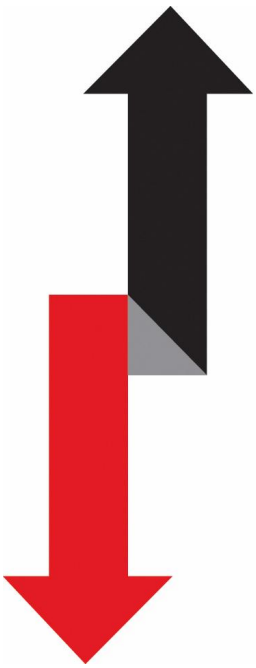

COLEMAN FUNG
RISK MANAGEMENT
RESEARCH CENTER

Working Paper # 2009 -06

Piercing the Veil of Ignorance

Shachar Kariv, University of California Berkeley
William R. Zame, University of California Los Angeles

September 15, 2009

University of California Berkeley

Piercing the Veil of Ignorance*

Shachar Kariv† William R. Zame ‡
UC Berkeley UCLA

September 15, 2008

Abstract

Theories of justice in the spirit of Harsanyi and Rawls argue that fair-minded people should aspire to make choices for society – that is, for themselves and for others – as if in the *original position*, behind a *veil of ignorance* that prevents them from knowing their own social and economic positions in society. While the original position is a purely hypothetical situation, developed as a thought experiment, the main result of this paper is that (under certain assumptions) preferences – hence choices – *behind* the veil of ignorance are determined by preferences *in front of* the veil of ignorance. This linkage between preferences behind and in front of the veil of ignorance has implications for distributive theories of justice and for theories of choice.

*We are grateful to Raymond Fisman, Douglas Gale and Daniel Markovitz for helpful conversations, and to seminar audiences at Bocconi, Brown, Caltech, Columbia, EUI, NYU, Oxford, Rice, UC Berkeley, UCL and UT Austin for many useful and encouraging comments. Financial support was provided by the National Science Foundation under grants SES-0317752, SES-0518936 and SES 0617027, by the UC Berkeley and UCLA Academic Senate Committees on Research, and by the Coleman Fung Risk Management Research Center (OpenLink Fund) at UC Berkeley. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agency.

†Department of Economics, University of California, Berkeley, 508-1 Evans Hall # 3880, Berkeley, CA 94720 (E-mail: kariv@berkeley.edu, URL: www.econ.berkeley.edu/~kariv/).

‡Department of Economics, University of California, Los Angeles, Bunche Hall, 405 Hilgard Avenue, Los Angeles, CA 90095 (E-mail: zame@econ.ucla.edu, URL: www.econ.ucla.edu/zame/).

JEL Classification Numbers: D63.

Key Words: Moral preferences, social preferences, distributional preferences, social choice, the original position, veil of ignorance.

1 Introduction

In a classic series of writings, John Harsanyi (1953, 1955) and John Rawls (1971) construct theories of social justice based on the choices that people would or should make for society (that is, for themselves and for others) in what Rawls terms the *original position*, behind a *veil of ignorance* that prevents people from "knowing their own social and economic positions, their own special interests in the society, or even their own personal talents and abilities (or their lack of them)." (Harsanyi, 1975; p. 594) The work of Harsanyi and Rawls, and of the many others who have followed them, has had broad-reaching influence across many disciplines, including philosophy, economics, and law.

Harsanyi and Rawls view preferences in the original position – which might be called *moral preferences* – as having a different nature from ordinary preferences for consumption or for risk or even for the consumption of others – which might be called *consumption preferences*, *risk preferences* and *social preferences*. Both Harsanyi and Rawls insist that moral preferences must conform to certain rationality requirements, and hence must have a special form – as opposed to consumption preferences and risk preferences and social preferences, which merely reflect taste and so could be quite arbitrary.[1]

This paper argues that moral preferences *cannot* occupy such a privileged position, because (modulo certain assumptions) *moral preferences are completely determined by risk preferences and social preferences*. If risk preferences and social preferences can be completely arbitrary reflections of taste, then moral preferences also can be completely arbitrary reflections of taste; if moral preferences must have a special form, then risk preferences and social preferences must have special forms as well.

Harsanyi and Rawls – and many other writers – view the original position

[1] As Harsanyi (1975,1978) points out, he and Rawls come to quite different conclusions about the form moral preferences should take, in large part because they view uncertainty very differently.

as a purely hypothetical environment, and hence view moral preferences as a purely intellectual construct. However, as this paper shows, preferences in the hypothetical environment of the original position are determined by preferences in environments that are not at all hypothetical. Put differently, choice behavior *behind* the veil of ignorance is determined by choice behavior *in front of* the veil of ignorance.

This linkage between preferences behind and in front of the veil of ignorance – between moral preferences and risk and social preferences – provides a new way of interpreting the theory of justice not just as a normative theory (how people ought to choose), but also as a descriptive theory (how people actually choose) and even as a prescriptive theory (as a practical aid to choice). This is important for the practical implications of broader theories of moral preferences, which are important in many contexts, including redistribution, taxation, development and globalization, among others. In all of these cases, understanding behavior requires understanding the preferences that lie behind it.

The remainder of the paper is organized as follows. Section 2 introduces the choice environments and the basic assumptions on preferences, Section 3 provides the decomposition and some simple corollaries, Section 4 discusses alternative interpretations and Section 5 provides examples that explicate the role of the assumptions. Section 6 concludes by relating this work to some of the literature and suggesting directions for future research.

2 Choice Environments

Society consists of N agents, of whom the Decision Maker is agent 1. (We abuse notation and write N for both the set of agents and its cardinality.) We are interested in preferences (equivalently, choice behavior) of the Decision Maker in three environments. In the first, which we term the Risk environment, the objects of choice are random (risky) allocations for the Decision Maker. In the second, which we term the Social Choice environment, the objects of choice are deterministic allocations for all the members of society, including the Decision Maker. In the third, which we term the Moral Choice environment, the objects of choice are again allocations for all the members of society, including the Decision Maker, but in a setting in

which the Decision Maker does not know his/her position in the society (nor indeed the positions of others). For the moment, we interpret the outcomes of decisions as *income*; different interpretations are discussed in Section 4.

We formalize choice spaces for these environments as follows:

- The choice space $\mathcal{R}$ in the RISK environment consists of all *lotteries* with non-negative payoffs; that is, formal expressions $\sum p_j x_j$ where (p_j) is a probability vector and each $x_j \in \mathbb{R}_+$. The lottery $\sum p_j x_j$ yields the Decision Maker income x_i with probability p_j.[2]
- The choice space $\mathcal{S}$ in the SOCIAL CHOICE environment consists of all vectors $x \in \mathbb{R}_+^N$. The vector x yields agent $i \in N$ the income x_i, and in particular, yields x_1 to the Decision Maker.

To describe the choice space in the MORAL CHOICE environment, let $\mathrm{Perm}(N)$ be the group of permutations (bijections) $\sigma : N \to N$. A vector $x \in \mathbb{R}_+^N$ may be viewed as a function $x : N \to \mathbb{R}_+$ so if $x \in \mathbb{R}_+^N$ and $\sigma \in \mathrm{Perm}(N)$ then the composition $x\sigma$ is again an element of $\mathbb{R}_+^N$; x assigns income x_i to agent i so $x\sigma$ assigns income $x_{\sigma(i)}$ to agent i.

- The choice space $\mathcal{M}$ in the MORAL CHOICE environment consists of all lotteries
$$\sum_{\sigma \in \mathrm{Perm}(N)} p_\sigma (x\sigma)$$
where (p_σ) is a probability distribution on $\mathrm{Perm}(N)$ and $x \in \mathbb{R}_+^N$. This lottery yields agent $i \in N$ the income $x_{\sigma(i)}$ with probability p_σ, and in particular yields the Decision Maker income $x_{\sigma(1)}$ with probability p_σ.

In the RISK environment, the Decision Maker is to choose (a lottery yielding) random income for the Decision Maker alone. In the SOCIAL CHOICE environment, the Decision Maker is to choose deterministic income for every agent in society. In the MORAL CHOICE environment, the Decision Maker is to choose a *deterministic* distribution of income across society but with a

[2] We could enlarge the environment to include compound lotteries as well as simple lotteries; this would complicate the notation and some of the arguments without changing the analysis or conclusions.

random assignment of agents to places in society. The MORAL CHOICE environment (with equal probabilities) coincides *exactly* with Harsanyi's (1953, 1955) formalization of the original position.

In the RISK, SOCIAL CHOICE and MORAL CHOICE environments, we describe preferences of the Decision Maker in terms of preference/indifference relations $\succeq_r$, $\succeq_s$ and $\succeq_m$, respectively, with associated strict preference relations $\succ_r$, $\succ_s$ and $\succ_m$ and indifference relations $\sim_r, \sim_s, \sim_m$; we refer to these as *risk preferences*, *social preferences* and *moral preferences*, respectively. We assume throughout that the (weak) preference relations satisfy the usual requirements: completeness, transitivity, reflexivity, negative intransitivity, and continuity.

The RISK, SOCIAL CHOICE and MORAL CHOICE environments involve entirely different choice spaces, so to link preferences in these environments, we construct a single choice space in which the RISK, SOCIAL CHOICE and MORAL CHOICE choice spaces can all be imbedded. The simplest choice space suitable for this purpose simply enriches the SOCIAL CHOICE environment by adding lotteries. Formally, write $\mathcal{L}$ for the set of lotteries $\sum_j p_j x^j$ where (p_j) is a probability vector and each $x^j \in \mathbb{R}^N_+$. This lottery yields agent $i \in N$ the income x^j_i with probability p_j, and in particular yields the Decision Maker income x^j_1 with probability p_j. A lottery in $\mathcal{L}$ represents a *random* allocation of income to each member of society.[3] As usual, if (p_j) is the degenerate probability distribution with $p_{j_0} = 1$ and $p_j = 0$ for $j \neq j_0$, then we identify the lottery $\sum p_j \cdot x^j$ with the certain outcome x^{j_0}.

Imbedding the RISK, SOCIAL CHOICE and MORAL CHOICE environments in $\mathcal{L}$ is straightforward. We identify the lottery $\sum p_j \in \mathcal{R}$ with the lottery $\sum p_j(x_j, 0, \ldots, 0)$, so identify $\mathcal{R}$ as the subset of $\mathcal{L}$ consisting of lotteries that yield all agents other than the Decision Maker the income 0 with probability 1. We identify $\mathcal{S}$ with the subset of $\mathcal{L}$ consisting of degenerate lotteries. Finally, we identify $\mathcal{M}$ as the subset of $\mathcal{L}$ consisting of lotteries $\sum p_\sigma x^\sigma$ with the property that $x^\sigma = x^1\sigma$ for each $\sigma \in \mathrm{Perm}(N)$.

In what follows, we make two assumptions about social preferences:

A1 (Worst Outcome) $x \succeq_s 0$ for every $x \in \mathcal{S}$.

[3] Again, we could allow for compound lotteries without changing the analysis or conclusions.

A2 **(Self-regarding)** For each $x \in \mathcal{S}$ there is a $t \in \mathbb{R}_+$ such that

$$(t, 0, \ldots, 0) \succeq_s x$$

Assumptions **A1** and **A2** limit the extent to which the Decision Maker is (respectively) spiteful or altruistic toward others; they seem very natural requirements but they are not entirely innocuous. It is certainly possible to imagine individuals who are so spiteful that **A1** fails or so altruistic that **A2** fails; for such individuals, Examples 1 and 2 below show that moral preferences are *not* determined by risk and social preferences.

3 Decomposing Preferences

Before stating our main result, we need a definition. Say that a preference relation $\succeq$ on $\mathcal{L}$ satisfies *Weak Independence* if for every probability vector (p_j) and choice arrays $(x^j), (y^j)$ we have

$$x^j \succeq y^j \text{ for each } j \Rightarrow \sum p_j \cdot x^j \succeq \sum p_j \cdot y^j$$

Weak Independence is a version of the familiar Independence Axiom: it allows substitution only of *outcomes* and not of lotteries. (Indeed, since we have not considered compound lotteries, substitution of lotteries would make no sense.) Weak Independence does *not* imply expected utility.

Theorem *For all risk preferences $\succeq_r$ and social preferences $\succeq_s$ that satisfy assumptions **A1** and **A2**, there is a unique preference relation $\succeq$ on $\mathcal{L}$ that satisfies Weak Independence and has the property that its restriction to $\mathcal{R}$ coincides with $\succeq_r$ and its restriction to $\mathcal{S}$ coincides with $\succeq_s$. In particular, if preferences $\succeq$ on $\mathcal{L}$ satisfy Weak Independence, then moral preferences $\succeq_m$ are determined by risk preferences $\succeq_r$ and social preferences $\succeq_s$.*

Proof. We provide a decomposition in terms of risk preferences over selfish equivalents.

Given $x \in \mathcal{S}$, say that $s \in \mathbb{R}_+$ is a *selfish equivalent* of x if $x \sim_s (s, 0, \ldots, 0)$. Our first task is to show that selfish equivalents exist.[4] To

[4]Because we have not assumed that social preferences are monotone in own consumption, selfish equivalents need not be unique, but that will not matter for our purposes.

this end, fix $x \in \mathcal{S}$. **A2** guarantees that there is some $t \in \mathbb{R}^N_+$ such that $(t, 0, \ldots, 0) \succeq_s x$. Set

$$S(x) = \inf\{t : (t, 0, \ldots, 0) \succeq_s x\}$$

If $S(x) > 0$ then $x \succ_s (S(x) - \varepsilon, 0, \ldots, 0)$ for every $\varepsilon > 0$, so continuity guarantees that $(S(x), 0, \ldots, 0) \sim_s x$. If $S(x) = 0$ then **A1** guarantees that $x \succeq_s (0, 0, \ldots, 0)$. In either case, we conclude that $(S(x), 0, \ldots, 0) \sim_s x$, so that $S(x)$ is a selfish equivalent of x.

Now consider lotteries $\sum p_j \cdot x^j, \sum q^k \cdot y^k \in \mathcal{L}$. By construction,

$$x^j \sim_s (S(x^\sigma), 0, \ldots, 0) \text{ and } y^j \sim_s (S(y^\sigma), 0, \ldots, 0)$$

for each $\sigma \in \text{Perm}(N)$. Define a preference relation $\succeq$ on $\mathcal{L}$ by

$$\begin{array}{ccc} \sum p^j x^j & \succeq & \sum q^k y^k \\ & \Updownarrow & \\ \sum p^j (S(x^j), 0, \ldots, 0) & \succeq_r & \sum q^k (S(y^k), 0, \ldots, 0) \end{array}$$

Because risk preferences and social preferences are complete, reflexive, transitive, negatively intransitive and continuous, the preference relation $\succeq$ has the same properties. It is evident that $\succeq$ satisfies weak independence and that the restriction of $\succeq$ to $\mathcal{R}$ coincides with $\succeq_r$ and that the restriction of $\succeq$ to $\mathcal{S}$ coincides with $\succeq_s$. This is the desired decomposition.

Finally, Weak Independence implies that the extension $\succeq$ is unique. ■

Two simple corollaries are worth noting. The first corollary is motivated by the observation that our formulation allows for the possibility that the Decision Maker cares not only about the distribution of income to others, but also about which of the other agents receives which income. This would seem entirely natural, for instance, if the Decision Maker cares about the preferences of other agents, some (but not all) of whom care not only about their own income but also about their position in the income distribution. However, it follows immediately from the above proof that if the Decision Maker cares only about the distribution of income to others in the SOCIAL CHOICE environment then she cares only about the distribution of income to others in the MORAL CHOICE environment as well – indeed, in all of $\mathcal{L}$.

To formalize this result, write $\text{Perm}_1(N)$ for the subgroup of permutations in $\text{Perm}(N)$ that fix 1; that is, $\text{Perm}_1(N) = \{\sigma \in \text{Perm}(N) : \sigma(1) = 1\}$. Note that if $x \in \mathcal{S}$ and $\sigma \in \text{Perm}_1(N)$ then x and $x\sigma$ yield the same distribution of income to all other agents.

Corollary 1 *If $\succeq_s$ has the property that $x \sim_s x\sigma$ for all $x \in \mathcal{S}$ and all $\sigma \in \text{Perm}_1(N)$, then $\succeq$ has the property that*

$$\sum p_\sigma x^\sigma \sim \sum p_\sigma (x^\sigma \tau^\sigma)$$

for all lotteries $p_\sigma \cdot x^\sigma \in \mathcal{L}$ and all arrays $\tau^\sigma \in \text{Perm}_1(N)$.

The second corollary records the simple fact that if the Decision Maker is perfectly *selfish* in the Social Choice environment, then preferences in the Risk environment coincide with preferences in the Moral Choice environment.

Corollary 2 *If social preferences are selfish, in the sense that*

$$x \sim_s (x_1, 0, \ldots, 0)$$

for all $x \in \mathcal{S}$, then $\succeq$ has the property that

$$\begin{array}{rcl} \sum p^j \cdot x^j & \succeq & \sum q^k \cdot y^k \\ & \Updownarrow & \\ \sum p^j \cdot (x_1^j, 0, \ldots, 0) & \succeq_r & \sum q^k (y_1^k, 0, \ldots, 0) \end{array}$$

for all lotteries $\sum p^j \cdot x^j, \sum q^k \cdot y^k \in \mathcal{L}$.

4 Alternative Interpretations

We have interpreted elements $x \in \mathcal{S}$ as vectors of income, but other interpretations are possible. For instance, we might interpret x as the vector of utility obtained from a particular underlying physical choice, so that x_i is the utility obtained by agent i. In that interpretation, the Decision Maker's

preference ordering of over $\mathcal{S}$ should by definition coincide with the ordering of first components; i.e., $x \succeq_{\mathcal{S}} y$ exactly when $x_1 \geq y_1$. With this interpretation, social choices are selfish almost by definition – because utility already encompasses attitudes toward others – so that Corollary 2 applies.

A more subtle – but perhaps more problematic – approach is to interpret elements $x \in \mathcal{S}$ as vectors of *personal welfares* associated with underlying physical choices, so that x_i is the personal welfare obtained by agent i, but where personal welfare does *not* encompass attitudes toward others. In this interpretation, the preference ordering over $\mathcal{S}$ expresses the Decision Maker's attitude toward the relative importance of her own personal welfare and that of others.

5 Examples

The following examples show that our main result fails if assumptions **AA** or **A2** do not obtain; i.e., the Decision Maker is too spiteful or too altruistic.

Example 1 (Too spiteful) For simplicity, take $N = 2$. Let $f : \mathbb{R} \to \mathbb{R}$ be any continuous, strictly increasing function with the property that $f(t) = t$ for $t \geq 0$. Define a utility function $U_f : \mathcal{L} \to \mathbb{R}$ by

$$U_f(p_1(x_1, y_1) + p_2(x_2, y_2) = p_1 f(x_1 - y_1) + p_2 f(x_2 - y_2).$$

The risk preferences induced by U_f on $\mathcal{R}$ do not depend on f (because $f(t) = t$ whenever $t \geq 0$), and the social preferences induced by U_f on $\mathcal{S}$ do not depend on f (because f is strictly increasing). However, the moral preferences induced by U_f on $\mathcal{M}$ *do* depend on f (because the weight given to inequality depends on f). In particular, if $f_1(t) = t$ for all t but

$$f_2(t) = \begin{cases} t & \text{if} \quad t \geq 0 \\ 2t & \text{if} \quad t < 0 \end{cases}$$

then U_{f_1} and U_{f_2} do *not* induce the same moral preferences. Notice that the preferences induced by any such U_f satisfy **A2** on $\mathcal{S}$ and satisfy weak independence on $\mathcal{L}$ – but that **A1** fails: there is no worst outcome.

Example 2 (Too altruistic) For simplicity, take $N = 2$. Let $g : \mathbb{R} \to \mathbb{R}$ be any continuous, strictly increasing function with the property that $g(t) = t$

for $t \leq 0$. Define a utility function $W_g : \mathcal{L} \to \mathbb{R}$ by

$$W_g(p_1(x_1, y_1) + p_2(x_2, y_2) = p_1 g(-\exp(x_1) + y_1) + p_2 g(-\exp(x_2) + y_2)$$

The Risk preferences induced by W_g on $\mathcal{R}$ do not depend on g (because $g(t) = t$ whenever $t \leq 0$) and the social preferences induced by W_g on $\mathcal{R}$ do not depend on g (because g is strictly increasing). However the moral preferences induced by W_g on $\mathcal{M}$ *do* depend on g (because the weight given to inequality depends on g). In particular, if $g_1(t) = t$ for all t but

$$g_2(t) = \begin{cases} \frac{t}{2} & \text{if} \quad t \geq 0 \\ t & \text{if} \quad t \leq 0 \end{cases}$$

then W_{g_1} and W_{g_2} do *not* induce the same moral preferences. Notice that the preferences induced by any such W_g on $\mathcal{L}$ satisfy **A1** on $\mathcal{S}$ and satisfy weak independence on $\mathcal{L}$ – but do not satisfy **A2**: they are not self-regarding.

6 Conclusion

This paper contributes to the vast body of research on moral preferences. Harsanyi (1953, 1955) uses von Neumann and Morgenstern (1947) Expected Utility Theory to develop an axiomatizations of utilitarian ethics. Harsanyi's equiprobability model for moral value judgments is a special case of our MORAL CHOICEs environment. This model is Harsanyi's version of the concept of the original position because "he (the Decision Maker) would clearly satisfy the impartiality and impersonality requirements to the fullest possible degree" (Harsanyi, 1978; p. 227). Rawls (1971) rejects the use of expected utility in moral value judgments; predictably, Harsanyi (1975, 1977a, 1977b, 1978, 1979) disagrees. There has also been considerable controversy over Harsanyi's work, as in the debate between Sen (1976, 1977, 1986) and Harsanyi (1975, 1977c). Weymark (1991) provides an excellent review of this debate and clarifies the significance of Harsanyi's contributions.

It is perhaps most important to emphasize that the existence of a link between risk and social preferences and moral preferences seems quite unexpected. On the face of it, risk preferences, social preferences and moral preferences seem conceptually quite distinct. Moreover, as we have already noted, Harsanyi and Rawls go so far as to argue that, while risk and social

preferences are merely matters of taste, moral preferences reflect a deeper underlying rationality. Our results suggest that such a position may not be tenable. Given the considerable heterogeneity of risk and social preferences within and across societies, there seems no conceptual reason to expect that moral preferences should be consistent with any particular notion of rationality – or theory of justice.

The linkage established here between risk and social preferences on the one hand and moral preferences on the other hand suggests a number of promising directions for future research. Most obviously, this linkage means that the techniques of economic analysis may be brought to bear on modeling and predicting behavior governed by these preferences, including testing for consistency within and across environments and identifying the underlying structure of preferences.

References

[1] Harsanyi, J. (1953) "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking." *Journal of Political Economy*, 61, pp. 434-435.

[2] Harsanyi, J. (1955) "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy*, 63, pp. 309-321.

[3] Harsanyi, J. (1975) "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory." *American Political Science Review*, 69, pp. 594-606.

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

[5] Harsanyi, J. (1977b) "Nonlinear Social Welfare Functions: A Rejoinder to Professor Sen." In *Foundational Problems in the Special Sciences*, R. Butts and J. Hintikka (eds.), pp. 293-296.

[6] Harsanyi, J. (1977c) "Morality and the Theory of Rational Behavior." *Social Research*, 44, pp. 623-656.

[7] Harsanyi, J. (1978) "Bayesian Decision Theory and Utilitarian Ethics," *American Economic Review Papers & Proceedings*, 68, pp. 223-228.

[8] Harsanyi, J. (1979) "Bayesian Decision Theory, Rule Utilitarianism, and Arrow's Impossibility Theorem," *Theory and Decision*, 11, pp. 289-317.

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

[10] Sen, A.K. (1976) "Welfare Inequalities and Rawlsian Axiomatics." *Theory and Decision*, 7, pp. 243-262.

[11] Sen, A.K. (1977) "Non-linear Social Welfare Functions: a Reply to Professor Harsanyi." In *Foundational Problems in the Special Sciences*, R. Butts and J. Hintikka (eds.), pp. 297–302.

[12] Sen, A.K. (1986) "Social Choice Theory." In the *Handbook of Mathematical Economics*, vol. III, K. Arrow and M. Intriligator (eds.), pp. 1073–1181.

[13] von Neumann, J. and O. Morgenstern (1947) *The Theory of Games and Economic Behavior, 2nd ed.* Princeton: Princeton University Press.

[14] Weymark, J. (1991) "A Reconsideration of the Harsanyi-Sen Debate on Utilitarianism." In *Interpersonal Comparisons of Well-Being*. J. Elster and J. Roemer (eds.), pp. 255–320.