social policy that gives the better education outright to the older brother yields the life chance δ^{x_1} with probability one, and so induces the birth lottery over life chances $P_1 := [(\delta^{x_1}; 1)]$. The social policy that allocates the better education by flipping a fair coin yields the life chance $\frac{1}{2}\delta^{x_1} + \frac{1}{2}\delta^{x_2}$, and so induces the birth lottery over life chances $P_2 := \left[\left(\frac{1}{2}\delta^{x_1} + \frac{1}{2}\delta^{x_2}; 1\right)\right]$. Both (z, P_1) and (z, P_2) are elements of $\triangle(\mathcal{I}) \times \triangle_0^D(\triangle(\mathbf{X}))$; that is, the only randomness resolved at birth is that over identity. We could also consider a third possibility in which the policy of whether to allocate the good outright to the elder brother or to the younger brother is determined with equal probability by a hypothetical pre-birth lottery. This possibility yields life chances δ^{x_1} or δ^{x_2} each with equal pre-birth probability, and so induces the birth lottery over life chances $P_3 := \left[\left(\delta^{x_1}; \frac{1}{2}\right), \left(\delta^{x_2}; \frac{1}{2}\right)\right]$. Notice that the birth-life lotteries (z, P_2) and (z, P_3) reduce to the same uniform distribution over the four identity/outcome pairs, but in (z, P_3) all randomness is resolved in the hypothetical stage before birth while recall that in (z, P_2) only identity is resolved at birth. Harsanyi would regard these two as equivalent, but we would argue

that only (z, P_2) gives each brother a real "fair shake".

To allow for preferences that might distinguish such birth-life lotteries, we endow the impartial observer with a preference relation $\succsim$ defined over $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$. We assume throughout that $\succsim$ is complete, transitive and continuous, and so admits a continuous representation $V : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$. That is, for any pair of birth-life lotteries, (z, P) and (z', P'), $(z, P) \succsim (z', P')$ if and only if $V(z, P) \geq V(z', P')$.

3 The Generalized Utilitarian Impartial Observer

In this section, we consider axioms that restrict the preferences of the impartial observer. We start from what are essentially Harsanyi's axioms adapted to our framework. From these we obtain Harsanyi's utilitarian form of social welfare function. We then relax axioms in order to accommodate Diamond's critique, and so obtain a generalized utilitarian representation.

The first axiom is the analog of Harsanyi's 'acceptance principle'. Consider a particular identity i and a degenerate birth-life lotteries of the form $\left(\delta^i, \delta^\ell\right)$. A lottery of this form yields identity i and life chance ℓ for sure at birth, so there is no pre-life uncertainty. Since the impartial observer knows that she is going to be identity i for sure, it is natural to require the impartial observer's preferences $\succsim$ to coincide with that individual's preferences $\succsim_i$ over life chances.

Axiom 1 (Acceptance Principle) *For all i in $\mathcal{I}$ and all ℓ, ℓ' in $\triangle(\mathbf{X})$, $\ell \succsim_i \ell'$ if and only if $\left(\delta^i, \delta^\ell\right) \succsim \left(\delta^i, \delta^{\ell'}\right)$.*

Second, Harsanyi assumes that each individual i's preferences satisfy the standard independence axiom on the set of life chances, $\triangle(\mathbf{X})$. We refer to this as post-birth independence.

Axiom 2 (Individuals' Post–birth Independence) *For all i in $\mathcal{I}$, for all life chances ℓ, ℓ', ℓ'' in $\triangle(\mathbf{X})$, and all α in $(0, 1]$, $\ell \succsim_i \ell'$ if and only if $\alpha\ell + (1-\alpha)\ell'' \succsim_i \alpha\ell' + (1-\alpha)\ell''$.*

Third, Harsanyi assumes that the impartial observer's preferences also satisfy independence, but on the set $\triangle(\mathcal{I} \times \mathbf{X})$. In our framework, the impartial observer's preferences are over the product space $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$, so there are two dimensions in which probability mixtures are well

defined: mixtures over birth lotteries and over future life chances. The following axiom adapts Safra & Weissengrin's (2002) substitution axiom.

Axiom 3 (Impartial Observer's Pre-Birth Independence) *Fix elements z, z', $\tilde{z}$ and $\tilde{z}'$ of $\triangle(\mathcal{I})$ and elements $P, P', \tilde{P}$ and $\tilde{P}'$ of $\triangle_0(\triangle(\mathbf{X}))$. If $(z,P) \sim (z',P')$, then for all α in $(0,1]$,*

(a) $(\tilde{z},P) \succsim (\tilde{z}',P')$ *if and only if* $(\alpha\tilde{z}+(1-\alpha)z,P) \succsim (\alpha\tilde{z}'+(1-\alpha)z',P')$

(b) $(z,\tilde{P}) \succsim (z',\tilde{P}')$ *if and only if* $(z,\alpha\tilde{P}+(1-\alpha)P) \succsim (z',\alpha\tilde{P}'+(1-\alpha)P')$

(c) $(\tilde{z},P) \succsim (\tilde{z}',P')$ *if and only if* $(\alpha\tilde{z}+(1-\alpha)z,P) \succsim (z',\alpha\tilde{P}'+(1-\alpha)P')$

Finally, by use of his framework, Harsanyi is implicitly imposing the following reduction of compound lottery principle.

Axiom 4 (Reduction) *The impartial observer's preference relation $\succsim$ satisfies* reduction *if for all pairs of birth-lotteries over life-chances $P = \left[(\ell_m; q_m)_{m=1}^M\right]$ and $P' = \left[(\ell'_n; q'_n)_{n=1}^N\right]$ in $\triangle_0(\triangle(\mathbf{X}))$, $\sum_{m=1}^M q_m\ell_m = \sum_{n=1}^N q'_n\ell'_n$ implies $(z,P) \sim (z,P')$ for all z in $\triangle(\mathcal{I})$.*

If we impose all the above axioms then we get Harsanyi's utilitarian form of social-welfare function. In fact, individuals' post-birth independence is implied by the other axioms and so is redundant.

Proposition 1 (Utilitarianism) *The following are equivalent:*

(a) *The impartial observer's preferences $\succsim$ satisfies the acceptance principle, pre-birth independence, and reduction.*

(b) *There exists a function $U : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$ and functions $u_i : \mathbf{X} \to \mathbb{R}$ $i = 1,\ldots,I$, such that U represents $\succsim$; for each i, $\int u_i(x)\,\ell(dx)$ represents $\succsim_i$; and for all $\left((z_1,\ldots,z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right)$ in $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$,*

$$U\left((z_1,\ldots,z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = \sum_{i=1}^I \sum_{n=1}^N z_i q_n U_i(\ell_n).$$

where $U_i(\ell_n) := \int u_i(x)\,\ell_n(dx)$.

Moreover for any functions $\hat{U} : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$ *and* $\hat{u}_i : \mathbf{X} \to \mathbb{R}, i = 1, \ldots, I$, *such that* $\hat{U}$ *represents* $\succsim$; *for each* i, $\int \hat{u}_i(x)\,\ell(dx)$ *represents* $\succsim_i$, *then* $\hat{U}\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = \sum_{i=1}^I \sum_{n=1}^N z_i q_n \int \hat{u}_i(x)\,\ell_n(dx)$ *holds if and only if there exists* $a > 0$ *and* $b \in \mathbb{R}$, *such that* $\hat{u}_i = a u_i + b$, *for all* $i = 1, \ldots, I$.

Notice that if we restrict attention to the case where the birth lottery over identity is uniform and where the birth lottery over life chances is degenerate on ℓ, then this social welfare function reduces to the familiar utilitarian form $\frac{1}{I}\sum_{i=1}^I U_i(\ell)$.

Recall that reduction forces the impartial observer to treat as equivalent fictional randomizations that take place before birth and real randomizations (or life chances) that occur post birth. Suppose then that we drop reduction but keep the other Harsanyi axioms in place. When we drop reduction from the axioms of proposition 1, we no longer get individual's post-birth independence for free, so we have to impose it directly. When we do this, however, we get a generalized utilitarian form of social welfare function.

Proposition 2 (Generalized Utilitarianism) *The following are equivalent:*

(a) *The impartial observer's preferences* $\succsim$ *satisfy the acceptance principle, and pre–birth independence, and the individuals' preferences* $(\succsim_i)_{i \in \mathcal{I}}$ *satisfy post-birth independence;*

(b) *There exists a function* $V : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$, *functions* $u_i : \mathbf{X} \to \mathbb{R}$ *and increasing functions* $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, I$, *such that* V *represents* $\succsim$; *for each* i, $\int u_i(x)\,\ell(dx)$ *represents* $\succsim_i$, *and for all* $\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right)$ *in* $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$,

$$V\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = \sum_{i=1}^I \sum_{n=1}^N z_i q_n \phi_i \circ U_i(l_n).$$

where $U_i(\ell_n) := \int u_i(x)\,\ell_n(dx)$.

Moreover for any functions $\hat{V} : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$, $\hat{u}_i : \mathbf{X} \to \mathbb{R}$, *and* $\hat{\phi}_i : \mathbb{R} \to \mathbb{R}$ $i = 1, \ldots, I$, *such that* $\hat{V}$ *represents* $\succsim$; *for each* i, $\hat{U}_i(\ell) := \int \hat{u}_i(x)\,\ell(dx)$ *represents* $\succsim_i$, *then* $\hat{V}\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = \sum_{i=1}^I \sum_{n=1}^N z_i q_n \hat{\phi}_i \circ \hat{U}_i$ *holds if and only if there exists* $a > 0$ *and* $b \in \mathbb{R}$, *such that* $\hat{\phi}_i \circ \hat{U}_i = a\phi_i \circ U_i + b$, *for all* $i = 1, \ldots, I$.

If we restrict attention to the case where the birth lottery over identity is uniform and where the birth lottery over life chances is degenerate on ℓ, then this social welfare function reduces to the generalized utilitarian form $\frac{1}{I}\sum_{i=1}^{I}\phi_i \circ U_i(\ell)$.

So far, we have placed no restrictions on the shape of the transformation functions ϕ_i except that they are increasing. Notice also that there is a different function ϕ_i for each individual i. The reason for this will become clearer below. The next section addresses both Diamond's 'fairness' critique and the problem that the impartial observer should respect different agents' different risk-attitudes. The discussion suggests interpretations for the ϕ_i-functions, and restrictions of their shape.

4 Fairness and Risk Aversion.

Accommodating a preference for Fairness. The intuition of Diamond's example (and of class versus caste societies) suggest not only that reduction will fail but that the impartial observer may prefer randomizations to take place in the real post-birth world rather than the imaginary pre-birth thought experiment. Consider life chances $\overline{\ell}$, ℓ, and ℓ' in $\triangle(\mathbf{X})$ with $\overline{\ell} = \beta\ell + (1-\beta)\ell'$. Compare the birth lotteries over life chances $\delta^{\overline{\ell}}$ and $\beta\delta^{\ell} + (1-\beta)\delta^{\ell'}$ (or equivalently $[(\overline{\ell};1)]$ and $[(\ell;\beta),(\ell';1-\beta)]$). Since $\overline{\ell} = \beta\ell + (1-\beta)\ell'$, the two birth lotteries over life chances induce the same probability distribution over social states, and therefore would reduce to the same lottery in $\triangle(\mathbf{X})$. In the degenerate birth lottery over life chances $\delta^{\overline{\ell}}$, however, since the life chance $\overline{\ell}$ is chosen with probability one before birth, the entire randomization over social states resolves after birth. On the other hand, in the birth lottery over life chances $\beta\delta^{\ell} + (1-\beta)\delta^{\ell'}$, the randomization over social states is partially resolved before birth. Thus, less is determined pre-birth in the first case than in the second. This motivates the following definition.

Axiom 5 (Preference for Post-birth Randomization.) *The impartial observer's preference relation exhibits* preference for post-birth randomization*, if, for any pair of life chances ℓ and ℓ' in $\triangle(\mathbf{X})$, and any β in $[0,1]$, $(z, \delta^{\beta\ell+(1-\beta)\ell'}) \succsim (z,\ \beta\delta^{\ell} + (1-\beta)\delta^{\ell'})$ holds for all z in $\triangle(\mathcal{I})$.*

One could also imagine an impartial observer whose preferences go the other way. Accordingly define *preference for pre-birth randomization* by reversing the direction of preference (i.e., $(z,\ \beta\delta^{\ell} +$

$(1-\beta)\,\delta^{\ell'}) \succsim (z, \delta^{\beta\ell+(1-\beta)\ell'}))$ in the above axiom. Define neutrality between pre- or post-birth randomization as the conjunction of the two axioms.

If we impose preference for post-birth randomization (loosely speaking, as a weaker replacement for Harsanyi's reduction axiom), then we obtain the sometimes-desired concavity of the ϕ_i-functions in our generalized utilitarian representation.

Proposition 3 (Concave Generalized Utilitarianism) *The impartial observer's preferences and the individuals' preferences satisfy the conditions of Proposition 2 part (a) and the impartial observer's preferences satisfy preference for post- (pre-) birth randomization if and only if the ϕ_i-functions defined in Proposition 2 part (b) are concave (convex).*

If the transformation functions ϕ_i are linear (i.e., concave and convex) then we are back to Harsanyi's utilitarian form of social welfare function from Proposition 1. This concurs with the above discussion of the axioms: requiring an impartial observer to be neutral between pre- and post-birth randomization is equivalent to imposing the reduction axiom.

We next show that these concave generalized utilitarian social welfare functions accommodate Diamond's notion of fairness. To generalize the idea, suppose that from the point of view of the impartial observer, being agent i and obtaining life chance ℓ is equally good as being agent j and obtaining life chance ℓ'. Moreover, being agent i and obtaining life chance ℓ' is equally good as being agent j and obtaining life chance ℓ. Diamond's notion of fairness suggests that the impartial observer should prefer being i for sure and then having a post-birth coin flip between ℓ or ℓ', to having a pre-birth coin flip between being i or j and then obtaining the life chance ℓ for sure. Notice that, in Diamond's example, each particular society or social policy is associated with a particular life chance ℓ, ℓ' or the post-birth flip $\frac{1}{2}\ell + \frac{1}{2}\ell'$. Thus, in Diamond's example (unlike in our axiom), *pre-birth* randomization is not over whether ℓ, ℓ' or $\frac{1}{2}\ell + \frac{1}{2}\ell'$ obtains at birth. Instead, Diamond's pre-birth randomization just determines, in a given society, which individual i's identity the impartial observer will assume at birth.[5] Nevertheless, our axiom does the job.

[5] Formally, Diamond only needs to consider the space $\triangle(\mathcal{I}) \times \triangle_0^D(\triangle(\mathbf{X}))$.

Corollary 4 (Diamond's Preferences) *Consider any pair of individuals i and j, and any two life chances ℓ and ℓ', such that $(\delta^i, \delta^\ell) \sim (\delta^j, \delta^{\ell'})$ and $(\delta^i, \delta^{\ell'}) \sim (\delta^j, \delta^\ell)$. If the impartial observer's preferences satisfy the conditions of Proposition 3, then $(\delta^i, \delta^{\alpha\ell+(1-\alpha)\ell'}) \succsim (\alpha\delta^i + (1-\alpha)\,\delta^j, \delta^\ell)$ for all α in $[0,1]$.*

To summarize: If we replace reduction with preference for post-birth randomization, then our impartial observer respects Diamond's notion of fairness but still satisfies independence. All that Diamond's preference requires of our representation is that the ϕ_i-functions be concave.

One might conjecture that if we substitute the preference for pre-birth *identity* randomizations described in Corollary 4 in place of our preference for pre-birth *life-chance* randomizations in Axiom 5, we would still obtain the concave ϕ_i-transformations in Proposition 3. The problem is that, without more structure on outcomes and ex post preferences, we cannot be sure that for all individuals i and life chances ℓ there will exist a second individual j and a second life chance ℓ' such that $(\delta^i, \delta^\ell) \sim (\delta^j, \delta^{\ell'})$ and $(\delta^i, \delta^{\ell'}) \sim (\delta^j, \delta^\ell)$.

Accommodating different risk attitudes. We next turn to the problem that different individuals may have different attitudes toward risk. Recall that we might imagine the impartial observer being indifferent between being i with a good outcome for i and being j with a good outcome for j, and also being indifferent between being i with a bad outcome for i and being j with a bad outcome for j. But, if i is less risk averse than j, then we might expect that, if the impartial observer knows she will be faced with a lottery between the relevant good and bad outcomes, she would rather be agent i. Once we drop reduction, this is no problem. Loosely speaking, all we require is for the function ϕ_i to be a concave transformation of ϕ_j. The next proposition makes this more precise.

Proposition 5 (Different Risk Attitudes.) *Suppose that individuals and the impartial observer's preferences satisfy the conditions of Proposition 3 and so admit the representation $\sum_{i=1}^{I}\sum_{n=1}^{N} z_i q_n \phi_i \circ U_i(l_n)$. Consider any pair of individuals i and j. For any four life chances ℓ,ℓ',$\tilde{\ell}$, $\tilde{\ell}'$ if $(\delta^i, \delta^\ell) \sim (\delta^j, \delta^{\tilde{\ell}}) \succ (\delta^i, \delta^{\ell'}) \sim (\delta^j, \delta^{\tilde{\ell}'})$ implies $(\delta^i, \delta^{\alpha\ell+(1-\alpha)\tilde{\ell}}) \succsim (\delta^j, \delta^{\alpha\ell'+(1-\alpha)\tilde{\ell}'})$ for any α in $[0,1]$, then $\phi_i^{-1} \circ \phi_j$ is a convex function on the domain $[U_j(\tilde{\ell}'), U_j(\tilde{\ell})]$.*

We could imagine an impartial observer who rejects Diamond's fairness critique of Harsanyi but who nevertheless wishes to respect different individuals' risk attitudes. Such an impartial observer might not view accidents of birth and life chances as intrinsically different and so might accept reduction between pre- and post-birth randomization provided that both randomizations are faced by the same agent. Our framework can also accommodate this case. Suppose that (outside the thought experiment) this impartial observer was in fact agent i. Then from the position of the impartial observer, her social preferences would have a linear ϕ_i-function, but non-linear ϕ_j-functions to allow for the different risk attitudes of other agent's into whose identities she might be born.

5 Welfare and Inequality

In this section, we introduce an explicit notion of comparable welfare, and use it to interpret our functional form, and how it relates to fairness and risk aversion. Let $w_i : \triangle(\mathbf{X}) \to \mathbb{R}$ be agent i's welfare function, and let $w : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$ be the impartial observer's welfare function. These w_i's are functions of life chances rather than final outcomes, so think of them as defining *interim* (rather than *ex post*) welfares. Similarly, we can think of the impartial observer's welfare function w as defining her *ex ante* welfare. Let us assume that these welfare functions guide choice. That is,

Axiom 6 (Congruence) *For each individual i in $\mathcal{I}$ and for all ℓ, ℓ' in $\triangle(\mathbf{X})$, $\ell \succsim_i \ell'$ if and only if $w_i(\ell) \geq w_i(\ell')$. For the impartial observer, for all (z, P) and (z', P') in $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$, $(z, P) \succsim (z', P')$ if and only if $w(z, P) \geq w(z', P')$.*

Following Weymark (1991), let us further assume that the impartial observer adopts the welfare of agent i when she puts herself in the shoes of agent i.

Axiom 7 (Principle of Welfare Identity) *For each individual i in $\mathcal{I}$ and for all ℓ in $\triangle(\mathbf{X})$, $w_i(\ell) = w(\delta^i, \delta^\ell)$.*

For the remainder of this section, we assume that both of these axioms apply. As Weymark (1991) notes, taken together, Congruence and the Principle of Welfare Identity imply the Principle

of Acceptance. Furthermore, they entail that for any pair of individuals i and j, and any pair of life-chances ℓ and ℓ', the ranking between $\left(\delta^i, \delta^\ell\right)$ and $\left(\delta^j, \delta^{\ell'}\right)$ is completely determined by the ranking between $w_i(\ell)$ and $w_j(\ell')$. That is, the welfare functions $(w_1(.), \ldots, w_I(.))$ are at least *ordinally measurable* and *fully comparable*.

To relate these welfare measures to the representation of the Impartial Observer's preferences obtained in Proposition 2, define for each individual i, the function $g_i : \mathbb{R} \to \mathbb{R}$ that maps individual i's interim welfare to his von Neumann-Morgenstern utility. That is, for each individual i, and for all ℓ in $\triangle(\mathbf{X})$:

$$g_i(w_i(\ell)) \equiv U_i(\ell). \tag{1}$$

Similarly, let $g : \mathbb{R} \to \mathbb{R}$ denote the mapping from the impartial observer's ex ante welfare to her von Neumann-Morgenstern utility. Thus, if the conditions of Proposition 2 apply, then for all $\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right)$ in $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$:

$$g\left(w\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right)\right) \equiv \sum_{i=1}^{I}\sum_{n=1}^{N} z_i q_n \phi_i \circ U_i(l_n) \tag{2}$$

Given this, we can now re-interpret the functions ϕ_i in terms of welfare. Applying the Principle of Welfare Identity, we get $\phi_i \circ U_i(l_n) = g\left[w(\delta^i, \delta^{\ell_n})\right] = g\left[w_i(\ell_n)\right] = g\left[g_i^{-1}(U_i(\ell_n))\right]$. Thus, for each individual i, the function ϕ_i is given by the function $g \circ g_i^{-1}$.

We can re-express the social welfare function from Proposition 2 in terms of our g-functions and welfare to yield

$$w\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = g^{-1}\left(\sum_{i=1}^{I}\sum_{n=1}^{N} z_i q_n g(w_i(\ell_n))\right). \tag{3}$$

With only *ordinally measurable* welfares, the shape and hence degree of curvature of g can vary as one considers different (common) monotonic transformations of $(w_1(.), \ldots, w_I(.))$. As such utility only has meaning as a representation of social preferences. In this case (following Sen (1977)), we can no more interpret the shape of g than can Harsanyi interpret his social welfare function as being linear in welfare. However, since ϕ_i is equal to $g \circ g_i^{-1}$, it remains *invariant* to any common increasing transformations of the welfare functions $(w_1(.), \ldots, w_I(.))$. That is, if we take $(\hat{w}_1(.), \ldots, \hat{w}_I(.))$, where $\hat{w}_i = h \circ w_i$ and $h : \mathbb{R} \to \mathbb{R}$ is an increasing function, then the

functions that map the transformed welfares to the von Neumann-Morgenstern utilities are now given by $\hat{g}_i = g_i \circ h^{-1}$ and $\hat{g} = g \circ h^{-1}$. And so, $\hat{\phi}_i \equiv \hat{g} \circ \hat{g}_i^{-1} = g \circ h^{-1} \circ h \circ g_i = g \circ g_i^{-1}$. This provides some intuition for why the ϕ_i-functions are unique up to positive affine transformations.

Suppose, however, that welfares are cardinally measurable. In this case, we can give more interpretation to the g-functions. In particular, if we restrict attention to the case where the birth lottery over identity is uniform and where the birth lottery over life chances is degenerate on ℓ (that is, there is no pre-birth uncertainty about the social policy ℓ), we can associate the representation in expression (3) with a Bergson-Samuelson social welfare function W that maps vectors of individual interim welfares to 'social welfare' (that is, the impartial observer's ex ante welfare, before she knows whom she will become):

$$W(w_1(\ell), \ldots, w_I(\ell)) := w\left(\left(\frac{1}{I}, \ldots, \frac{1}{I}\right), \delta^\ell\right) = g^{-1}\left(\frac{1}{I}\sum_{i=1}^{I} g\left(w_i(\ell)\right)\right).$$

In this representation, since we have imposed cardinal measurability, g is uniquely-defined up to positive affine transformations. In this case, the degree of concavity of g may be interpreted as measuring the degree of the impartial observer's aversion to *interim* welfare inequality or her attitudes toward the risks embodied in accidents of birth.

This is exactly what we should expect. Had we started from the viewpoint of a Bergson-Samuelson social welfare function, then we would immediately have interpreted Diamond's notion of fairness as aversion to interim welfare inequality. This in turn would have led us to a social welfare function which is concave in individual interim utilities. Instead we started from the viewpoint of Harsanyi's impartial observer but dropped reduction to accommodate Diamond's critique. If we now impose cardinally measurable welfare, we arrive at the same point.

An explicit notion of welfare also helps us interpret Proposition 5. Recall that the impartial observer is more willing to take on risks in the identity of i than in the identity of j if and only if ϕ_i is a concave transformation of ϕ_j. If welfare is cardinally measurable, then ϕ_i is a concave transformation of ϕ_j if and only if g_i is a convex transformation of g_j. This corresponds to our usual notion of income risk aversion except that instead of being risk averse over income, our individuals are risk averse over their final (ex post) welfares. Individual i is more welfare risk averse than individual j. In other words, each function g_i captures individual i's attitudes toward

the welfare risk embodied in her life chances. If we want the functions ϕ_i to be the same for all individual's i then we require all individuals to have the same attitudes toward welfare risk.

Once we allow our social welfare function to take into account that different agents may have different degrees of welfare risk aversion, we may in fact no longer wish to accept Diamond's fairness axiom. That is, there are cases where the impartial observer may actually prefer accidents of birth to life chances. Suppose our two brothers are extremely (ex post) welfare risk averse, but our impartial observer is only mildly (interim) welfare inequality averse. In welfare terms, the functions g_1 and g_2 might be more concave than is the function g. In this case, the ϕ_i functions are convex. The impartial observer, anticipating the discomfort that real-life uncertainty would cause the brothers, prefers to absorb the risk in the imaginary world before they are born. More generally, if there are at least some very welfare risk averse agents in the real world then it no longer follows that we would prefer to expose them to real world ex post welfare risk. If we want each ϕ_i function to be concave, we require the function g to be more concave than each of the individual's g_i functions.

Although the case of convex ϕ_i functions may seem odd, it corresponds to an intuition sometimes used to defend caste-like societies. Preferring accidents of birth to life chances corresponds to preferring risks that resolve early. In another paper (Grant, Kajii & Polak (1998)), for a general dynamic setting in which all risks are real and not imaginary, we argued that a preference for early resolution corresponds to an instrinsic preference for information. Anxious agents may prefer to know their fate soon. In the context of the impartial observer, if individuals are highly risk averse over their ex post welfares, they may prefer for uncertainty to have been resolved by the time they are born. They might prefer "to know their place".

The following Proposition summarizes the above discussion.

Proposition 6 (Welfare Interpretation) *Suppose that individuals and the impartial observer's preferences satisfy the conditions of Proposition 3 and so admit the representation* $\sum_{i=1}^{I}\sum_{n=1}^{N} z_i q_n \phi_i \circ U_i(l_n)$. *Suppose further that the Principle of Welfare Identity and Congruence are satisfied, and that welfare is cardinally measurable. Then each* $\phi_i = g \circ g_i^{-1}$ *where*

(a) *the function* g_i *maps individual* i*'s welfare to* i*'s von Neumann-Morgenstern utility. The*

concavity of this function reflects i's attitude toward the welfare risks embodied in life chances.

(b) *the function g maps the impartial observer's welfare (which is identical to individual interim welfare) to her von Neuman-Morgenstern utility. The concavity of this function captures the impartial observer's attitudes toward interim welfare inequality and/or the risks embodied in accidents of birth.*

(c) *The impartial observer's preferences will satisfy preference for post-both randomization (and hence respect Diamond's notion of fairness) if and only if g is a concave transformation of g_i for all i; that is, the impartial observer's aversion to interim welfare inequality and/or to accident of birth risk outweighs each individual's aversion to life chance risk.*

(d) *The functions ϕ_i will be identical if and only if the functions g_i are identical; that is, each individual i has the same attitude toward welfare risk.*

6 A More General Representation Result

Recall that, in Proposition 2, without reduction, we had to impose that individuals' preferences satisfy post-birth independence. If we do not impose expected utility on the individuals then we still get a representation that is *additive* across monotonic transformations of the individuals' life-chance utilities as well as being *bi-linear* in the pre-birth (first-stage) probabilities (i.e. the z_i's and q_n's).

Theorem 7 *The following are equivalent:*

(a) *The impartial observer's preferences $\succsim$ satisfy the acceptance principle and pre-birth independence.*

(b) *There exist a function $V : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$ and functions $V_i : \triangle(\mathbf{X}) \to \mathbb{R}$ $i = 1, \ldots, I$, such that V represents $\succsim$; for each i, V_i represents $\succsim_i$; and for all $\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right)$ in $\triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$,*

$$V\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = \sum_{i=1}^{I} \sum_{n=1}^{N} z_i q_n V_i(\ell_n).$$

Moreover for any functions $W : \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$ *and* $W_i : \triangle(\mathbf{X}) \to \mathbb{R}, i = 1, \ldots, I$, *such that* W *represents* $\succsim$*; for each* i, W_i *represents* $\succsim_i$*, then* $W\left((z_1, \ldots, z_I), \left[(\ell_n; q_n)_{n=1}^N\right]\right) = \sum_{i=1}^I \sum_{n=1}^N z_i q_n W_i(\ell_n)$ *holds if and only if there exists* $a > 0$ *and* $b \in \mathbb{R}$*, such that* $W_i = aV_i + b$.

If we restrict attention to the case where the birth lottery over identity is uniform and where the birth lottery over life chances is degenerate on ℓ, then this social welfare function reduces to the form $\frac{1}{I}\sum_{i=1}^I V_i$.

Clearly, the representations of section 3 are special cases of the above representation. But according to Theorem 7, $V_i(.)$ may be any function on $\triangle(\mathbf{X})$, and so in particular each $\succsim_i$ need not conform to expected utility theory. In this sense Theorem 7 is a generalization of both Karni & Safra (2000) and Safra & Weissengrin (2001).

If we now impose preference for post-birth randomization, then the additional restriction forces each V_i function to be concave.

Proposition 8 (Concavity) *The impartial observer's preferences satisfy the conditions of Theorem 7 part (a) and satisfy preference for post- (pre-) birth randomization if and only if the* V_i*-functions defined in Theorem 7 part (b) are concave (convex).*

If we impose neutrality between pre- and post-birth randomization then we get back the conditions of Proposition 1 and hence obtain a utilitarian social welfare function.

7 Related Literature

Other papers that model moral value judgements from the viewpoint of a sympathetic impartial observer such as Eichberger and Pethig (1994), Karni and Weymark (1998), Karni and Safra (2000, 2002) and Safra and Weissengrin (2002) all require the impartial observer's preference to satisfy reduction. With the exception of Karni and Safra (2000, 2002) all these papers also require the preferences of individuals to conform to expected utility and so cannot accommodate the fairness considerations exemplified by Diamond's critique. Karni and Safra's (2000, 2002) model, on the other hand, can accommodate Diamond's critique by allowing individuals themselves to possess

an inherent sense of fairness which is manifested in choice behavior that violates expected utility in a systematic manner.

In the related literature on Harsanyi's aggregation theorem, in which ethical considerations are superimposed on personal preferences through a social preference relation defined on the same domain as those of the preferences of the individuals, of particular interest is the contribution by Epstein and Segal (1992). Independence is weakened in a way that favors randomizations over any pair of social states for which two individuals' interests are opposed and leads to a quadratic social welfare function defined over the individuals' (expected) utilities. As they observe, if their framework is extended to a dynamic setting where uncertainty resolves in multiple stages through time, by imposing reduction and "dynamic consistency" the conditional preference derived from a quadratic social welfare function given the resolution of some previous uncertainty, may now entail a strict preference for the allocation of an indivisible good to one individual over another, and exhibit a *strict* aversion to further randomization. In our setting, given an imputed randomization pre-birth, this might then lead to a strict aversion to post-birth randomization.

The closest in spirit to our work is Karni (1996). In his treatment, social states are "allocation procedures" which are modelled as randomizations or lotteries over final allocations. So the set of (simple) lotteries over social-states are therefore *two-stage* lotteries, i.e. elements of $\triangle_0(\triangle(\mathbf{X}))$ in our notation. The Harsanyi aggregation theorem in this setting results in the following representation for the social preference relation: for each i, there exists a function $V_i : \triangle(\mathbf{X}) \to \mathbb{R}$, and a positive weight $a_i > 0$, such that $\sum_{n=1}^{N} q_n V_i(P_n)$ represents individual i's preferences defined over $\triangle_0(\triangle(\mathbf{X}))$ and $\sum_{i=1}^{I} a_i \sum_{n=1}^{N} q_n V_i(P_n)$ represents the social preference relation. The justification he provides for this setup actually seems more appropriate for our framework, in which it is only the impartial observer who has preferences defined over a two-stage lottery space. Indeed the section in which he articulates his justification is entitled "The veil of ignorance and the veil of amnesia" and contains the following paragraph (where we have changed the notation so as to conform with what we have used in this paper):

> "To grasp this claim consider the following thought experiment. Imagine that the members of a society convene in a presocietal stage to decide on allocation procedures to be used by a society of which they will be members. Suppose that, at this presocietal convention, the participants are aware that after they have decided on the allocation

procedure to be employed, they enter the societal state. Moreover, *once in the societal state they will have no recollection of the lotteries that were used to select the procedure and will know only which procedure is actually employed.* Consequently, in the societal state, they will regard δ^{x_1} as an arbitrary procedure favoring one individual over the other while they will consider the societal-state $[\frac{1}{2}\delta^{x_1} + \frac{1}{2}\delta^{x_2}]$ to be a fair procedure. With this foresight, the individuals *in the presocietal stage* strictly prefer $[\frac{1}{2}\delta^{x_1} + \frac{1}{2}\delta^{x_2}; 1]$ over $[(\delta^{x_1}; \frac{1}{2}), (\delta^{x_2}; \frac{1}{2})]$." (pp491-2, emphasis in the original).

The essential difference between our approach and Karni's (as well as that of Karni and Safra, 2002) is that he requires the notion of procedural fairness to be embodied in the individuals' preference relations $\{\succsim_i\}_{i\in\mathcal{I}}$ which becomes reflected in the societal preference relation by the aggregation procedure. We, on the other hand, have not imposed any ethical rules on individual preferences as they are defined on $\triangle(\mathbf{X})$.[6] In our setting, as is the case for Harsanyi, concern for procedural fairness is viewed as an ethical or moral concern that is encoded in the preferences of the hypothetical impartial observer.[7] Indeed, as a general principle, we would argue that it is easier for an individual to contemplate procedural fairness from the viewpoint of an impartial observer. That is, procedural fairness concerns arise when an individual undertakes the thought experiment of putting himself outside his own identity and trying to act morally or ethically.

[6] Of course, once we add in welfare and assume congruence, then we are explicitly assuming selfishness.

[7] In this respect, our approach is similar to Segal (2000) who imputes a separate preference relation for an individual when he or she is considering choice among different social policies which stands apart from his or her own personal preferences over the social alternatives.

Appendix

Although the results in the paper appear in order of increasing generality, it is convenient to prove the more general results first and show how the others follow as special cases.

Proof of Theorem 7. It is straightforward to check that (b) implies (a). We show that (a) implies (b) in the following. We use Safra and Weissengrin's (2001) representation theorem: it shows that if the domain of $\succsim$ were $\triangle(\mathcal{I}) \times \triangle_0(D)$ for a compact metrizable set D, the Acceptance Principle and the impartial observer's pre-birth independence implies that there is $V : \triangle(\mathcal{I}) \times \triangle_0(D) \to \mathbb{R}$ and continuous and non-constant functions $W_i : \triangle(D) \to \mathbb{R}$, $i = 1, \ldots, I$, such that V represents $\succsim$; for each i, W_i represents $\succsim_i$; and for all $(z, p) \in \triangle(\mathcal{I}) \times \triangle_0(D)$, $V(z, p) = \sum_{i=1}^{I} z_i W_i(p)$ and W_i is unique up to positive affine transformation as in the statement. In our setup, since $\triangle(\mathbf{X})$ is a compact metrizable space, Safra - Weissengrin's theorem implies that we have continuous $W_i : \triangle_0(\triangle(\mathbf{X})) \to \mathbb{R}$ $i = 1, \ldots, I$, such that V represents $\succsim$; for each i, W_i represents $\succsim_i$; and for all $(z, P) \in \triangle(\mathcal{I}) \times \triangle_0(\triangle(\mathbf{X}))$, $V(z, P) = \sum_{i=1}^{I} z_i W_i(P)$ and the functions W_i are jointly determined unique up to positive affine transformation, that is, $V'(z, P) = \sum_{i=1}^{I} z_i W_i'(P)$ and V' represents $\succsim$ and each W_i' represents $\succsim_i$, then there are $a > 0$ and b such that $W_i = aW_i' + b$.

Now pre-birth independence further implies that the relation induced by W_i has an expected utility representation where $\triangle_0(\triangle(\mathbf{X}))$ is viewed as the set of lotteries over outcomes ℓ in $\triangle(\mathbf{X})$. So W_i must be a continuous, increasing transformation of an expected utility function: that is, there exist an increasing continuous function ϕ_i and a continuous function V_i defined on $\triangle(\mathbf{X})$ such that $W_i(P) := W_i(\left[(\ell_n; q_n)_{n=1}^{N}\right]) = \phi_i(\sum_{n=1}^{N} q_n V_i(\ell_n))$. But since W_i in the representation is unique up to positive affine transformation, each ϕ_i must itself be a positive affine function. Hence relabeling as necessary we conclude that there is a continuous function V_i such that $W_i(\left[(\ell_n; q_n)_{n=1}^{N}\right]) = \sum_{n=1}^{N} q_n V_i(\ell_n))$ holds for any $\left[(\ell_n; q_n)_{n=1}^{N}\right] \in \triangle_0(\triangle(\mathbf{X}))$. By the Acceptance Principle, each V_i must represent $\succsim_i$. Then such a representation is jointly unique up to positive affine transformation as stated in condition 2, since functions W_i are unique up to positive affine transformation. This completes the proof. ■

Proof of Proposition 8 If the impartial observer's preferences satisfy the conditions for Theorem 7, then $(z, P') \succsim (z, P)$ if and only if $\sum_{i=1}^{I} z_i V_i(\ell_m) \geq \sum_{i=1}^{I} z_i(\beta V_i(\ell') + (1-\beta) V_i(\ell''))$. Let z be the lottery which yields identity i with probability one. By the acceptance principle (Axiom 1), we have $V_i(\ell) \geq \beta V_i(\ell') + (1-\beta) V_i(\ell'')$ for any lotteries ℓ, ℓ', and $\ell'' \in \triangle(\mathbf{X})$ with $\ell = \beta\ell' + (1-\beta)\ell''$. So each function V_i is concave on $\triangle(\mathbf{X})$. Conversely, if each V_i is a concave function, the required inequality above holds by definition for any z. If P is obtained by a finite series of elementary bifurcation of P', then repeating the inequality above finitely many times we have $(z, P') \succsim (z, P)$ for any z as desired. ■

Proof of Proposition 2 Clearly, (b) implies (a). To show that (a) implies (b), assume (a). Then by Theorem 7, we have a generalized utilitarian representation of the form $\sum_{i=1}^{I} \sum_{n=1}^{N} z_i q_n V_i(\ell_n)$. Since by the acceptance principle each V_i represents $\succsim_i$ which by post-birth independence admits a representation of the form $\int u_i(x)\, \ell(dx)$ whose range is $[0,1]$, there exist increasing, continuous functions $\phi_i : [0,1] \to \mathbb{R}$, $i = 1, \ldots, I$, such that $V_i(\ell) = \phi_i(\int u_i(x)\, \ell(dx))$ for any $\ell \in \triangle(\mathbf{X})$. This gives (b). ■

Proof of Proposition 3 By Proposition 8, $\succsim$ exhibits a preference for post-birth resolution of uncertainty if and only if V_i is concave for all i in $\mathcal{I}$. But $V_i(\ell) = \phi_i(\int u_i(x)\, \ell(dx))$ is concave if and only if ϕ_i is concave, since $\ell \mapsto \int u_i(x)\, \ell(dx)$ is linear. ■

Proof of Proposition 1 Again (b) implies (a) is immediate. To see that (a) implies (b) consider the preference relation $\hat{\succsim}$ defined on $\triangle(\mathcal{I}) \times \triangle(\mathbf{X})$ induced by a preference relation over birth-life lotteries $\succsim$ as follows:

$$\left(z, \left[(\ell_n; q_n)_{n=1}^{N}\right]\right) \succsim \left(z', \left[(\ell'_{n'}; q'_{n'})_{n'=1}^{N'}\right]\right) \text{ implies } \left(z, \sum_{n=1}^{N} q_n \ell_n\right) \hat{\succsim} \left(z', \sum_{n'=1}^{N'} q'_{n'} \ell'_{n'}\right)$$

If $\succsim$ satisfies reduction then $\hat{\succsim}$ inherits the ordering and continuity properties of $\succsim$. Furthermore, if $\succsim$ satisfies the acceptance principle, and pre-birth independence, then $\hat{\succsim}$ satisfies the properties necessary and sufficient for Safra & Weissengrin's (2002) representation theorem and thus $\succsim$ admits the above representation. ■

Proof of Corollary 4. By axiom 5 (preference for Post-Birth Randomization) $\left(\delta^i, \delta^{\alpha\ell+(1-\alpha)\ell'}\right) \succsim \left(\delta^i, \alpha\delta^\ell + (1-\alpha)\,\delta^{\ell'}\right)$. The right side can be re-expressed as $\alpha\left(\delta^i, \delta^\ell\right) + (1-\alpha)\left(\delta^i, \delta^{\ell'}\right)$. Since $\left(\delta^i, \delta^{\ell'}\right) \sim \left(\delta^j, \delta^\ell\right)$, this is indifferent to $\alpha\left(\delta^i, \delta^\ell\right) + (1-\alpha)\left(\delta^j, \delta^\ell\right) = \left(\alpha\delta^i + (1-\alpha)\,\delta^j, \delta^\ell\right)$ ■

Proof of Proposition 5. Fix v,w in $\left[U_j\left(\widetilde{\ell}'\right), U_j\left(\widetilde{\ell}\right)\right]$ and without loss of generality assume $v > w$. Since $U_i(.)$ and $U_j(.)$ are affine, there exist unique β^j, γ^j, β^i and β^i each in $[0,1]$, for which

$$\begin{aligned}
\beta^j U_j\left(\widetilde{\ell}\right) + \left(1-\beta^j\right) U_j\left(\widetilde{\ell}'\right) &= v \\
\gamma^j U_j\left(\widetilde{\ell}\right) + \left(1-\gamma^j\right) U_j\left(\widetilde{\ell}'\right) &= w \\
\beta^i U_i(\ell) + \left(1-\beta^i\right) U_i(\ell') &= \phi_i^{-1} \circ \phi_j(v) \\
\gamma^i U_i(\ell) + \left(1-\gamma^i\right) U_i(\ell') &= \phi_i^{-1} \circ \phi_j(w)
\end{aligned}$$

Applying the representation $\sum_{i=1}^{I}\sum_{n=1}^{N} z_i q_n \phi_i \circ U_i(\ell_n)$ it follows that by construction we have

$$\left(\delta^i, \delta^{\beta^i\ell+\left(1-\beta^i\right)\ell'}\right) \sim \left(\delta^j, \delta^{\beta^j\widetilde{\ell}+\left(1-\beta^j\right)\widetilde{\ell}'}\right) \succ \left(\delta^i, \delta^{\gamma^i\ell+\left(1-\gamma^i\right)\ell'}\right) \sim \left(\delta^j, \delta^{\gamma^j\widetilde{\ell}+\left(1-\gamma^j\right)\widetilde{\ell}'}\right).$$

Hence by hypothesis

$$\left(\delta^i, \delta^{\alpha\left(\beta^i\ell+\left(1-\beta^i\right)\ell'\right)+(1-\alpha)\left(\gamma^i\ell+\left(1-\gamma^i\right)\ell'\right)}\right) \succsim \left(\delta^j, \delta^{\alpha\left(\beta^j\widetilde{\ell}+\left(1-\beta^j\right)\widetilde{\ell}'\right)+(1-\alpha)\left(\gamma^j\widetilde{\ell}+\left(1-\gamma^j\right)\widetilde{\ell}'\right)}\right),$$

for all α in $[0,1]$. From the representation, this holds if and only if

$$\begin{aligned}
& U_i\left(\alpha\left(\beta^i\ell + \left(1-\beta^i\right)\ell'\right) + (1-\alpha)\left(\gamma^i\ell + \left(1-\gamma^i\right)\ell'\right)\right) \\
\geq\ & \phi_i^{-1} \circ \phi_j\left(U_j\left(\alpha\left(\beta^j\widetilde{\ell} + \left(1-\beta^j\right)\widetilde{\ell}'\right) + (1-\alpha)\left(\gamma^j\widetilde{\ell} + \left(1-\gamma^j\right)\widetilde{\ell}'\right)\right)\right).
\end{aligned}$$

But since $U_i(.)$ and $U_j(.)$ are affine, this implies

$$\begin{aligned}
& \alpha\left(\beta^i U_i(\ell) + \left(1-\beta^i\right) U_i(\ell')\right) + (1-\alpha)\left(\gamma^i U_i(\ell) + \left(1-\gamma^i\right) U_i(\ell')\right) \\
\geq\ & \phi_i^{-1} \circ \phi_j\left(\alpha\left(\beta^j U_j\left(\widetilde{\ell}\right) + \left(1-\beta^j\right) U_j\left(\widetilde{\ell}'\right)\right) + (1-\alpha)\left(\gamma^j U_j\left(\widetilde{\ell}\right) + \left(1-\gamma^j\right) U_j\left(\widetilde{\ell}'\right)\right)\right).
\end{aligned}$$

Substituting $\phi_i^{-1} \circ \phi_j(v)$ for $\beta^i U_i(\ell) + \left(1-\beta^i\right) U_i(\ell')$, $\phi_i^{-1} \circ \phi_j(w)$ for $\gamma^i U_i(\ell) + \left(1-\gamma^i\right) U_i(\ell')$, v for $\beta^j U_j\left(\widetilde{\ell}\right) + \left(1-\beta^j\right) U_j\left(\widetilde{\ell}'\right)$ and w for $\gamma^j U_j\left(\widetilde{\ell}\right) + \left(1-\gamma^j\right) U_j\left(\widetilde{\ell}'\right)$, yields

$$\alpha\phi_i^{-1} \circ \phi_j(v) + (1-\alpha)\,\phi_i^{-1} \circ \phi_j(w) \geq \phi_i^{-1} \circ \phi_j(\alpha v + (1-\alpha)\,w),$$

for all α in $[0,1]$. Since v and w were arbitrarily chosen from the interval $\left[U_j\left(\widetilde{\ell}'\right), U_j\left(\widetilde{\ell}\right)\right]$, the last inequality corresponds to the convexity of $\phi_i^{-1} \circ \phi_j$ on this interval. ■

References

Atkinson, Anthony B. (1970): "On the Measurement of Inequality," *Journal of Economic Theory*, 2, 244-63.

Deschamps, Robert and Louis Gevers (1979): "Separability, Risk-Bearing and Social Welfare Judgements", in Jean-Jacques Laffont (ed.) *Aggregation and Revelation of Preferences*, North Holland.

Diamond, Peter A. (1967): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment." *Journal of Political Economy* 75, 765-66.

Eichberger, Jurgen and Rudiger Pethig (1994): "Constitutional choice of rules" *European Journal of Political Economy* 10, 311-37.

Epstein, Larry G. and Uzi Segal (1992): "Quadratic Social Welfare," *Journal of Political Economy* 100, 691-712.

Peter C. Fishburn (1988): *Nonlinear Preference and Utility Theory.* Baltimore: Johns Hopkins University Press.

Grant, Simon, Atsushi Kajii and Ben Polak (1998): "Intrinsic Preference for Information", *Journal of Economic Theory*, 83, 233-259.

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

Harsanyi, John C. (1955): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment." *Journal of Political Economy* 63, 309-21.

Harsanyi, John C. (1975): "Nonlinear Social Welfare Functions: Do Welfare Economists Have a Special Exemption from Bayesian Rationality?" *Theory and Decision* 6, 311-32.

Harsanyi, John C. (1977): "Nonlinear Social Welfare Functions: A Rejoinder to Professor Sen." In *Foundational Problems in the Social Sciences*, edited by Robert E. Butts and Jaakko Hintikka. Dordrecht: Reidel.

Karni, Edi (1996): "Social welfare functions and fairness." *Social Choice and Welfare* 13, 487-96.

Karni, Edi and Zvi Safra (2000): "An extension of a theorem of von Neumann and Morgenstern with an application to social choice." *Journal of Mathematical Economics* 34, 315-27.

Karni, Edi and Zvi Safra (2002): "Individual sense of justice: a utility representation" *Econometrica* 70(1), 263-284.

Karni, Edi and John Weymark (1998): "An informationally parsimonious impartial observer theorem", *Social Choice and Welfare* 15(3), 321-32.

Myerson, Roger (1981): "Utilitarianism, egalitarianism and the timing effect in social choice problems," *Econometrica* 49, 883-97.

Pattanaik, Prasanta K. (1968): "Risk, Impersonality, and the Social Welfare Function." *Journal of Political Economy* 76, 1152-69.

Pearce, David, G. (1995): "Arrow's Theorem on its Head: A Bayesian Perspective on Social Choice," Yale University Discussion Paper.

Safra, Zvi and Einat Weissengrin (2002): "Harsanyi's impartial observer theorem with a restricted domain." Forthcoming *Social Choice and Welfare*.

Segal, Uzi (1990): "Two-Stage Lotteries without the Reduction Axiom", *Econometrica* 58, 349-377.

Segal, Uzi (2000): "Let's agree that all dictatorships are equally bad", *Journal of Political Economy* 108, 569-89

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

Sen, Amartya, K. (1977): "On weights and measures: informational constraints in social welfare analysis." *Econometrica* 45, 1539-72.

Sen, Amartya, K. (1992): *Inequality Reexamined*, Harvard University Press.

Weymark, John (1991): "A reconsideration of the Harsanyi-Sen debate on utilitarianism" in *Interpersonal Comparisons of Well-being* edited by Jon Elster and John Roemer. Cambridge: CUP.