A generalized representation theorem for Harsanyi's ('impartial') observer

Simon Grant, Atsushi Kajii, Ben Polak & Zvi Safra

Social Choice and Welfare

ISSN 0176-1714
Volume 39
Number 4

Soc Choice Welf (2012) 39:833-846
DOI 10.1007/s00355-011-0563-0

Springer

Soc Choice Welf (2012) 39:833–846
DOI 10.1007/s00355-011-0563-0

ORIGINAL PAPER

A generalized representation theorem for Harsanyi's ('impartial') observer

Simon Grant · Atsushi Kajii · Ben Polak · Zvi Safra

Received: 3 June 2010 / Accepted: 24 June 2011 / Published online: 16 July 2011

Abstract We provide an axiomatization of an additively separable social welfare function in the context of Harsanyi's impartial observer theorem. To do this, we reformulate Harsanyi's setting to make the lotteries over the identities the observer may assume independent of the social alternative.

1 Introduction

This article revisits Harsanyi (1953, 1955, 1977) utilitarian impartial observer theorem. Consider a society of individuals $\mathcal{I}$. The society has to choose among a set $\mathcal{A}$ of social alternatives, with a generic element denoted by a. Each individual i has preferences $\succsim_i$ over these social alternatives. These preferences are known and they differ. In Harsanyi's set up, the set of social alternatives are identified with the set

S. Grant (✉)
Department of Economics, Rice University, Houston, TX, USA
e-mail: simongrant@uq.edu.au

S. Grant
School of Economics, University of Queensland, Brisbane, QLD, Australia

A. Kajii
Institute of Economic Research, Kyoto University, Kyoto, Japan

B. Polak
Department of Economics, School of Management, Yale University, New Haven, CT, USA

Z. Safra
University of Exeter, Exeter, UK

Z. Safra
The College of Management, Tel Aviv University, Tel Aviv, Israel

of probability measures (that is, 'lotteries') over a set of final social outcomes, which we shall denote by $\mathcal{X}$, with generic element x. In that case i's preference relation $\succsim_i$ corresponds to her risk preferences over the set of lotteries on $\mathcal{X}$, denoted by $\triangle(\mathcal{X})$.

To help choose among social outcomes, Harsanyi proposes that each individual should imagine herself as an '(impartial) observer' who does not know which person she will be. From the perspective of the observer, not only does she face uncertainty about which social outcome in $\mathcal{X}$ will obtain, but she also faces uncertainty over which identity in $\mathcal{I}$ she will assume. Harsanyi takes the observer's preferences to be over the set of lotteries defined on the set of *extended outcomes*, $\mathcal{I} \times \mathcal{X}$, that we shall denote by $\triangle(\mathcal{I} \times \mathcal{X})$. In forming preferences over $\triangle(\mathcal{I} \times \mathcal{X})$, the observer is forced to make interpersonal comparisons; for example, she is forced to compare being person i in (final) social outcome x with being person j in (final) social outcome x'.

Harsanyi assumes the acceptance principle: that is, when the observer imagines herself being person i she adopts person i's preferences over the social alternatives which in his case are the outcome lotteries in $\triangle(\mathcal{X})$. He also assumes that all individuals are expected utility maximizers, and that they continue to be so in the role of an observer. Harsanyi argues that these Bayesian rationality axioms force the observer to be a (weighted) utilitarian in the following sense: over all extended lotteries $e \in \triangle(\mathcal{I} \times \mathcal{X})$, the observer's preferences admit a representation of the form

$$V(e) = \sum_i z_i^e U_i\left(\ell_i^e\right), \tag{1}$$

where for each i in $\mathcal{I}$, z_i^e is the probability assigned by the extended lottery e to assuming the identity i, ℓ_i^e is the distribution over $\mathcal{X}$ associated with e conditional on identity i being assumed, and $U_i(\ell) := \int_{\mathcal{X}} u_i(x)\ell(\mathrm{d}x)$ is person i's expected utility for the lottery ℓ over final social states.

By considering extended lotteries with a *uniform* marginal distribution over identities, that is, a marginal distribution over identities that assigns an equal-chance $1/I$ for the observer assuming each identity i in $\mathcal{I}$ and by assuming that the outcome lotteries are public (that is, all individuals face the same outcome lottery or equivalently e is a product lottery with $\ell_i^e = \ell_j^e$ for all i and j), Harsanyi argues that the function $V(\cdot)$ given in (1), can be used to obtain an 'impartial' ranking of the outcome lotteries (and hence, the associated social policies). That is, for any pair of outcome lotteries ℓ and ℓ' in $\triangle(\mathcal{X})$, the observer 'impartially' ranks ℓ above ℓ' if $\sum_i I^{-1} \times U_i(\ell) > \sum_i I^{-1} \times U_i\left(\ell'\right)$. This is essentially the method Weymark (1991) uses to obtain an impartial ranking over social policies in his formalization of Harsanyi's impartial observer theorem.[1]

As we pointed out in our companion paper (Grant et al. 2010, pp. 1957–1958), Harsanyi's approach requires that the observer engages in what he refers to as imaginative empathy:

[1] Roemer (1992) observes that such a notion of impartiality reduces any consideration of social justice to simply one of rational prudence on the part of the observer. He, among others, contends that this constitutes an inadequate basis for a theory of social justice.

> This must obviously involve [her] imagining [her]self to be placed in individual i's *objective position*, i.e., to be placed in the objective positions (e.g., income, wealth, consumption level, state of health, social position) that i would face in social situation x. But it must also involve assessing these objective conditions in terms of i's own *subjective attitudes* and *personal preferences*… rather than assessing them in terms of [her] own subjective attitudes and personal preferences. [Harsanyi 1977, p. 52: notation changed to ours but emphasis in the original][2]

The knowledge and acceptance of individual preferences embodied in imaginative empathy implies agreement across all potential observers for pairs of the form (i, ℓ) and (i, ℓ') since by acceptance this is determined solely by $\succsim_i$. As Broome (1993) and Mongin (2001) have pointed out (and as Harsanyi (1977, p. 57) himself concedes), however, it does not imply agreement in ranking pairs of the form (i, ℓ) and (j, ℓ') where $i \neq j$. For example, each observer can have her own rankings across others' subjective and objective positions.

In this article, we shall retain the assumption that an observer can engage in imaginative empathy. As the focus of our analysis is on the representation such an observer's preferences may admit, we do not address the issue of the lack of agreement across potential observers.[3] Rather our point of departure from Harsanyi is that we do not require the set of social alternatives, $\mathcal{A}$, to correspond to a set of lotteries over some set of final social states. Hence individual i's preferences $\succsim_i$ need not correspond to her 'risk' preferences. As a consequence, we shall only require an observer's preferences to be defined over the set of identity lottery/social alternative pairs. That is, we take the domain of the observer's preferences to be $\triangle(\mathcal{I}) \times \mathcal{A}$, with a generic element denoted by the pair (z, a), where z is an identity lottery in $\triangle(\mathcal{I})$ and a is a social alternative in $\mathcal{A}$.

First notice that this still entails that the observer can make interpersonal comparisons between pairs of the form (i, a) and (j, a') for any $i \neq j$ and any pair of alternatives a and a'. However, it does not require her preferences to extend to the set $\triangle(\mathcal{I} \times \mathcal{A})$ the set of lotteries defined on identity/social alternative pairs which is strictly larger than $\triangle(\mathcal{I}) \times \mathcal{A}$.

From a conceptual viewpoint, one advantage of working with preferences defined on a smaller domain is that it makes less strenuous demands on the 'empathetic imagination' of the observer. Furthermore, if we take the view that randomizations involving known or objective probabilities are themselves purely hypothetical constructs used to help calibrate the observer's comparisons of interpersonal utility, then our domain for the observer's preferences has the virtue that such randomizations only take place in the (already) hypothetical setting in which the observer is not yet aware of the identity

[2] Rawls also appeals to such imaginative empathy: A competent judge … must not consider his own de facto preferences as the necessarily valid measure of the actual worth of those interests which come before him, but…be both able and anxious to determine, by imaginative appreciation, what those interests mean to persons who share them, and to consider them accordingly. [Rawls 1951, p. 179 quoted in Pattanaik (1968, pp. 1157–1158)]. See also Sen's (1979) behavioral and introspective bases for interpersonal comparisons of welfare.

[3] So we shall not claim the special designation of impartial for the observer's preferences.

she will assume. Even if randomizations involving known or objective probabilities are well defined in the real world, so that for example, following Harsanyi, we take $\mathcal{A} = \triangle(\mathcal{X})$, we argue in Grant et al. (2010) that the adoption of $\triangle(\mathcal{I} \times \mathcal{X})$ as the domain for the observer's preferences entails conflating 'hypothetical' randomizations with 'real' randomizations which may well put us at odds with the acceptance principle.

As an illustration of this last point, consider a society comprising two individuals i and j, and a perfectly divisible private good that is to be allocated between them. We take the set of final social outcomes to be $\mathcal{X} = [0, 1]$, with the interpretation that $x \in [0, 1]$, is the social outcome in which i gets the fraction x of the good and j gets the remaining $1 - x$. There exists an objective randomizing device, so we take the set of alternatives $\mathcal{A} = \triangle([0, 1])$ with generic element denoted by ℓ. Suppose each individual cares only about the amount of the good she receives and the observer's own interpersonal assessments are such that $(i, x) \sim (j, 1 - x)$ for all x in $[0, 1]$. That is, she is indifferent between being person i with share s of the good and being person j with the same share s of the good. Let ℓ' denote the outcome lottery for which there is a half chance x equals 1 and a half chance x equals 0. That is, ℓ' is the lottery in which each person has a half chance of getting the whole good and a half chance of getting none. Now if the observer's preferences are defined on $\triangle(\{i, j\} \times [0, 1])$ and she is an expected utility maximizer then it readily follows from her interpersonal assessments $(i, 1) \sim (j, 0)$ and $(i, 0) \sim (j, 1)$ that she must be indifferent as to which identity she assumes when facing the lottery ℓ', that is, $\left(j, \ell'\right) \sim \left(i, \ell'\right)$. But suppose $\succsim_i$ exhibits risk aversion and can be represented by the expected utility functional $U_i(\ell) = \int_0^1 \sqrt{x}\,\ell(dx)$ while $\succsim_j$ is risk neutral, and so can be represented by the functional $U_j(\ell) = \int_0^1 (1 - x)\,\ell(dx)$. It follows that person i is indifferent between ℓ' and the (degenerate) lottery in which $x = 1/4$ for sure. Person j on the other hand is indifferent between ℓ' and the (degenerate) lottery in which $x = 1/2$ for sure. Again combining acceptance with the observer's interpersonal assessments, we now have that the observer must prefer facing the lottery ℓ' as person j. That is, $\left(j, \ell'\right) \succ \left(i, \ell'\right)$, contradicting the previous conclusion that the observer must be indifferent as to which identity she assumes when facing ℓ'.

It may seem tempting to attribute the problem to the assumption that the observer's preferences conform to (standard) expected utility. And indeed preferences defined on $\triangle(\mathcal{I} \times \mathcal{X})$ that exhibit so-called 'issue preference' can accommodate the pattern of preference exhibited in the example.[4] The advantage of restricting the domain as we do in this article is that it keeps these issues apart from the start. It is possible for the observer to treat them as equivalent but that becomes a property of the preferences. Our main result, however, is that one can conduct Harsanyi's analysis within this restricted domain. Indeed, we show (see Theorem 1 below) that by imposing acceptance and a form of independence that applies to mixtures of identity lotteries, this leads to the existence of an additively separable representation of the form:

4 See, for example, Ergin and Gul (2009).

$$V(z, a) = \sum_i z_i V_i(a),$$

where z_i is again the probability of assuming person i's identity and $V_i(a)$ is a representation of person i's preferences $\succsim_i$.

2 Set up and notation

Let society consist of a finite set of individuals $\mathcal{I} = \{1, \ldots, I\}$, $I \geq 2$, with generic elements i and j. The set of social alternatives is denoted by $\mathcal{A}$ with generic element a. The set $\mathcal{A}$ is assumed to have more than one element and to be a compact metrizable space.

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

In Harsanyi's story, an observer imagines herself behind a veil of ignorance, uncertain about which identity she will assume in the given society. Let $\triangle(\mathcal{I})$ denote the set of *identity lotteries* on $\mathcal{I}$. Let z denote the typical element of $\triangle(\mathcal{I})$, and let z_i denote the probability assigned by the identity lottery z to individual i. They represent the imaginary risks in the mind of the observer of being born as someone else.[5] With slight abuse of notation, we will let i or sometimes $[i]$ denote the degenerate identity lottery that assigns probability weight 1 to the observer's assuming the identity of individual i.

As discussed above, we assume that the identity lotteries faced by the observer are independent of the social alternative; that is, she faces a identity lottery/social alternative pair $(z, a) \in \triangle(\mathcal{I}) \times \mathcal{A}$.

The observer is endowed with a preference relation $\succsim$ defined over $\triangle(\mathcal{I}) \times \mathcal{A}$. We assume throughout that $\succsim$ is complete, transitive and continuous, and that its asymmetric part $\succ$ is non-empty, and so it admits a (non-trivial) continuous representation $V : \triangle(\mathcal{I}) \times \mathcal{A} \to \mathbb{R}$. That is, for any two pairs, (z, a) and (z', a'), $(z, a) \succsim (z', a')$ if and only if $V(z, a) \geq V(z', a')$.

3 Generalized utilitarianism

In this section, we adapt the axioms from Harsanyi's observer theorem to apply to the framework of identity lottery/social alternative pairs.

The first axiom is Harsanyi's acceptance principle. In degenerate identity lottery/social alternative pairs of the form (i, a) or (i, a'), the observer knows she will assume

[5] Recall our discussion in the introduction of the 'imaginative empathy' required by the observer to formulate her preferences. For a more extensive discussion on what it is an individual imagines and knows when she imagines herself in the role of the observer see Grant et al. (2010, Sect. 7).

identity i for sure. The acceptance principle requires that, in this case, the observer's preferences $\succsim$ must coincide with that individual's preferences $\succsim_i$ over social alternatives.

The acceptance principle. *For all i in $\mathcal{I}$ and all $a, a' \in \mathcal{A}$, $a \succsim_i a'$ if and only if $(i, a) \succsim (i, a')$.*

Next, in the spirit of Harsanyi, we assume that the observer's preferences also satisfy independence. Here, however, we need to be careful. First, the set of identity lottery/social alternative pairs $\triangle(\mathcal{I}) \times \mathcal{A}$ is not a convex set. Not all probability mixtures of identity lottery/social alternative pairs are well defined. Therefore, the axiom we adopt applies independence over pairs of lottery/social alternatives for which probability mixtures are well defined.[6]

Independence over identity lotteries (for the impartial observer). *Suppose (z, a), $(z', a') \in \triangle(\mathcal{I}) \times \mathcal{A}$ are such that $(z, a) \sim (z', a')$. Then, for all $\tilde{z}, \tilde{z}' \in \triangle(\mathcal{I})$: $(\tilde{z}, a) \succsim (\tilde{z}', a')$ if and only if $(\lambda\tilde{z} + (1-\lambda) z, a) \succsim (\lambda\tilde{z}' + (1-\lambda) z', a')$ for all λ in $(0, 1]$.*

To understand this axiom, first notice that the two mixtures on the right side of the implication are identical to $\lambda(\tilde{z}, a) + (1-\lambda)(z, a)$ and $\lambda(\tilde{z}', a') + (1-\lambda)(z', a')$, respectively. These two mixtures of identity lottery/social alternative pairs are well defined since they each involve the mixture of two identity lotteries holding the social alternative fixed. Second, notice that the identity lottery/social alternative pairs, (z, a) and (z', a'), that are 'mixed in' with weight $(1-\lambda)$ are themselves indifferent. The axiom states that 'mixing in' two indifferent identity lottery/social alternative pairs (with equal weight) preserves the original preference between $(\tilde{z}, a)$ and $(\tilde{z}', a')$ prior to mixing.

One technical remark that might interest some readers. In the axiom, we allow the mixing of identity lotteries to occur at two different social alternatives; that is, we do not restrict a to equal a'. We could define a weaker axiom—call it *conditional independence*—that simply imposes independence over identity lotteries at each *fixed* social alternative $\bar{a}$. That is, for all $\bar{a} \in \mathcal{A}$, if $z, z' \in \triangle(\mathcal{I})$ are such that $(z, \bar{a}) \sim (z', \bar{a})$ then for all $\tilde{z}, \tilde{z}' \in \triangle(\mathcal{I})$, $(\tilde{z}, \bar{a}) \succsim (\tilde{z}', \bar{a})$ if and only if $(\alpha\tilde{z} + (1-\alpha) z, \bar{a}) \succsim (\alpha\tilde{z}' + (1-\alpha) z', \bar{a})$ for all α in $(0, 1]$. Our stronger axiom is necessary for the representation result that follows. To show this, Example 1 in the Appendix presents an example in which the set of alternatives is a set of lotteries over some set of final outcomes (that is, $\mathcal{A} = \triangle(\mathcal{X})$), each individual is an expected utility maximizer and the observer's preferences satisfy both the acceptance principle and conditional independence over identity lotteries but does *not* satisfy the (unconditional) independence axiom over identity lotteries defined above.

As our main result states, in conjunction with acceptance, independence over identity lotteries is all we need to ensure the observer's preferences admit a generalized utilitarian representation.

[6] Our axiom is similar to Karni and Safra (2000) 'constrained independence' axiom, but their axiom applies to all joint distributions over identities and social alternatives, not just to identity-lottery/social alternative pairs.

Theorem 1 *The following are equivalent:*

(a) *The observer's preferences $\succsim$ satisfy the acceptance principle and independence over identity lotteries.*
(b) *There exist a continuous function $V : \triangle(\mathcal{I}) \times \mathcal{A} \to \mathbb{R}$ and, for each i in $\mathcal{I}$, a function $V_i : \mathcal{A} \to \mathbb{R}$, such that V represents $\succsim$; for each i, V_i represents $\succsim_i$; and for all (z, a) in $\triangle(\mathcal{I}) \times \mathcal{A}$*

$$V(z, a) = \sum_{i=1}^{I} z_i V_i(a).$$

The proof appears in an Appendix but let us just note here that the main step is to show that we can always find a countable number of 'preference intervals' with 'end-points' consisting of identity lottery/social alternative pairs of the form $(i, \hat{a}) \succ (j, \hat{a})$ that cover the indifference sets. For a given preference interval with end points $(i, \hat{a}) \succ (j, \hat{a})$, independence over identity lotteries enables us to construct a representation $\hat{V}(z, a)$ that is affine in z, for all identity lottery/social alternative pairs (z, a) that lie in this preference interval (that is, for which $(i, \hat{a}) \succsim (z, a) \succsim (j, \hat{a})$). We then construct a representation that is affine in identity lotteries using these intervals. One complication is that the intervals may not overlap. Without such overlapping, we do not obtain uniqueness. The precise condition guaranteeing uniqueness is given in terms of the representation in Case 1 of the proof of Theorem 1. To express this condition as a property of the preferences first let us say an identity lottery/social alternative pair (z, a) is non-extreme (wrt $\succsim$) if there exists two other pairs identity lottery/social alternative pairs (z', a') and (z'', a'') such that $(z', a') \succ (z, a) \succ (z'', a'')$. Loosely speaking, the condition on the observer's preferences guaranteeing uniqueness is that any non-extreme identity lottery/social alternative pair must lie in the 'interior' of some 'preference interval' of the form discussed above. More formally, we require:

Property U For each non-extreme (wrt $\succsim$) identity lottery/social alternative pair (z, a) in $\triangle(\mathcal{I}) \times \mathcal{A}$, there exists a social alternative $\bar{a}$ in $\mathcal{A}$ and two individuals i and j in $\mathcal{I}$ such that $(i, \bar{a}) \succ (z, a) \succ (j, \bar{a})$.

In Grant et al. (2010, Theorem 1, p. 1947), we obtained uniqueness for our generalized utilitarian representation by imposing a richness condition on the domain of individual preferences that assumed none of the social alternatives under consideration was Pareto dominated.

Absence of unanimity For all a, a' in $\mathcal{A}$, if $a \succ_i a'$ for some i in $\mathcal{I}$, then there exists j in $\mathcal{I}$ such that $a' \succ_i a$.

We argued in that article the condition was perhaps a natural restriction in the context of Harsanyi's thought experiment for a society of expected utility maximizers. In the current context where the theorem deals with the social alternatives themselves and not lotteries over social alternatives it might again be argued this is a natural restriction since the exercise is motivated by the need to make social choices when agents disagree. We do not need to imagine ourselves as an impartial observer facing an identity lottery to rule out social alternatives that are Pareto dominated.[7]

The next proposition shows that in the context of Theorem 1 absence of unanimity is sufficient to ensure Property U holds and so the generalized utilitarian representation is unique up to common affine transformations of the functions V_i.

Proposition 1 *Suppose the observer's preferences $\succsim$ satisfy the acceptance principle and independence over identity lotteries. Then absence of unanimity implies Property U.*

4 Applications and related literature

If, as Harsanyi did, we take $\mathcal{A}$ to be the set of lotteries over some set of social outcomes $\mathcal{X}$, then the observer preferences $\succsim$ are defined on the set of product lotteries $\triangle(\mathcal{I}) \times \triangle(\mathcal{X})$, and Theorem 1 implies there exists a generalized utilitarian representation of the form

$$V(z, \ell) = \sum_{i=1}^{I} z_i V_i(\ell).$$

This is the same representation we obtain in our companion paper (Grant et al. 2010, Lemma 8). And since we also assume there absence of unanimity we are able to show this representation is unique (up to common affine transformations of the functions V_is). The combination of acceptance and that each individual is an expected utility maximizer also means that each function $V_i(\ell)$ can be expressed as a monotonic transformation of that individual's expected utility. That is, $V_i(\ell) = \phi_i \circ U_i(\ell)$, where $U_i(\cdot)$ is an expected utility representation of $\succsim_i$ (see Grant et al. 2010, Theorem 1). Safra and Weissengrin (2003) also obtain a unique representation. Instead of requiring an existential axiom, however, they strengthen Independence over Identity Lotteries so that the independence property holds for *any* mixture of two product lotteries that is itself a product lottery. For example, if $(z, \ell) \sim (z', \ell')$ then they require $(z, \tilde{\ell}) \succsim (\tilde{z}', \ell')$ if and only if

$$\begin{aligned}\lambda(z, \tilde{\ell}) + (1-\lambda)(z, \ell) = \left(z, \lambda\tilde{\ell} + (1-\lambda\ell)\right) &\succsim \left(\lambda\tilde{z}' + (1-\lambda)z', \ell'\right)\\ &= \lambda\left(\tilde{z}', \ell'\right) + (1-\lambda)\left(z', \ell'\right),\end{aligned}$$

for all λ in $(0, 1)$. Moreover, in combination with acceptance their stronger independence axiom *implies* each individual is an expected utility maximizer and each function $V_i(\ell)$ is itself an expected utility representation of $\succsim_i$.

Karni and Safra (2000) can also be recast as a special case of Theorem 1 for which $\mathcal{A} = \triangle(\mathcal{X})^n$. That is, the observer may be viewed as having preferences over $\triangle(\mathcal{I}) \times$

[7] However, if $\mathcal{A} = \triangle(\mathcal{X})$ and the individuals are not expected utility maximizers, then it could be the case an outcome lottery ℓ in $\triangle(\mathcal{X})$ might not be Pareto dominated and so potentially might be ranked optimal by the observer even if every pure alternative in its support was Pareto dominated.

$\triangle(\mathcal{X})^n$, each individual has a personal conditional outcome lottery ℓ_i in place of the common outcome lottery ℓ and so the representation obtained takes the form, $V(z, \ell_1, \ldots, \ell_n) = \sum_{i=1}^{I} z_i V_i(\ell_i)$.[8]

Finally, consider the situation where individual preferences are defined over purely subjectively uncertain acts. In particular, let S be a state space (possibly infinite), let Σ be a $(\sigma-)$algebra of subsets of S (the set of measurable events) and assume that $\succsim_i$ are defined over the set of simple Σ-adapted acts $\mathcal{F}$.[9] Taking $\mathcal{A} = \mathcal{F}$, Theorem 1 implies that $\succsim$ admits a representation of the form $V(z, f) = \sum_{i=1}^{I} z_i V_i(f)$, where V_i is a representation of $\succsim_i$. If we required each individual to be a subjective expected utility maximizer then, for each i, there would exist a probability measure π_i defined on Σ, a (Bernoulli) utility index $u_i : \mathcal{X} \to \mathbb{R}$, and a strictly increasing real-valued function ϕ_i, such that $V_i(f) = \phi_i\left(\sum_{x\in\mathcal{X}} \pi_i\left(f^{-1}(x)\right) u_i(x)\right)$. Corresponding representations would be generated if an individual conformed to any of the generalizations of subjective expected utility that have appeared in the literature over the past 25 years to accommodate Allais and/or Ellsberg type 'paradoxical' behaviors.

Acknowledgments We thank the editor and two referees for their constructive criticisms and suggestions. We also thank Jurgen Eichberger, Edi Karni, Bart Lipman, Stephen Morris, Heve Moulin, David Pearce, John Quiggin, John Roemer, and John Weymark for many helpful comments. Atsushi Kajii thanks Grant-in-Aid for Scientific Research S (grant no. 90152298) and the Inamori foundation for their support. Zvi Safra thanks the Israel Science Foundation (grant no. 1299/05) and the Henry Crown Institute of Business Research for their support.

Appendix A: Example

Example 1 shows that the observer's preferences can satisfy the acceptance principle and *conditional* independence over identity lotteries for the impartial observer, moreover every individual is an expected utility maximizer and yet the observer's preferences do not satisfy the (unconditional) independence axiom over identity lotteries. Hence by Theorem 1 these preferences do not admit a generalized utilitarian representation.

In this example, $\mathcal{I} = \{1, 2\}$ and $\mathcal{A} = \triangle(\{x_1, x_2\})$. To simplify notation, for each $z \in \triangle(\mathcal{I})$, let $q = z_2$; and for each outcome lottery $\ell \in \mathcal{A}$ let $p := \ell(x_2)$. Then, with slight abuse of notation, we write $(q, p) \succsim \left(q', p'\right)$ for $(z, \ell) \succsim \left(z', \ell'\right)$, and write $V(q, p)$ for $V(z, a)$.

Example 1 Let agent 1 's preferences be given by $U_1(p) = -(2p-1)/4$, and let agent 2's preferences be given by $U_2(p) = 3(2p-1)/4$. Notice that these individual preferences satisfy independence. Let the impartial observer's preferences be given by $V(q, p) = (2p-1)\left(q^2 - 1/4\right)$.

[8] Notice there is a natural homeomorphism $h : \triangle(\mathcal{I}) \times \triangle(\mathcal{X})^n \to \triangle(\mathcal{I} \times \mathcal{X})$, where $h(z, \ell_1, \ldots, \ell_n) = e$ is the extended lottery defined by setting $e(i, x) := z_i \times \ell_i(x)$. Thus this example also encompasses Harsanyi's original setting where the impartial observer's preferences are defined over the set of extended lotteries $\triangle(\mathcal{I} \times \mathcal{X})$.

[9] That is, $f : S \to \mathcal{X}$ is in $\mathcal{F}$, if the range of f is finite and for every x in $\mathcal{X}$, $f^{-1}(x) \in \Sigma$.

By construction, these preferences satisfy the acceptance principle. To show that they satisfy conditional independence over identity lotteries, notice that for each fixed $\bar{p}$, the function $V(q, \bar{p})$ is monotone in q. If $\bar{p} = 1/2$, then the impartial observer is indifferent over all q and conditional independence follows trivially. If $\bar{p} > 1/2$, then $(\tilde{q}, \bar{p}) \succsim (\tilde{q}', \bar{p})$ if and only if $\tilde{q} \geq \tilde{q}'$. Thus, for all $\alpha \in (0, 1]$, $(\alpha\tilde{q} + (1-\alpha) q, \bar{p}) \succsim (\alpha\tilde{q}' + (1-\alpha) q, \bar{p})$ if and only if $\tilde{q} \geq \tilde{q}'$. The case for $\bar{p} < 1/2$ is similar.

To show that these preferences violate (unconditional) independence over identity lotteries, let $p = 0$ and let $p' = 1$. Let $q = q' = 1/2$ so that $V(q, p) = V(q', p') = 0$. Let $\tilde{q} = 0$ and let $\tilde{q}' = 1/\sqrt{2}$ so that $V(\tilde{q}, p) = V\left(\tilde{q}', p'\right) = 1/4$. Let $\alpha = 1/2$. Then $V(\alpha\tilde{q} + (1-\alpha) q, p) = -\left((1/4)^2 - 1/4\right) = 3/16$. But $V(\alpha\tilde{q}' + \left(1-\alpha\right)q', p') = \left(1/4 + 1\Big/\left(2\sqrt{2}\right)\right)^2 - 1/4 = \left(1/16 + 1/8 + 1/\left(4\sqrt{2}\right) - 1/4\right) = \left(3/16 + \left(1-\sqrt{2}\right)\Big/\left(4\sqrt{2}\right)\right) < 3/16$, violating (unconditional) independence.

□

Appendix B: Proofs

Proof of Theorem 1 Since the representation is affine in identity lotteries and each V_i represents $\succsim_i$, it is immediate that the represented preferences satisfy the axioms. We will show that the axioms imply the representation.

The idea of the proof is to cover the space $\triangle(\mathcal{I}) \times \mathcal{A}$ with preference intervals of the form

$$\left\{(z, a) \in \triangle(\mathcal{I}) \times \mathcal{A} : \left(i, a'\right) \succ (z, a) \succ \left(j, a'\right) \text{ for some } i, j \in \mathcal{I} \text{ and } a' \in A\right\},$$

then to define an affine function for every such interval and, finally, to glue these functions to obtain the desired representation. But there might exist problem product lotteries (z, a) (other than just the best or the worst, which exist by compactness and continuity) such that no social alternative a' and individuals i and j exist with the property above. If (z, a) is a such a problem identity lottery/social alternative pair then all identity lottery/social alternative pairs in its indifference set have the same problem.

Let $\tilde{V}$ be a continuous utility function representing $\succsim$ that we can use as a benchmark to label indifference sets and let T denote its image. Without loss of generality, assume that T is a subset of $[0, 1]$ with $\{0, 1\} \in T$. Note that T is closed. Let us define the set of problem indifference levels as follows

$$\begin{aligned} \mathcal{V} = &\left\{v \in T\backslash\{0, 1\} : \nexists i, j \in \mathcal{I} \text{ and } a \in \mathcal{A} \text{ s.t. } \tilde{V}(i, a) > v > \tilde{V}(j, a)\right\} \\ &\cup \{0\} \ \left(\text{if } \tilde{V}(i, a) = 0 \Rightarrow \forall j \in I \ \ \tilde{V}(j, a) = 0\right) \\ &\cup \{1\} \ \left(\text{if } \tilde{V}(i, a) = 1 \Rightarrow \forall j \in I \ \ \tilde{V}(j, a) = 1\right) \end{aligned}$$

The set $\mathcal{V}$ is closed. Suppose not. That is, let $v_n \to v$ be a sequence in $\mathcal{V}$ where $v \in T\backslash\mathcal{V}$ and assume first that $v \neq 0, 1$. Hence there exist $i, j \in \mathcal{I}$ and $a \in \mathcal{A}$ such that $\tilde{V}(i,a) > v > \tilde{V}(j,a)$. Then for sufficiently large n, $\tilde{V}(i,a) > v_n > \tilde{V}(j,a)$: a contradiction. If $v = 0$, assume there exist $i \in \mathcal{I}$ and $a \in \mathcal{A}$ such that $\tilde{V}(i,a) > 0$ and then, for sufficiently large n, $\tilde{V}(i,a) > v_n > 0$: a contradiction. Similarly, if $v = 1$ let $j \in \mathcal{I}$ and $a \in \mathcal{A}$ be such that $1 > \tilde{V}(j,a)$ and continue as above.

By the definition of $\mathcal{V}$, if $\tilde{V}(z,a) = v \in \mathcal{V}$ then either $\min_i \tilde{V}(i,a) \geq v$ or $\max_i \tilde{V}(i,a) \leq v$. By independence over identity lotteries, $\max_i \tilde{V}(i,a) \geq \tilde{V}(z,a) \geq \min_i \tilde{V}(i,a)$. Hence, there exists at least one individual j such that $\tilde{V}(j,a) = v$. And, using independence over identity lotteries again, $\tilde{V}(z,a) = v$ implies $V(i,a) = v$ for all individuals i in the support of the identity lottery z.[10]

The proof proceeds in three cases.

Case 1 Assume that the set $\mathcal{V}$ is empty. For each t in $T \setminus \{0,1\}$ we can find a social alternative a^t for which there exist individuals i^t and j^t such that $\tilde{V}\left(i^t,a^t\right) > t > \tilde{V}\left(j^t,a^t\right)$. Similarly, for $t = 0$ there exists a^0 and individuals i^0 and j^0 with $\tilde{V}\left(i^0,a^0\right) > 0 = \tilde{V}\left(j^0,a^0\right)$ and, for $t = 1$, there exists a^1 and individuals i^1 and j^1 with $\tilde{V}\left(i^1,a^1\right) = 1 > \tilde{V}\left(j^1,a^1\right)$. Note that, by independence over identity lotteries, the intervals $(\tilde{V}\left(j^t,a^t\right), \tilde{V}\left(i^t,a^t\right))$, $[\tilde{V}\left(j^0,a^0\right), \tilde{V}\left(i^0,a^0\right))$ and $(\tilde{V}\left(j^1,a^1\right), \tilde{V}\left(i^1,a^1\right)]$ are subsets of T and hence T is open (relative to $[0,1]$). As T is also closed, it follows that $T = [0,1]$.

Let

$$BL^0 = \left\{(z,a) \in \triangle(\mathcal{I}) \times \mathcal{A} : \tilde{V}\left(i^0,a^0\right) > \tilde{V}(z,a) \geq \tilde{V}\left(j^0,a^0\right)\right\}$$
$$BL^1 = \left\{(z,a) \in \triangle(\mathcal{I}) \times \mathcal{A} : \tilde{V}\left(i^1,a^1\right) \geq \tilde{V}(z,a) > \tilde{V}\left(j^1,a^1\right)\right\}$$

and, for $t \in (0,1)$

$$BL^t = \left\{(z,a) \in \triangle(\mathcal{I}) \times \mathcal{A} : \tilde{V}\left(i^t,a^t\right) > \tilde{V}(z,a) > \tilde{V}\left(j^t,a^t\right)\right\}$$

By compactness of $\triangle(\mathcal{I}) \times \mathcal{A}$, its open cover $\left\{BL^t\right\}_{t\in[0,1]}$ contains a finite cover $\left\{BL^{t_1},\ldots,BL^{t_K}\right\}$, where the intersection of any two adjacent sets is non-empty.

Next we show that, for each k, a representing function for $\succsim$ of the form $V^k(z,a) = \Sigma_i z_i V_i^k(a)$ can be constructed on the closure of BL^{t_k}.[11] For each $(z,a) \in BL^{t_k}$, let $V^k(z,a)$ be defined by

$$\left(V^k(z,a)\left[i^{t_k}\right] + \left(1 - V^k(z,a)\right)\left[j^{t_k}\right], a^{t_k}\right) \sim (z,a).$$

[10] Moreover, for all v in the interior of $\mathcal{V}$, $\tilde{V}(i,a) = v$ for all individuals i. Let j satisfy $\tilde{V}(j,a) = v$ and suppose, by way of negation and without loss of generality, that $\tilde{V}(i,a) < v$ for some i. Then for all v' such that $\tilde{V}(i,a) < v' < v$, we have $\tilde{V}(i,a) < v' < \tilde{V}(j,a)$. A contradiction to v being interior.

[11] This is similar to Case 1 in Safra and Weissengrin (2003, p. 184). This case is also analogous to Case 1 of Karni and Safra (2000, p. 320) except that, in their setting, the analog of a^{t_k} is a *vector* of outcome lotteries, with a different outcome lottery for each agent.

By continuity and independence over identity lotteries, such a $V^k(z,a)$ exists and is unique. To show that this representation is affine, notice that if $(V^k(z,a)[i^{t_k}] + (1-V^k(z,a))[j^{t_k}], a^{t_k}) \sim (z,a)$ and $\left(V^k(z',a)\left[i^{t_k}\right] + \left(1 - V^k(z',a)\right)\left[j^{t_k}\right], a^{t_k}\right) \sim \left(z',a\right)$ then independence over identity lotteries implies $([\alpha V^k(z,a) + (1-\alpha) V^k(z',a)][i^{t_k}] + [1-\alpha V^k(z,a)-(1-\alpha)V^k(z',a)][j^{t_k}], a^{t_k}) \sim \left(\alpha z + (1-\alpha) z', a\right)$. Hence $V^k\left(\alpha z + (1-\alpha) z', a\right) = \alpha V^k(z,a) + (1-\alpha) V^k(z',a)$. Since any identity lottery z in $\Delta(\mathcal{I})$ can be written as $z = \sum_i z_i [i]$, proceeding sequentially on $\mathcal{I}$, affinity implies $V^k(z,a) = \sum_i z_i V^k(i,a)$. Finally, by acceptance, $V^k(i,\cdot)$ agrees with $\succsim_i$ on the relevant social alternatives in $\mathcal{A}$. Hence, if we define $V_i^k(a) = V^k(i,a)$, then V_i^k represents individual i's preferences on the relevant social alternatives in $\mathcal{A}$. On this subset, the functions V_i^k are unique up to common affine transformation.[12]

Now, as any two adjacent intervals $BL^{t_k}, BL^{t_{k+1}}$ overlap, we can apply an affine re-normalization of either V^k or V^{k+1} such that the (re-normalized) representations agree on the 'overlap'. Since both functions are affine, the re-normalized representation is affine on $BL^{t_k} \cup BL^{t_{k+1}}$ and has the form $V^{k,k+1}(z,a) = \sum_i z_i V_i^{k,k+1}(a)$ as before. Again, uniqueness follows from standard arguments.

Finally, doing it for any two adjacent intervals yields the desired (unique) representation function $V(z,a) = \Sigma_i z_i V_i(a)$ for $\succsim$, where, for each i in $\mathcal{I}$ (by acceptance) the function $V_i(a) := V(i,a)$ represents $\succsim_i$ on $\mathcal{A}$.

Case 2 Assume that the set $\mathcal{V}$ is finite. We can write $\mathcal{V} = \{v_1, \ldots, v_{K-1}\}$ where $k' > k$ implies $v_{k'} > v_k$. Assume, without loss of generality, that $0, 1 \notin \mathcal{V}$ and define $v_0 := 0$ and $v_K := 1$. Fix an interval of the form $\left[v_{k-1}, v_k\right], k = 1, \ldots, K$. If the intersection $\left[v_{k-1}, v_k\right] \cap T$ is empty then there is nothing to do. Assume that it is non-empty and note that, similarly to the argument used in Case 1 to show that $T = [0,1]$, it now follows that $\left[v_{k-1}, v_k\right] \cap T = \left[v_{k-1}, v_k\right]$. By independence over identity lotteries, if $\tilde{V}(z,a) \in (v_{k-1}, v_k)$ then $\min_i \tilde{V}(i,a) < v_k$. Hence, by the continuity of $\succsim$ and the definition of $\mathcal{V}$, $\max_i \tilde{V}(i,a) \leq v_k$. Similarly, $\max_i \tilde{V}(i,a) > v_{k-1}$ and hence $\min_i \tilde{V}(i,a) \geq v_{k-1}$. That is, if $\tilde{V}(z,a) \in (v_{k-1}, v_k)$ then $\tilde{V}(i,a) \in \left[v_{k-1}, v_k\right]$ for all i. Moreover, if $\tilde{V}(z,a) = v_{k-1}$ (resp. v_k) then $\tilde{V}(i,a) = v_{k-1}$ (resp. v_k) for all i in the support of z. Therefore, following the method of case 1, we can construct a function $V^k(z,a) = \Sigma_i z_i V_i^k(a)$ that represents $\succsim$ on $\left\{(z,\ell) \in \Delta(\mathcal{I}) \times \mathcal{A} : \tilde{V}(z,a) \in \left[v_{k-1}, v_k\right]\right\}$ and we can re-normalize it so that its image equals $\left[\frac{k-1}{K}, \frac{k}{K}\right]$. To conclude, let $\tilde{V}(z,a) = V^k(z,a)$ if $V^k(z,a) \in \left[v_{k-1}, v_k\right]$. Notice that, as this construction suggests, we do not have uniqueness in this case.

Case 3 Assume that $\mathcal{V}$ is infinite. For every $\tau \in T \backslash \mathcal{V}$, if exist, define

$$\tau^+ = \begin{cases} \min\{v \in \mathcal{V} : v > \tau\} & \text{if } \{v \in \mathcal{V} : v > \tau\} \neq \varnothing \\ 1 & \text{otherwise} \end{cases}$$

$$\tau^- = \begin{cases} \max\{v \in \mathcal{V} : v < \tau\} & \text{if } \{v \in \mathcal{V} : v < \tau\} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

[12] The uniqueness argument is standard: see, for example, Karni and Safra (2000, p. 321).

Clearly, if $\tau \notin \{0, 1\}$ then $\tau^- < \tau < \tau^+$ (recall $\mathcal{V}$ is a closed set). As in Cases 1 and 2, define functions V^τ and V_i^τ on $[\tau^-, \tau^+]$. The set $T \backslash \mathcal{V}$ is covered by a countable number of disjoint intervals of the form (τ^-, τ^+), and hence functions V and V_i can be constructed (inductively and continuously) on their closed union. In this way, the functions are also defined for $\mathcal{V}^0$, the set of $\mathcal{V}$'s boundary points.

Let v be an interior point of $\mathcal{V}$ and let

$$v^+ = \begin{cases} \min\left\{v' \in \mathcal{V}^0 : v' > v\right\} & \text{if } \left\{v' \in \mathcal{V}^0 : v' > v\right\} \neq \varnothing \\ 1 & \text{otherwise} \end{cases}$$
$$v^- = \begin{cases} \max\left\{v' \in \mathcal{V}^0 : v' < v\right\} & \text{if } \left\{v' \in \mathcal{V}^0 : v' < v\right\} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

Clearly, $v^- < v < v^+$ and all $\succsim_i$ agree on $\{a \in \mathcal{A} : \tilde{V}(z, a) \in (v^-, v^+)$ for some $z\}$. Choose V that (with continuity) agrees with V of the former step at the indifference sets that are associated with v^- and v^+, such that it represents $\succsim$ on this set. To conclude, define

$$V_i : \{a \in \mathcal{A} : \tilde{V}(z, a) \in (v^-, v^+) \text{ for some } z\} \to \mathbb{R}$$

by $V_i(a) = V(z, a)$ and note that $V(z, a) = \Sigma_i z_i V_i(a)$ is trivially satisfied. As the number of (non trivially) open components of $\mathcal{V}$ is also countable, the construction of the desired functions can be carried out easily.

Finally, the proof for $\mathcal{V} = T$ is derived as a special case of the previous arguments, as in this case $v^- = 0$ and $v^+ = 1$. □

Proof of Proposition 1 Consider any non-extreme (with respect to $\succsim$) $(z, x) \in \Delta(\mathcal{I}) \times \mathcal{A}$. Let $(\hat{z}, \hat{a})$ be maximal with respect to $\succsim$ and (z', a') be minimal with respect to $\succsim$ on $\Delta(\mathcal{I}) \times \mathcal{A}$. These maxima and minima exist because $\succsim$ is continuous and both $\Delta(\mathcal{I})$ and $\mathcal{A}$ are compact. Because $(\hat{z}, \hat{a})$ is maximal, by independence over identity lotteries, there exists an $i \in \mathcal{I}$ such that $(i, \hat{a}) \sim (\hat{z}, \hat{a})$. Similarly, there exists a $j \in \mathcal{I}$ such that $(j, a') \sim (z', a')$. (It could be that $\hat{z}$ is the degenerate identity lottery $[i]$ or that z' is the degenerate identity lottery $[j]$.) Because (z, a) is non-extreme, we therefore have $(i, \hat{a}) \succ (z, x) \succ (j, a')$. If $(i, \hat{a}) \sim (i, a')$, the result follows by setting $\bar{a} = a'$. Otherwise, $(i, \hat{a}) \succ (i, a')$ and, therefore, $\hat{a} \succ_i a'$ by acceptance. By absence of unanimity, there then exists a $k \in \mathcal{I}$ such that $a' \succ_k \hat{a}$. By acceptance, $(k, a') \succ (k, \hat{a})$. If $(z, a) \succeq (k, a')$, then $(i, \hat{a}) \succ (z, a) \succ (k, \hat{a})$ and the proof is complete. Otherwise, we have $(k, a') \succ (z, a) \succ (j, a')$. □

References

Broome J (1993) A cause of preference is not an object of preference. Soc Choice Welf 10:57–68

Ergin H, Gul F (2009) A theory of subjective compound lotteries. J Econ Theory 144(3):899–929

Grant S, Kajii A, Polak B, Safra Z (2010) Generalized utilitarianism and Harsanyi's impartial observer theorem. Econometrica 78(5):1939–1971

Harsanyi JC (1953) Cardinal utility in welfare economics and in the theory of risk-taking. J Political Econ 61:434–435

Harsanyi JC (1955) Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: comment. J Political Econ 63:309–321
Harsanyi JC (1975) Nonlinear social welfare functions: do welfare economists have a special exemption from Bayesian rationality? Theory Decis 6:311–332
Harsanyi JC (1977) Rational behavior and bargaining equilibrium in games and social situations. Cambridge University Press, Cambridge
Karni E, Safra Z (2000) An extension of a theorem of von Neumann and Morgenstern with an application to social choice. J Math Econ 34:315–327
Mongin P (2001) The impartial observer theorem of social ethics. Econ Philos 17:147–179
Rawls J (1951) Outline for a decision procedure for ethics. Philos Rev 40:177–197
Roemer JE (1992) Utilitarianism, the difference principle, and the veil of ignorance: an application of the theory of social situations. In: Selten R (ed) Rational interaction: essays in honor of John C. Harsanyi. Springer, Berlin pp 337–351
Safra Z, Weissengrin E (2003) Harsanyi's impartial observer theorem with a restricted domain. Soc Choice Welf 20(2):95–111
Weymark JA (1991) A reconsideration of the Harsanyi–Sen debate on utilitarianism. In: Elster J, Roemer JE (eds) Interpersonal comparisons of well-being. Cambridge University Press, Cambridge pp 255–320