Economics and Philosophy, **17** (2001) 147–179

THE IMPARTIAL OBSERVER THEOREM OF SOCIAL ETHICS

Philippe Mongin
Ecole Polytechnique & Centre National de la Recherche Scientifique

I would like to dedicate this paper to the memory of Louis-André Gérard-Varet. His premature death in January 2001 deprived us all of a genuinely broad theoretical economist with deep foundational interests. He was one of the few to be conversant with the intricacies of the theory of subjective probability, and he himself co-authored a formal reconstruction of the Impartial Observer Theorem (d'Aspremont and Gérard-Varet, 1991). At the time of publishing this alternative reconstruction, I cannot refrain from remembering his exceptionally active intelligence, as well as the fruitful interchanges we had.

1. GENERAL

Vickrey (1945) and Harsanyi (1953) are credited for having independently introduced the following argument. One should compare income distribution vectors from the viewpoint of an observer who, by assumption, knows the income values, but does not know who has what, and in particular does not know his own income. It is also assumed that this observer gives an equal chance to the outcome of landing in each possible position. Then, applying the von Neumann–Morgenstern theory of risk to this special informational context, one concludes that income distribution vectors must be ranked according to the mean rule of utilitarianism. The argument as a whole is often referred to as the

I acknowledge helpful discussions with Claude d'Aspremont on the subject of this paper, and I am grateful to Richard Bradley, John Broome, Marc Fleurbaey, Philippe Fontaine, Francesco Guala, Edi Karni, Serge Kolm, Robert Leonard, Isaac Levi, Wlodek Rabinowicz, and John Weymark for detailed comments on earlier drafts. I have also benefitted from remarks made during a seminar and two conference sessions in Cergy-Pontoise, Montréal and Lund.

Impartial Observer Theorem. It has no well-agreed formulation. Vickrey's supposedly seminal paper has just one paragraph on it, and it is quite informal (1945, in 1994, pp. 24–5). So is the slightly expanded restatement in Vickrey (1960). Harsanyi's 1953 contribution has only two pages without any symbolism, and his 1955 restatement is again very terse. It is only later that Harsanyi (1977a) came to restate his insights more formally – though, as we will see, not satisfactorily. At the same time, he put them more generally, dealing with abstract social states rather than just income distribution vectors. When we contrast Vickrey's and Harsanyi's versions in this paper, the latter will always mean Harsanyi's mature version.

We provide a reconstruction of the 'theorem', not in order to turn it into a piece of mathematics, which it cannot be, but to precisely identify all of the assumptions which have to be defended if it is to be regarded as a serious ethical argument. Our reconstruction is, like Harsanyi's in (1977a), based on the 'extended preference' framework of social choice theory, but differs from his in several respects. We emphasize the need to assume *uniformity of extended preferences* among individual observers; otherwise, the ordinary utilitarian formula cannot be derived. We argue that uniformity of extended preferences is undefended in Harsanyi's framework, and we are thus led to investigate weaker variants of the conclusion in which each ethical observer adopts a utilitarian formula of his own, with utility representations of the others' preferences depending on the particular observer.

We also depart from the historical versions of the 'theorem' in considering *subjective probability assessments* instead of Vickrey's and Harsanyi's equiprobable lotteries. Laplace's principle has raised innumerable objections, and this provides a serious, if only negative, reason for considering a Bayesian variant in which subjective priors can differ. Following this heuristics, we offer a novel and more sophisticated formalization of the Impartial Observer. Conceptually, this variant will be seen to have an effect similar to the previous one, that is, it entails observer-dependent additive formulas and thus falls short of the utilitarian objective. An advantage of this variant, however, is that it provides a partial answer to an objection classically raised against the use of the von Neumann–Morgenstern theorem in social ethics, that is, that there is no conceptual – in the sense of preference-based – reason for selecting the specific utility representations provided by the theorem.

Another line of argument we examine is to take *utility assessments rather than preference judgements* as primitive data for the axiomatic construction. Following this direction, one can reach an observer-independent formula by virtue of what we call the *causal account* of interpersonal comparisons. But as we explain, this formula is of the

'generalized' utilitarian sort – that is, it does not entail identical weights for the individuals. Also, the causal account does not deliver a theorem in a real sense. It is rather an addition of claims that are philosophically debatable. Those who accept this objective account would be better off in taking a more direct ethical approach than that of the Impartial Observer Theorem.

So the simplest message of the paper is this. There is no way in which the Impartial Observer Theorem can bridge the whole gap from impartiality to utilitarianism, even making generous allowance for technical assumptions. But it is possible to conclude that at least in the subjective version explained here, the reasoning proves *something* – even if the result is a long way from the official objective.

There is another point of general significance. Broadly speaking, impartiality amounts to disregarding what is irrelevant in the peculiarities of a case when making a judgement on this case. In the present context, impartiality has received a more determinate meaning. Vickrey and Harsanyi equate 'disregarding' with 'not knowing', that is, they interpret impartial judgements as being those made in a situation of hypothetical ignorance. This is the seminal idea underlying the Impartial Observer Theorem; it then leads to the surprising application of decision theory to an ethical context. At the same time, Vickrey's and Harsanyi's commentators understand them as also employing a commonsensical notion of impartiality that recommends equal treatment of the individuals. Our analysis will reveal a tension between the two notions of impartiality. We will argue that the Impartial Observer Theorem is best understood by starting only with the first, and then examining whether or not the second can be derived. We do not mean to suggest that impartiality as equal treatment lacks normative warrant. Quite the contrary. The point is that to *assume* it at the same time as the other concept takes the edge off the argument. In order to make the best of the Impartial Observer Theorem it seems methodologically sound to minimize the number of purely ethical postulates.

The paper is organized as follows. Section 2 briefly relates the Impartial Observer Theorem to the philosophical tradition of basing ethical judgements on impartiality. Section 3 provides a relatively literal reconstruction of Vickrey's version, and Section 4 moves to Harsanyi's (it is a minor contribution of this paper to clarify the differences between Vickrey and Harsanyi. As will be explained in Section 4, Harsanyi's version needs three axioms besides the VNM one, to be called here, *Equal Chance*, *Consideration of Others*, and *Uniform Extended Preference*. Section 5 discusses the last two axioms, and Section 6 discusses the first while explaining our subjective probability reconstruction. Section 7 provides a summary assessment of the Impartial Observer Theorem. The technical details are covered in the appendix.

2. THE ETHICS OF IMPARTIALITY

There is an important philosophical tradition which emphasizes impartiality as a distinctive origin of moral judgements on collective life – notably, but not exclusively, in matters of justice. Allegedly, these judgements should not depend on individuals' identities and other particular circumstances. It is also argued that the symmetry or interchangeability requirements implied by impartiality go a long way towards determining not only the form, but the *content* of moral judgements. As far as we can see, this broad impartiality tradition borrows from two major sources – that is, the eighteenth-century Scottish writers, especially Adam Smith in *The Theory of Moral Sentiments* (1759), and Kantianism rather than Kant himself (because it uses only edulcorated versions of the Categorical Imperative and the Universalization Maxim). Encompassing though it is, the impartiality approach must be kept distinct from that of state-of-nature (or contractarian) theories. This difference is not sufficiently well reflected in today's textbook comparison between Rawls and Harsanyi, which treats them both as if they unproblematically belonged to the impartiality tradition.[1] We eschew the task of arguing for these broad claims and focus instead on the philosophical background of the Impartial Observer Theorem.

In 1953 Harsanyi identified the foundations of morality with 'nonegoistic impersonal judgements of preference' – a statement reiterated in all his subsequent work. For instance, he claimed in (1977a, p. 49) that 'the moral point of view is essentially the point of view of a *sympathetic* but *impartial* observer'. Notice that in Harsanyi's mature formulation, impersonality or impartiality has become compounded with sympathy. The passage just quoted only refers to Adam Smith, but in a 1958 paper Harsanyi explicitly endorsed a version of Kant's universalization maxim. His work thus reproduces the combination of Scottish and Kantian elements that is typical of the impartiality tradition. It is also likely, though not entirely clear-cut, that Harsanyi recognizes the difference between the mental experiment involved in his observer construction, and the hypothetical histories that underlie the state-of-nature approach. Finally, even if the 1953 paper is ostensibly concerned with income distribution, Harsanyi appears to strive towards a complete system of ethics rather than just a theory of economic justice.[2] All in all, Harsanyi, if perhaps not Vickrey, should count as a major representative of the ethics of impartiality among twentieth-century writers.

[1] Rawls's theory, not Harsanyi's, is the problem here. It does not belong solely to the impartiality tradition. Hampton (1980) has discussed the sense in which it also belongs to the social contract tradition.

[2] Witness the distinctions he makes between ethics and other forms of rational behaviour in (1977a).

We aim at analysing the specific contribution of these authors' 'theorem' to the impartiality tradition, and specifically, at clarifying the stark contrast between the weak philosophical premiss that philosophically motivates the reasoning, and its strong and questionable conclusion. How does one proceed from impartiality, possibly compounded with sympathy, to the utilitarian mean rule? It *must* be the case that the assumptions of the 'theorem' are more than just a formal dressing of the intuition of the sympathetic-but-impartial-observer. The analytical task, then, is first to delineate the added logical content, and second, to assess its conceptual significance. We have already pointed out that the added assumptions include that of von Neumann–Morgenstern (VNM) rationality in order to model the observer's judgements. Both in Vickrey and Harsanyi this assumption drives the additive form of the social evaluation rule.[3] Critics of VNM rationality will then dismiss the utilitarian looking conclusion of the 'theorem' as being irrelevant. We will put aside this sweeping criticism, and despite well recognized difficulties, assume that the VNM axioms provide a satisfactory construal of rational preference under risk. We will assume that the related set of axioms introduced by Anscombe and Aumann's (1963) for the uncertainty case is equally acceptable. There are further and less obvious difficulties in the way of the argument, as will soon become clear.

3. VICKREY AND HARSANYI CONTRASTED[4]

Vickrey gives no reason why the observer should give an equal chance to each position in the society. In Harsanyi this Equal Chance (EC) principle is to some extent argued for. Harsanyi (1953) claims that an impartial or impersonal observer's judgements can be reproduced as any individual's judgements of preference in a situation of

> complete ignorance of what his own position, and the position of those near to his heart, would be within the system chosen. (1976, p. 4)

He then *adds* that this state of ignorance

> would be the case if he had exactly the same chance of obtaining the first position . . . or the second or the third, etc., up the last position. (ibid.)

This is as much as Harsanyi is willing to say in order to defend the EC principle. It seems clear that there are two steps in the argument, one

[3] Of the two, only Harsanyi exploits the fact that the additive conclusion can be phrased in terms of mean utilitarianism. Vickrey has in mind the Benthamite sum rule rather than the mean rule.

[4] What is said of Vickrey in this section is meant to apply to both his 1945 discussion, and the 'potential immigrant' discussion of his 1960 paper (in 1994, pp. 44–5). As we read it, the latter is but a brilliant illustrative restatement of the 1945 argument.

from impartiality to ignorance, and the other from ignorance to EC itself. Neither step is logically compelling, but the second one is specially easy to criticize. It amounts to Laplace's application of the 'principle of insufficient reason': complete ignorance should be modelled as equiprobability. There are famous objections against it. Rawls (1971, Section 28) endorses them, albeit in passing, because his conclusion is much more drastically that one should not use probability at all in order to model the observer's ignorance. We emphasize the following intermediate possibility: to reject Laplace's application of the 'principle of insufficient reason', while remaining within the confines of a probabilistic ('Bayesian') decision theory. This line of argument amounts to accepting the first step (from impartiality to ignorance) while rejecting the second (from ignorance to EC). It is pursued in Section 6.

Both Vickrey and Harsanyi adhere to the von Neumann–Morgenstern theory of preference under risk. This VNM assumption is of course distinct from EC. One may accept Laplace's principle, while disbelieving that lotteries should be evaluated in the linear way implied by the von Neumann–Morgenstern axioms. Rawlsian critics, who reject all the Vickrey–Harsanyi assumptions at once, should attend to these obvious distinctions – they do not always make them. Supposing now that equiprobable distributions and VNM preferences are relevant in modelling the impartial observer's judgements, one gets different formalizations, as well as significantly different ethical implications, depending on how one draws the line between what the observer is supposed to know and not to know. In sum, the contrast between a Rawlsian 'fair' observer and a Harsanyian 'impersonal' observer will eventually depend on *both* the analytical treatment *and* the factual content of ignorance.[5] In the case of Harsanyi *versus* Vickrey the distinction simply boils down to the content of ignorance, analysed in one and the same way.

The Vickreyan observer chooses an income distribution as would any individual, 'were he asked which of various variants of the economy he would like to become a member of, assuming that once he selects a given economy *with a given distribution of income*, he has an equal chance of landing in the shoes of each member of it' (1945, p. 329; own emphasis). Thus, Vickrey merely requires that members of society ignore their position on the income distribution ladder.[6] In 1953, Harsanyi's

[5] A point well recognized by Levi (1977) in his comparison of ignorance in Rawls and Harsanyi.

[6] Vickrey (1960, in 1994, pp. 44–5) compares the observer with a prospective immigrant who contemplates various communities to migrate to, and is uncertain as to what income he will receive in each of these. As before, income is the only variable of interest to the observer.

delineation of the content of ignorance was not clearly different from Vickrey's, but, as early as 1955, he turned in a different direction.

In 1955, Harsanyi summarized the content of his earlier paper, adding the following essential footnote: [Impersonality requires that the observer have an equal chance of] 'being put in the place of any individual member of the society, with regard not only to his objective social (and economic) conditions, but *also to his subjective attitudes and tastes*. In other words, he ought to judge the utility of another individual's position not in terms of his own attitudes and tastes but rather in terms of the attitudes and tastes of the individual actually holding this position' (1955, fn. 16, in 1976, p. 22; own emphasis). The point was reiterated, though somewhat differently, in (1977a, p. 52). Following these very clear suggestions, we conclude that, for Harsanyi, the observer's ignorance must extend to i's subjective features, including his preferences, and not only to the usual objects of individual preference comparisons, such as money incomes or consumption levels.

Enough has now been said to formally separate Vickrey's version from Harsanyi's. To model the Vickreyan observer let us suppose, very simply, that each individual $i = 1, \ldots, n$ compares, in terms of his actual preference ordering, the distribution vector x with an equal chance of receiving any of the n components of x, and the vector y with an equal chance of receiving any of the components of y. Formally, consider a finite (the assumption is for convenience) set of possible income levels A and construct the set X of conceivable income distributions by assigning an income level to each individual in all possible ways. Define $\Delta(A)$ to be the set of all lotteries (i.e., probability measures) on A. We single out for consideration the equiprobable lotteries, that is, for any $x = (x_1, \ldots x_n) \in X$,

$$L_x = (1/n(x_1), \ldots, 1/n(x_n)).$$

Each individual $i = 1, \ldots, n$ is endowed with von Neumann–Morgenstern (VNM) preferences $\geq i$ on $\Delta(A)$, so that L_x can be evaluated as:

$$v_i(L_x) = 1/n\ v_i(x_1) + \ldots + 1/n\ v_i(x_n)$$

for some VNM utility representation v_i of the individual preference.[7] We routinely identify outcomes in A with sure lotteries, so that we may use the same symbol v_i on the right- and left-hand sides. The expression 'VNM utility function' will refer sometimes to the utility representation on the given lottery set, sometimes to its implied restriction to the outcome set. By assumption, each individual is also endowed with a

[7] For a precise statement of von Neumann and Morgenstern's axioms and the ensuing representation theorem, the reader may consult Fishburn (1970). His treatment also provides background material for the subjective probability variant of Section 6.

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

moral preference $*\geq^{i}$ on X; it will give rise to utility representations w_i. Now, Vickrey's version of Equal Chance is:

(*) for all i, and all x, y in X, $x *\geq^{i} y$ iff $L_x \geq^{i} L_y$

We also need a notion of actual preference $^{\mathrm{A}}\geq^{i}$ on income distribution vectors. We introduce it here by following Vickrey's implicit assumption that the individual only cares about his own component in each distribution vector:

(**) for all i, and all x, y in X, $x\ ^{\mathrm{A}}\geq^{i} y$ iff $x_i \geq^{i} y_i$.

Then:

Vickrey's Impartial Observer Theorem: If all individuals i have identical VNM preferences on the lottery set $\Delta(A)$, there exist common utility representations u and w of, respectively, the individuals' actual and moral preferences on X, such that for all $x \in X$,

$$w(x) = 1/n\Sigma_{i=1}^{n}\ u(x).$$

This is indeed a utilitarian formula, but it has been obtained, quite trivially, by assuming the individual *von Neumann–Morgenstern* preferences to be uniform. As Pattanaik (1968) has emphasized in discussing Vickrey (1960), the assumption is hard to accept. Individuals actually entertain varying risk attitudes, and moving to the normative side, there appears to be no reason why they should not. Harsanyi's version avoids this restriction, though at a price, as we will shortly see.

There is another, perhaps deeper, reason for taking leave of Vickrey's version and moving to Harsanyi's. It is dubious that an individual who retains his own preference to assess social states – even not knowing his position in the social state – manifests impartiality or impersonality to an extent sufficient to ground a *moral* judgement. The Impartial Observer should step outside himself. He should take account of his own interests no more and no less than if he were another individual. It follows that one's *actual* VNM preferences must not be used directly to assess social states morally. Axiom **(*)** has no ethical grounding. In Harsanyi's deeper construction, a new preference concept – extended preference – will mediate between moral preferences and actual VNM preferences. The critical point against **(*)** somehow gets lost because of Vickrey's uniformity assumption. Had he not made this unpalatable restriction, it would have become obvious that his axiom was too blunt. Vickrey's confusion between actual and ethically relevant assessments is probably facilitated by the fact that under standard economic assumptions,[8] a utilitarian formula with *identical* utility functions in the sum automatically recommends equality of income; so that it does not matter for the

[8] I. e., diminishing marginal utility and a well-behaved domain for the utility function.

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

conclusion what utility function is chosen. This independence of the rule from the specific utility function is limited to a highly particular case.[9]

4. EXTENDED PREFERENCE AND HARSANYI'S IMPARTIAL OBSERVER THEOREM

To formalize Harsanyi's version, one might want to borrow from the following, independently developed construction of social choice theory. After Arrow (1963, pp. 114–15), various writers in social choice theory have discussed 'extended sympathy', that is to say, judgements of the following sort: Alternative x is better (or worse) for individual i than alternative y for individual j. Arguing that it is possible and meaningful to make such judgements, Suppes (1966), Sen (1970, Chapter 9), Kolm (1972), Arrow (1977), Suzumura (1983) and others have formally elaborated the *extended preference approach*, which will be extensively employed here. This approach endows each member of the society i with both an actual preference relation defined on some set X of alternatives (social states), and a relation defined on suitably modified alternatives (x, j), to be interpreted as 'to be in social state x and in the position of j'.

That i can rank these 'extended alternatives' may be a light or heavy assumption, depending on how the individual's 'position' is construed. If it refers to the individual's position in the income distribution, or such similar objective features, there is perhaps nothing very problematic about it. At the other extreme, the 'position' may be construed so as to include all subjective features of the individual. Reading (x, j) in the latter way, extended preference allows for numerous interpersonal comparisons – to wit, i is able not only to compare what it means for j to be in x and to be in y, but even to compare what it means for j to be in x and what it means for k to be in y. To illustrate the wide range of meanings of 'extended preference', recall Sen's (1970, pp. 149–50) example in which j is a devout Muslim and k is a devout Hindu, while x and y are the states in which the individual, whoever he is, eats pork or beef, respectively. Sen's discussion was literally concerned with Suppes's principle of justice, which involves only the weak construal of the word 'position'. However, this well-known example can also be – and has been – read in accordance with the strong construal. Then, it says in effect that

[9] In view of the criticisms in this section, it seems permissible to disagree with the editors of Vickrey's *Selected Papers* when they claim that, 'as early as 1945 he sketched the basis of modern utilitarianism, later developed by Harsanyi' (1994, p. 5). Without the crucial step of endowing the observer with a special preference concept, Vickrey's VNM analysis has no ethical significance. Vickrey's editors also credit him for having 'not only the germ but the whole idea' of Rawls's original position (1994, p. 14). Again, this seems to be an overstatement.

i should be able to decide whether it is better or worse to break Hindu law while participating in the personal features of a Hindu (hence, adhering to Hindu law) than to break the Muslim law while participating in the features of a Muslim (hence, accepting Muslim law).

Once extended preference is introduced in a social choice theory, it must somehow be connected with actual preference. The usual linkage in the literature is that extended preference judgements should conform with actual preference judgements whenever this is possible, that is, any time it comes to comparing alternatives of the form (x, j) and (y, j) for some given j. We call this principle Consideration Of Others (CO). In the existing literature it is sometimes called the 'acceptance principle', and it is typically defended on the normative grounds of either 'nonpaternalism' or 'consumer's sovereignty'.[10]

Extended preference is a natural tool for modelling the impartial observer. Harsanyi's strong assumption on the content of the observer's ignorance calls for a correspondingly strong interpretation of this observer's 'position'. This still leaves open various theoretical possibilities, which we may classify as follows: (i) the second variable in (x, j) is the name of individual j; (ii) it refers to properties that j's preferences satisfy; (iii) it refers to causal factors which determine j's preferences. The distinction between the first interpretation and the other two should be clear: it is not the same to name 'John' and to list properties or factors bearing on John. The latter might provide a way of referring to John, but it would be a roundabout one, and arguably, an unsatisfactory one. There is a sense in which John's identity exceeds any description of John in terms of abstract features. The distinction between the last two interpretations is more elusive, if only because preference theory is in an unsettled state and specialists do not always agree on causal imputations in this area. But this distinction can at least be exemplified. As a relevant *property* of j's preferences, take his index of risk-aversion; as a possible *cause* for the value of his index, consider his wealth (since, according to standard theory, wealth influences risk-attitudes).

There may be a further problem with the proper way to understand (iii).[11] On the first reading, the second variable refers to those causal factors which bring it about that there is an individual j having the preferences he has. On the second reading, it refers to those causal factors which bring it about that j has the preferences he has, and it is then implied that j could have had other preferences while still being j. We take the second reading as allowing also for the possibility that k (different from j) could be subjected to the same causal factors as j, and therefore have the same preferences as j, while still being k. The second

[10] See, e.g., Sen (1970, p. 156).

[11] A problem of this sort was pointed out by Isaac Levi in correspondence.

reading offers more flexibility than the first. Essentially, it assumes that *j*'s and *k*'s identities can be defined independently of what their preferences are. We will take it for granted that it provides a suitable interpretation for (iii).

As we read it, the bulk of the extended preference literature is concerned with extended alternatives in the sense of either (i) or (ii), and sometimes ambiguously so.[12] It is conceivable to reconstruct the Impartial Observer Theorem on the basis of these particular interpretations.[13] But neither of them is Harsanyi's. In (1977a, pp. 58–9), he clearly differentiates between an individual's 'subjective attitudes (including his preferences)' and 'all the objective causal variables needed to explain these subjective attitudes', and he goes on to suggest that extended preference is best understood in terms of the latter. In keeping with this important comment, Harsanyi's (1977a, pp. 53–5) formal notion of extended preference, as well as the corresponding restatement of the 'theorem', should be read with interpretation (iii) in mind. That is, for each observer *i*, to make an extended preference judgement amounts to comparing, 'to be in social state *x* and under the influence of the factors determining *j*'s preferences', with 'to be social state *y* and under the influence of the factors determining *k*'s preferences'.

Let us call this version of extended preference the *causal* one. Harsanyi propounds it because it seems to convey effectively the notion that interpersonal comparisons of preferences can be objective in nature. The latter claim is recurrent in his work. He made it forcefully long before he thought of employing the extended preference apparatus. In his 1955 paper, where for the first time he introduced the famous distinction between two kinds of preferences ('subjective' and 'ethical', to be later called 'empirical' and 'moral'), Harsanyi discussed at some length interpersonal comparisons of utility. He claimed that they could be predicted from earlier psychological data :

> If two individuals have opposite preferences between two situations, we try to find out the psychological differences responsible for this disagreement and on the basis of our general knowledge of human psychology, try to judge to what extent these psychological differences are likely to increase or decrease their satisfaction derived from each situation. (1955, in 1976, p. 17)[14]

[12] Compare this interpretation with Suzumura's (1983, pp. 133–6).

[13] As perhaps Sen (1970, p. 150) had meant to suggest.

[14] There is an interesting and little noticed connection between Harsanyi's long-standing view that interpersonal comparisons are objective and his construction of the *type* concept in his pioneering work on games of incomplete information, see Mongin and d'Aspremont (1998). The game-theoretic variant of the Impartial Observer Theorem proposed by d'Aspremont and Gérard-Varet (1991) makes direct use of the concept of a player's type.

The only significant difference between this early formulation and the later ones in Harsanyi (1977a) is that the former does and the latter do not employ the language of extended preferences. Clearly, Harsanyi came to think that this notion would help him to convey his long-standing conviction more strongly and convincingly, especially among economists.

Having motivated the principles, we now introduce them axiomatically. Denote the initial alternative set X and the set of all individuals by $N = \{1, \ldots, n\}$. Then, $X \times N$ is the set of *extended alternatives*. We also introduce the set $\Delta(X \times N)$ of *extended lotteries*, among which are the extended equiprobable lotteries, that is: For any $x \in X$,

$$L_x = (1/n(x,1), \ldots, 1/n(x,n)).$$

Each individual $i = 1, \ldots, n$ is endowed with extended preferences $\mathrm{E} \geq i$ on $\Delta(X \times N)$ satisfying the VNM axioms, so that L_x can be evaluated as:

$$v_i(L_x) = 1/n\ v_i(x) + \ldots + 1/n\ v_i(x)$$

for some VNM utility representation v_i of i's extended preference.

Now, each individual is also endowed with an actual and a moral preference relation on X, respectively, $\geq i$ and $* \geq i$. (Since there is no possible confusion now, we drop the index A in the symbol of actual preference $\mathrm{A} \geq i$.) Utility representations for these preferences will be denoted by u_i and w_i, respectively. Harsanyi's Equal Chance principle connects the moral with the extended preference, and is formally stated like Vickrey's:

EC for all i, and all x, y in X, $x * \geq i\ y$ iff $L_x\ \mathrm{E} \geq i\ L_y$.

Harsanyi's Consideration of Others principle connects extended with actual preference:

CO for all i and j, and all x, y in X, $(x, j)\ \mathrm{E} \geq i\ (y, j)$ iff $x \geq j\ y$.

Finally, we introduce a principle – Uniformity Of Extended Preference (UEP) – which was not part of the earlier motivations. It says that extended preferences are the same from one observer to the other:

UEP for all i, j, and any two lotteries l, l' in $\Delta(X \times N)$, $l\ \mathrm{E} \geq i\ l'$ iff $l\ \mathrm{E} \geq j\ l'$.

Harsanyi's Impartial Observer Theorem: Assume that for each individual, the extended preference relation satisfies the VNM axioms, and that EC, CO and UEP hold. Then, there exists a common utility representation w of the individuals' moral preferences on X, such that: for all $x \in X$,

$$w(x) = 1/n\Sigma_{j=1}^{n}\ u_j(x),$$

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

where $u_1, \ldots, u_n$ are utility representations of the individuals' actual preferences.

The present statement is mathematically crude, but it strips Harsanyi's 'theorem' to its logical bare bones. It serves especially to highlight the role of UEP. Without it, the conclusion does not hold, and we only get that for each i, there exists a moral utility function $1/n\Sigma_{j=1}^{n} u_{ij}$, where u_{ij} is a utility representation for j's preference *depending on the particular observer* i. Harsanyi's own formal exposition (1977a, pp. 54–5) may suggest that the other three assumptions are sufficient to derive a set of utility representations obeying the mean rule of utilitarianism. This is not the case. Relevant details are provided in the appendix.

This (easy) logical point being granted, one should ask whether or not UEP is *conceptually* justified. Is it not the case that extended preferences can unproblematically be taken to be uniform from individual to individual? Before moving on to this and related discussions in Section 5, we record another problem that should be raised in connection with the above formalization. The statements just given of Vickrey's and Harsanyi's 'theorems' are purely existential. They do not include any uniqueness restriction on the derived utility representations w and u_j, $j = 1, \ldots, m$. In fact, a glance at the proof in the appendix shows that there are other choices than these representations that are compatible with the axioms, but do not deliver the desired additive representation. The functions w and u_j, $j = 1, \ldots, m$, have been selected among many to make the utilitarian-looking formula come right. Prima facie, this casts a doubt on the ethical significance of the Impartial Observer Theorem. This point has been made clearly and forcefully by Weymark (1991) in connection with his own formalization of the Impartial Observer Theorem.[15]

More specifically, we may consider replacing the VNM indexes u_j by *nonlinear* transforms of them, thus destroying the additive form of the social rule. Sen (1986) had argued that since this can always be done, Harsanyi's two theorems, namely, the Impartial Observer Theorem and the no less famous 1955 Aggregation Theorem, are only superficially connected with utilitarianism. A good deal of Weymark's (1991) reexamination of Harsanyi's work amounts to endorsing this objection. It is true that the preference relations of traditional theory are purely ordinal; so, if they are taken to be the only primitives, there can be no conceptual reason for selecting VNM indexes rather than any ordinal transforms of them. If, however, the notion of preference is understood in a different sense, the primitives may include a further relation to represent the *intensity* of preferences. Then adding suitable axioms, it becomes possible to restrict the range of permissible utility representa-

[15] See Weymark's (1991) Theorem 9 and the comments following it.

tions to the set of VNM indexes, and the Sen–Weymark objection can be countered, at least at the logical level. We have discussed this line of analysis elsewhere and will not pursue it here.[16]

Even granting that an argument can be made for VNM indexes, the choice of the particular VNM index clearly influences the symmetric form of the additive rule. A non-uniform linear rescaling of the u_j would still deliver a VNM representation for each individual, with all the axioms being satisfied, but with weights other than $1/n$. Hence, even if the Sen–Weymark objection is answered along the lines we suggested, there remains a significant arbitrariness problem. This further problem can be addressed within the confines of the present paper. In Section 6 we show how subjective probability theory can be put to use in order to fix the weights in a non-arbitrary way. The resulting formula, however, does not involve equal weights.

5. THE CAUSAL ACCOUNT AND UNIFORM EXTENDED PREFERENCE

Following the causal interpretation of extended preferences, $(x, j) \overset{E}{\geq} i$ (y, k) should be read as:

(*) 'individual i prefers to be in social state x and under the influence of the factors determining j's preferences than to be in social state y and under the influence of the factors determining k's preferences'. Compare this reading with the more standard one in which j and k directly refer to individuals:

(**) 'individual i prefers to be j in social state x than to be k in social state y'.

Preference judgements of the form (**) are typically analysed by saying that the observer i identifies himself with j's personality, or at least, manages to reproduce in himself j's preference attitudes. The words 'sympathy' and 'empathy' have been used to cover these psychological experiences. Something of this sort seems to be needed to explicate (**). By contrast, the causal account underlying (*) does not require the observer's entering a particular mental state. Here, the process of identifying oneself with the other gives way to the process of deducing what one's own preference would be under certain ideal conditions. Both the causal and the more standard account in terms of identification must be contrasted with the notion that *i records j's choices*

[16] In Mongin (1994) and, in more detail, in Mongin and d'Aspremont (1998) the possibility of cardinalizing the *preference relation* constitutes the gist of an argument in favour of Harsanyi's two theorems. Recently, Harvey (1999) has made this possiblity more concrete by formally reconstructing the Aggregation Theorem in terms of a cardinal preference relation. Note carefully that this is not the same reconstruction as that which makes *cardinal utilities* the primitives in Harsanyi's analysis.

in his extended preferences. A choice-based interpretation along this line could perhaps be all right as long as (x, j) and (y, j) are compared, but there is no way of extending it to comparisons between (x, j) and (y, k).

Most writers on extended preferences use *sympathy* and *empathy* interchangeably. If one is concerned with distinguishing between the two,[17] Harsanyi's account must definitely be put on the empathy side. Roughly speaking, sympathy has to do with the observer's ability to be affected in his own welfare by another's situation. Empathy is rather an ability to understand another person, possibly by swapping places with him and reproducing his experiences in oneself, but possibly also by purely deductive means. Empathy does not imply benevolence, just interest. The outcome of the empathetic exercise is a piece of knowledge that the observer (say, the historian, but it may also be the man in the street) can use for any purpose of his own, and not necessarily for the benefit of the observed individual. On this account, and barring several textual vacillations, Harsanyi would exemplify the intellectual extreme of the empathy concept. The only feature of sympathy that is important for him, as for any other extended preference theorist, is that by CO, the observer *positively* correlates some of his preference judgements with those of the observed person. In this limited sense, the impartial observer of social ethics, including Harsanyi's, is also sympathetic. How the correlation between two sets of judgements is obtained is a matter different from, and conceptually prior to, the direction of correlation. To reiterate, in Harsanyi it is an intellectual empathizing process that brings about the positive correlation. To decide whether it is better to be rich while being under the influence of expensive tastes or to be poorer while being under the influence of more economical tastes, the observer will use his knowledge of psychological laws and deductive ability.

A definite advantage of the causal account is that it may save the impartial observer theory from hazardous discussions of personal identity. Typically, those who understand extended comparisons in terms of one's identification with another's attitudes have stumbled on the following problem: how much of the observer's identity is preserved by sympathetic identification or by empathetic identification of the non-deductive sort? Is there enough left, as it were, to warrant the claim that it is the observer who makes the preference judgement? The point has been put forward that if i must effectively enter j's or k's mental state to make extended preference judgements, it cannot be i, after all, who

[17] As Fontaine (1997) usefully does; we are indebted to him here. A majority of writers on the Impartial Observer Theorem do not attempt to distinguish between 'sympathy' and 'empathy'.

makes them. There would be something self-destructive in the way identification works.[18] A thought-provoking claim like this depends on a detailed conception of personal identity, and not just on the way personal identity relates to preference maps. Luckily, Harsanyi does not have to delve into these deep waters of metaphysics. Following formula (*), the observer's preferences are plainly *this* observer's preferences. Harsanyi would simply have to make it clear that he understands preferences broadly enough. The observer's preferences are *considered* preferences, and they are defined over objects which in the particular instance are states of affairs, not objects of choice – a clear departure from 'revealed preference' theory.

If in (*) we take $j = k$, we get the following class of extended preference judgements:

'individual i prefers to be in social state x and under the influence of the factors determining j's preferences than to be in social state y and under the influence of the factors determining j's preferences'. It is the class to which CO applies. The axiom states that the previous statement must hold if and only if:

'individual j actually prefers to be in social state x than to be in social state y'. This equivalence can only be a matter of stipulation, not of logical necessity. Harsanyi, who calls CO 'the principle of acceptance' or of 'consumer's sovereignty' defends it as follows: 'it requires us to accept each individual's own personal preferences as the basic criterion for assessing the utility (personal welfare) that he will derive from any given situation' (1977a, p. 52). Essentially, CO plays the same role as the Pareto principle does in standard welfare economics and in Harsanyi's Aggregation Theorem. Even granting the 'qualifications' that he is willing to introduce in respect of both CO and the Pareto principle, both are surrounded with normative difficulties. These difficulties are neither new nor specific to the causal interpretation adopted here for extended preferences. So we gloss over them.

The benefits that Harsanyi (1977a, p. 58) hopes to reap from his causal account have to do exclusively with axiom UEP. Using the language of utility functions rather than preferences, he claims that the causal account implies an *identical* utility function for all observers. The rest of this section is devoted to examining this claim. We will argue that extended preferences cannot be independent of the particular observer. We will salvage a version of the claim in which utilities stand for themselves, without representing preferences.

[18] This claim is Rothenberg's (1961) main objection against Harsanyi, and it also arises in Kaneko (1984) and McKay (1986). Pattanaik (1968) expresses his disagreement with Rothenberg but does not really provide an alternative account.

To clarify the locus of the disagreement, our objection against Harsanyi does *not* relate to his extraordinary assumption that there exist psychological laws sufficiently precise to deliver predictions on the individuals' dispositions, and that all observers both know these laws and can make the correct inferences from them. Harsanyi himself is willing to admit that his assumption is, indeed, far-fetched. But as we read him, he would consider a failure of this assumption to be the *only* reason why extended preferences might after all differ from one observer to another. Differences among extended preferences would be completely explained by the primitive stage of psychological knowledge. They would not count as differences in principle. But we will argue that there are other reasons for doubting the conceptual validity of UEP, and these reasons are of a permanent sort.[19]

Take two observers i and h. On the face of it, the statement that:

$$(x, j)\mathrm{E} \geq^{i} (y, k) \text{ if and only if } \mathrm{E} \geq^{h} (y, k)$$

follows automatically neither in interpretation (*) nor in interpretation (**) – except, of course, whenever $j = k$ and CO is granted. Think of a rich man with expensive tastes and a poor man with economical tastes. Suppose that they are clever and knowledgeable so that they fully understand what it means to be rich with expensive tastes and poor with economical tastes. There is nothing in the theory of extended preference to exclude that these two individuals wish to swap their positions and subjective features. In a (doctored) version of La Fontaine's famous fable, the Banker would prefer to be poor and lighthearted, like the Shoemaker, and the Shoemaker would prefer to be rich and worried, like the Banker. Or take a Londoner who would like to live in Paris with the cultural tastes of a Parisian, while the Parisian would like to live in London, with the taste for success of a Londoner. There is nothing contradictory in these imagined situations.

But perhaps we are relying here on the ordinary interpretation (*), and the causal interpretation (**) would deliver a different conclusion. Let us try and give it another chance. Underlying the causal account is the thesis that all ordinary preference judgements are causally determined and all observers make full use of their supposedly complete knowledge of causes. Accordingly, the following claim is also part of the causal account:

[19] Compare with Broome (1993) who also argues against Harsanyi (and a few others, like Kolm, 1972) that extended preferences cannot be taken to be uniform from one individual to another. Broome reaches his conclusion by emphasizing the distinction between an object and a cause of preference, and we will proceed differently here, emphasizing instead the distinction between preference judgements and utility amounts. Broome's argument has apparently failed to convince Kolm (see his 1994 answer). Hopefully, the present argument will.

'All individuals i can predict which of x and y individual j will prefer'. With some effort, one can devise an analogue of this claim for extended comparisons involving different observed individuals j and k, as would be required in order to defend UEP:

(+) 'All individuals i can predict whether x satisfies j's preferences more than, less than, or equally as y satisfies k's preferences'.

We discard the difficulties introduced by comparisons of the degree, or the extent, to which the preferences of distinct individuals can be satisfied – a notion needed in order to state (+) – since there is a simpler critical point to make. Predictions relative to preferences, be they for one observed individual or for two, do not normally have the form of *preference* judgements. It would be fanciful to claim that (*) and (+) are synonymous. Still, some readers might overlook the difference between a preference judgement and a prediction of a preference judgement, and incorrectly conclude that the preference judgement in (*) is unanimous, starting from the correct understanding that in the causal account, predictions of (+) are unanimous.

Let us try to exploit the causal account in yet another way. Statement (*) refers to a preference judgement which, by the causal account or a plausible extension of it, must also be causally determined. It follows that any two observers, say l and m, will be able to predict (*) identically from their prior knowledge of the causes of this judgement. So the following holds:

'All individuals l can predict that i prefers to be in social state x and under the influence of the factors determining j's preferences than to be in social state y and under the influence of the factors determining k's preferences'. However, identical predictions made on the same observer i by different meta-observers are not at all what we were after. Supposing that we could deduce it, a more relevant fact would be this: all meta-observers make the same prediction for i and for any other observer h. But the causal account does not warrant any conclusion of the sort. It can be stretched to the point of saying that there are causal factors underlying i's and j's extended preferences, not that these factors are the same.

The previous hints are all that we could think of to defend UEP on the basis of the causal account. But what about Harsanyi's own defence? We do not think that he is confused to the point of reading a preference into a prediction. The following important point has gone generally unnoticed among commentators.[20] When Harsanyi really

[20] An exception is Weymark (1991) at several points in his thorough commentary. In his Theorem 10, Weymark exploits this observation to reconstruct the Impartial Observer Theorem in terms of cardinal utilities taken as primitive concepts. He does not consider its implications for the uniformity of moral assessments, as we do here.

argues about extended preferences, as opposed to just mentioning them in passing, or gesturing towards them, he replaces them with extended utility assessments. It is in *this* language that he makes an argument for uniformity:

> the extended utility function v_i should really be written as $v_i = v_i\,(x, R_j)$ . . . Written in this form, the utility function indicates the utility that individual i would assign to the objective position x if the causal variables determining his preferences were R_j. Because the mathematical form of this function is defined by the basic psychological laws governing people's choice behavior, this function must be the same for all individuals. (1977a, p. 58; we have adapted the notation)

Thus, utility values measure the individual's preference satisfaction, given the objective variables (e.g., income) and subjective variables (the individual's preference parameters) influencing it. By assumption, subjective variables act causally in a stable and recognizable way. It follows that the mapping from objective variables to numerical degrees of satisfaction is perceived in the same way by all observers. So far so good for this utility-based version of the causal account. But it does not deliver uniformity of extended *preferences*, only uniformity of the considered utility amounts. The statement that

$$v_i(x, R_j) \geq v_i(y, R_k)$$

might well hold for all i without representing any observer's preferences. Why should it? More utility can be obtained in x under causal circumstances R_j than in y under causal circumstances R_k. This does not imply that there is anybody who prefers, or should prefer, the first extended alternative to the second.

But is it not possible to improve on Harsanyi's utility-based argument? There is an obvious connection between utility and preference which we have not yet exploited, namely, the utility amounts $v_i(x, R_j)$ and $v_i(y, R_k)$ represent, respectively, the degree to which x satisfies j's preferences, and the degree to which y satisfies k's preferences. (There are difficulties with any conception which is concerned with social ethics, and takes utility numbers to represent degrees of satisfaction rather than well-being. But we are keeping as close as possible to Harsanyi's own argument, and thus we by-pass these difficulties here.) Accordingly, we may couch the statement that

$$v_i(x, R_j) \geq v_i(y, R_k) \text{ hold for all } i$$

in the language of preference, and reach the statement that:

(+) 'All individuals i can predict whether x satisfies j's preferences more, less, or equally than y satisfies k's preferences'. We have already encountered this statement. It leads nowhere. We conclude that there is

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

no way of bridging the gap between Harsanyi's argument in terms of utilities and axiom UEP.

This paper is concerned with the Impartial Observer Theorem, not with extended preference and extended utility theories in general, but it might help to reinforce the critique of this section if we briefly discuss writers other than Harsanyi.[21] Consider again Arrow's definition of 'extended sympathy' in 1963 (pp. 114–15). Literally understood, it does not deal with preference judgements, but with comparisons of individuals' well-being. It also seems clear that Kolm (1972) was concerned with well-being across individuals.[22] Let us formulate the *well-being* interpretation of extended comparisons explicitly, in order to contrast it with (*), (**), and (+):

(***) '*j* in social state *x* is better-off than *k* in social state *y*'. If extended comparisons are explicated by this statement, it is possible to take the further step that they are uniform from one observer to the other. Essentially, the argument for this conclusion will consist in saying that well-being is an objective concept and devising a causal account for it.

Although the objective perspective is only sketched in Arrow, Kolm and others, it is the most promising line for these writers to pursue. They sometimes come closer to making a different argument, which bears a definite resemblance to Harsanyi's and fails for the same general reason. They would no longer regard (***) as providing a *definition* of extended comparisons. Rather, they would construe them as being *preference* comparisons of some sort, and they would conclude from the alleged uniformity of these extended comparisons that well-being comparisons like (***) make good sense. The argument differs from Harsanyi's version of the 'theorem' in several respects. For one, Arrow, Kolm and most writers on extended comparisons take these comparisons to be only ordinal, whereas Harsanyi needs cardinal comparisons for his utilitarian conclusion to hold. For another, these writers are not so much interested in deriving a fully-fledged social rule as in justifying a relevant class of interpersonal comparisons. Finally, they are trying to reach well-being conclusions, as we have just said, whereas Harsanyi is best interpreted as a preference utilitarian. This said, the argument falls prey to the same objections as Harsanyi's version of the 'theorem'. It would be nice if comparisons of well-being could be deduced from some universally shared set of preferences, but there is no argument available to provide the desired set of preferences.

To return now to the Impartial Observer Theorem. Taking the notion of an extended *preference* at its face value, and because (UEP) is

[21] Compare again with Broome (1993).

[22] Kolm confirmed this to us. His answer (1994) to Broome is also rather explicit on this score.

indefensible on this reading, one is left with the following weaker version:

The 'Many Impartial Observers' Theorem: Assume that for each individual, the extended preference relation satisfies the VNM axioms, and that EC and CO hold. Then, for each i, there exists a representation w_i of his moral preferences on X, such that: for all $x \in X$,

$$w_i(x) = 1/n\Sigma_{j=1}^{n}\, u_{ij}(x),$$

where $u_{il}, \ldots, u_{in}$ are utility representations of the individuals' actual preferences.

6. EQUAL CHANCE

On a preliminary reading, Equal Chance is the axiom which in the Impartial Observer Theorem, corresponds to the informal idea of impartiality, while Consideration of Others would correspond to sympathy. This reading has become popular among Harsanyi's followers. In our interpretation there is no such simple pairing of informal ideas with axioms. If sympathy is anywhere, it is in CO, but the requisite of impartiality permeates the whole set of axioms, the arrangement of which is determined by the following powerful philosophical idea: *Impartiality must be analysed as a mode of ignorance*. Viewed in this light, EC is not a direct rendering of impartiality, but a step in an argument about impartiality. It is the connecting link between the informal notion of ignorance and the application of expected utility theory that eventually drives the conclusion. Accordingly, EC should *not* be seen as an equal treatment – that is, normative – axiom. It is an *epistemic* axiom – in effect, Laplace's principle of dealing with complete uncertainty in terms of equiprobable states of nature. Once this is accepted, important consequences follow, as this section will try to make clear.

Bayesian writers have always been divided on the significance of Laplace's principle. In case of ignorance, some – especially in econometrics – take 'diffuse priors' to be the starting point for Bayesian updating. But others are content with saying that priors are whatever they are. Neither the Dutch Book argument *à la* de Finetti, nor the more sophisticated axiomatic constructions by Savage (1954), Anscombe and Aumann (1963) or Jeffrey (1983), imply restrictions on the particular probability measures they derive. They just imply that there is a probability measure and that it is unique in some technically well-defined sense. When it comes to a multi-agent context of ignorance, as in the theory of games of incomplete information, Bayesians disagree on the question of whether or not there is a 'common prior'.[23] Foundational

[23] In a classic series of papers Harsanyi (1967–8) proposed to base the very notion of a game

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

issues of this sort should arise in any discussion of the Impartial Observer or related notions of the 'original position', as Rawls (1971, pp. 171–2) pointed out with respect to Laplace's principle. Without pursuing these complex issues in detail, we may stress that Harsanyi cannot avail himself of Bayesianism *tout court*, and that there is a plausible alternative within the confines of the doctrine, which replaces EC with the assumption of subjective probabilities, one for each observer i, whatever these probability measures are. In keeping with the spirit of axiomatic Bayesianism these measures should be *inferred* from antecedent preference conditions rather than taken for granted, and they should be *uniquely determined* by the inference process. The following construction is based on the Anscombe–Aumann version of axiomatic Bayesianism which turns out to be specially easy to relate to the Impartial Observer Theorem.[24]

We will introduce, and actually need for the proof, more structure than in the previous, more elementary versions. Let us replace the nondescript set of alternatives X by $X_1 \times \ldots \times X_n$, where the X_i are (finite) personalized sets of outcomes. Corresponding to each of these, there is a personalized lottery set $Y_i = \Delta(X_i)$, $i = 1, \ldots, n$. A *social state*, which will be the object of moral evaluation, is defined to be a mapping f: $N \to Y_1 \times \ldots \times Y_n$ such that for all i, $f(i) \in Y_i$. That is to say, a social state is an assignment of a personalized *lottery* to each individual; denote by $f(x_i, i)$ the probability value given by f to the event of i's getting outcome x_i. Thus, there are two added structural features compared with our earlier framework: outcomes are now individual-specific, and society assigns to each individual chances of getting these outcomes rather than a single outcome. One may, if one wishes, cancel the first feature while preserving the second, which is the technically important of the two. Then, take $X_1 = \ldots = X_n = A$, for example, a set of m feasible income levels, as in the above formalization of Vickrey. The set L of social states can be redescribed as the following set of vectors:

$$L = \{f \in (R+)^{mn} | \Sigma_{x \in A} f(x, i) = 1, \forall i \in N\}.$$

For each individual i we introduce a *subjective preference relation* $\overset{S}{\geq} i$ over L. It is this relation – instead of the extended preference relation $\overset{E}{\geq} i$ – that we will now make the basis of the individual's moral preferences. Hidden behind subjective preference judgements, there is one subjective probability measure for each i, which it is the purpose of the construction to deliver explicitly. These measures will replace the

of incomplete information on the 'common prior assumption'. Some game theorists have come to question this familiar assumption.

[24] This technical observation was made, but not yet put to use in Mongin and d'Aspremont (1998, p. 455). Meanwhile, it has been exploited by Karni (1998) in an interesting way; see the appendix.

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

uniform measure of the more standard version. Accordingly, EC will be substituted with the following Subjective Probability (SP) axiom:

SP for all i, and all f, g in L, $f^* \geq^i g$ iff $f \overset{S}{\geq}{}^i g$.

Notice that L is a convex set. As this is mathematically possible, we submit the relations $\overset{S}{\geq}{}^i$ (or equivalently, $^* \geq^i$) to the VNM axioms.

The notion of a *conditional* subjective preference $(\overset{S}{\geq}{}^i)_j$ will play an important role in the sequel. Intuitively, it is a restriction of i's subjective preference relation to individual j's personalized lotteries, disregarding what happens to the other individuals. Formally, it is defined from $\overset{S}{\geq}{}^i$ as follows. For all i and j,

$f\ (\overset{S}{\geq}{}^i)_j\ g$ iff $f' \overset{S}{\geq}{}^i g'$ for all f', g' such that $f'(j) = f(j)$ and $g'(j) = g(j)$, and f and g coincide with each other outside j.

We need to connect subjective with actual preferences in the same way as we previously connected extended with actual preferences. Hence the following version of Consideration Of Others (CO′):

CO′ for all i and j, and all f and g in L, $f\ (\overset{S}{\geq}{}^i)_j\ g$ iff $f \geq^j g$.

In the present variant, Consideration of Others has become a direct implication of the Strong Pareto principle of social choice theory. It says that if two social states differ only by the outcomes of some individual, one is subjectively preferred to the other if and only if it is preferred by that individual.[25]

Given CO′ our previous VNM assumption guarantees that the actual preference relations $\geq^i$ also satisfy the VNM axioms. The further condition we impose on the $\geq^i$ is the mild one that they are nontrivial:

NT for all i, there exist f_i, g_i in L such that $f_i >^i g_i$

(where as usual, $>^i$ is the strict preference relation derived from $\geq^i$).

We still need a notion of extended preferences. The relations $\overset{E}{\geq}{}^i$ are now defined on the set Δ' of all lotteries on extended alternatives of the form (x_j, j) where $x_j \in X_j$ and $j \in N$. As before, extended preferences will be assumed to satisfy VNM. Conceptually, they serve a different purpose than before. The $\overset{E}{\geq}{}^i$ relations now provide a quantitative benchmark for identifying the subjective probability measures p_i that underlie the $\overset{S}{\geq}{}^i$. The idea of comparing preference assessments of lotteries whose probability values are *numerically given* with preference assessments of prospects which do *not* involve preassigned probability numbers constitutes the distinctive feature of the Anscombe–Aumann approach to subjective probability (as opposed, say, to Savage's). We implement this heuristics by connecting $\overset{S}{\geq}{}^i$ with $\overset{E}{\geq}{}^i$ in the Probabil-

[25] Given the equivalence mentioned in this statement, preference can be understood either weakly or strongly.

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

istic Consistency (PC) axiom below.[26] Essentially, it says that subjective preferences conditional on j are identical with extended preferences conditional on the same individual j, where the latter are defined in the natural way (i.e., by disregarding what concerns the individuals other than j).

More formally, we define *conditional extended preferences* from $\overset{E}{\geq} i$ as follows. For all i and j, and all l and l' in Δ',

l $(\overset{E}{\geq} i)_j$ l' iff λ $\overset{E}{\geq} i$ λ' for all λ, λ' such that $\lambda(x_j, j) = l(x_j, j)$, $\lambda'(x_j, j) = l'(x_j, j)$ for all x_j, and λ and λ' coincide with each other outside the (x_j, j) values. Then, Probabilistic Consistency states that:

PC for all i and j, and all l, l' in Δ' such that the marginal probability distributions $\Sigma_{xj}\, l(x_j, .)$ and $\Sigma_{xj}\, l'(x_j, .)$ have full support,

$$l(\overset{E}{\geq} i)_j l' \text{ iff } H(l)\ (\overset{S}{\geq} i)_j\ H(l'),$$

where $H(l) = l(., j)\ /\ \Sigma_{xj}\, l(x_j, j)$ and $H(l') = l'(., j)\ /\ \Sigma_{xj}\, l'(x_j, j)$. The role of the H mapping here is to turn the $l(., j)$ and $l'(., j)$ functions into probability measures by normalizing them. Since the two conditional preferences $(\overset{E}{\geq} i)_j$ and $(\overset{S}{\geq} i)_j$ are not defined on the same mathematical objects, we need this transformation in order to relate them to each other.

Axiom PC says in words that whenever attention is restricted to individual j's personalized outcomes, it does not matter whether the observer assesses the chances of getting these outcomes in terms of his extended preferences or of his subjective preferences. When the observer disregards the possibility of being anyone other than j, the probability numbers assigned by extended lotteries to the different individuals do not matter any more, so there is no reason to expect comparisons based on $\overset{E}{\geq} i$ and $\overset{S}{\geq} i$ to be different. This is roughly the normative basis of the PC axiom.

We can now state:

The Impartial Observer Theorem (Subjective Probability Version): Assume that for each individual i, the subjective preference relation and the extended preference relation satisfy VNM, and the actual preference relation satisfies NT. Assume that SP, CO', and PC hold. Then, for each i, there exist a representation w_i of his moral preferences, a representation v_i of his extended preferences, representations u_{i1} (on X_1), ..., u_{in} (*on* X_n) of the individuals' actual preferences, and a (full support) subjective probability measure p_i on N, such that for all social states f, and all extended lotteries l:

[26] This axiom originates in an early construction due to Karni and Schmeidler (1981). Karni and Mongin (2000) have recently revived it. The interested reader is referred to the latter paper for further philosophical motivations as well as technical details. See also Schervish, Seidenfeld and Kadane (1991).

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

(#) $w_i(f) = \Sigma_{j\in N}\ \Sigma_{xj\in Xj}\ p_i(j)\ u_{ij}(x_j)\ f(x_j, j)$

and

(##) $b_i(l) = \Sigma_{j\in N}\ \Sigma_{xj\in Xj}\ u_{ij}(x_j)\ l(x_j, j)$

Any alternative set of functions $w_i^\circ, v_i^\circ, u_{i1}^\circ, \ldots, u_{in}^\circ,\ p_i^\circ$ satisfying properties (#) and (##) must be such that for some a_i>0 and $(b_{ij})_{j\in N}$

$(u_{ij}^\circ)_{j\in \mathrm{N}} = a_i(u_{ij})_{j\in N} + (b_{ij})_{j\in N}$,

$p_i^\circ = p_i$,

and w_i° and v_i° are positive affine transforms of w_i and v_i with a_i being the multiplicative factor of the transformations. (For a sketch of the proof, see the appendix.)

Compare the utility representation derived for i's moral preference with that of the 'Many Impartial Observers' Theorem of Section 5, taking account of the uniqueness properties stated here. The added value of the present variant is clear. First, we have avoided EC by embodying a subjective probability scheme into the ethical construction. Recall the overall motivation: impartiality is to be interpreted epistemically. When this is understood, EC appears to be doubtful, and gives way to SP and PC. As to CO', it is, like CO, a purely ethical axiom. Notice that PC and CO' together imply CO. The conclusion now is that the ethical observer i assesses social states in terms of some probability measure p_i. Following the Bayesian tenet, it does not matter what values p_i takes on. The only important thing is that this probability is uniquely identifiable from j's two sets of related preferences, as is spelled out in the uniqueness part of the result.

Second, there is is a welcome by-product of the construction. Again because of the uniqueness part, it overcomes the arbitrariness in the choice of the u_{ij} representations that plagued the previous versions of the Impartial Observer Theorem. The functions u_{ij} mentioned in the present statement are nearly uniquely (i.e., up to common positive affine transforms) those which 'reveal' the subjective probability p_i. They cannot be rescaled arbitrarily without destroying this subjective probability, and thus the possibility of a Bayesian interpretation of the individual's two preferences. The uniqueness of p_i also means that the individual weights in the utilitarian looking formula for w_i are fixed once and for all. Whether they are equal or not depends on the observer's subjective beliefs, as summarized by p_i. Thus, we have an answer to the earlier objection (at the end of Section 4) that the choice of VNM representations, or equivalently of individual weights, was an arbitrary one. *Weights are typically unequal, but there are reasons for that*.

Returning to the point made at the beginning of this section, it is common among Harsanyi's interpreters to match sympathy with a

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

version of CO, and impartiality with a version of EC. A reconstruction along this line trades on the following notion of impartiality: to be impartial is to treat the individuals equally. This is a normative, not an epistemic notion of impartiality. It is coarse but commonsensical enough. The reader would misunderstand us if he believed that by promoting an epistemic reading of impartiality, we are dismissing impartiality as equal treatment. Rather, we are dispensing with it. The trouble with axiom EC interpreted normatively is not that it is wrong – just that *it begs too much of the conclusion of the 'theorem'*. We end up with a reasoning that sounds hardly like an argument for utilitarianism. We already knew that the VNM assumptions went a long way towards begging the linear form of the rule. We now see that EC begs the equal weights in the sum. True, there remains at least one non-obvious step in the reasoning, which is to base moral judgements on extended comparisons and CO. But extended comparisons can only be justified in terms of ignorance. The ordinary interpretation is saddled with the task of defending its own ignorance scenario and making it plausible that ignorance and impartiality (normatively understood) can be assumed for the same observer.

In sum, it seems both more rewarding and more consistent to derive as much as possible from defining impartiality in terms of ignorance. Once this line is taken, it quite naturally leads to unequal weights, and a form of inequality appears to be justified in retrospect. The conclusion may strike one as surprising or even unpalatable. At least, we do have an *argument* now.

The Subjective Probability Version also runs counter to the desired result that moral assessments are uniform. For the sake of theoretical experimentation, we may superimpose UEP on the other axioms. The result would be a set of representations of moral preferences of the form:

$$w_i(f) = \Sigma_{j\in N}\Sigma_{x_j\in X_j} p_i(j)\ u'_j(x_j) f(x_j, j),$$

where p_i is i's uniquely defined subjective probability and $u'_1, \ldots, u'_n$ are essentially unique utility representations of the individuals' actual preferences that do not depend on i. This would be a 'Many Impartial Observers' formula again, the reason now being that the many observers' subjective probabilities, not their views of the others' utilities, are diverse.

There is an alternative plausible departure from EC that we may eventually consider. It starts by arguing that *objective frequencies* are the probabilistic data to rely on. This line takes us outside the confines of standard Bayesianism. However, there is also a minority school of 'objective' Bayesianism, and such a label will perhaps not be unsuitable for Harsanyi, given his double commitment to Bayesianism and an objective construal of extended preference. According to the causal account, the observer's state of ignorance is far from complete. The

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

observer does not know his own preference features but is endowed with all the required nomological knowledge about preferences. This knowledge bears on the causal factors, like tastes or wealth, that influence preferences. It is a further step, but one which is in line with the causal account, to assume that the observer also knows something about the empirical frequencies of the prevailing factors. These should induce the probability values to be taken into account in the expected utility formula. One can guess what the Impartial Observer Theorem is that corresponds to this alternative assumption. If (UEP) holds, it will lead to a utilitarian looking social rule independent from the particular observer, but in which individuals have generally non-uniform weights:

$$w(x) = \Sigma_{j \in N}\ \alpha_j\ u'_j(x)$$

In this formula α_j is the (objective) probability that the causal factors determining *j*'s preferences obtain.

There is no reason to expect the objective probability values α_j to be the same for each *j*. So once again, equal treatment does not emerge from the analysis, and impartiality viewed epistemically diverges from impartiality as it is more commonly envisaged. If we leave things at this unelaborate stage, the earlier objection that the u'_j are arbitrary will destructively apply. To supersede it, we should embody 'objective Bayesianism' into a 'revelation' subjective probability scheme in the style of the previous one. Conceivably, this can be done, although we do not know of any axiomatic work proceeding exactly along these lines.

7. WHAT IS LEFT OF THE IMPARTIAL OBSERVER THEOREM?

The Impartial Observer Theorem is surrounded with conceptual difficulties. They can be put in the summary form of a succession of dilemmas. Either the Impartial Observer Theorem is stated in Vickrey's limited form, and it has no ethical interest. Or it is stated in Harsanyi's more relevant form, and then it involves the difficulties of extended preferences. One version of extended preferences, which is not Harsanyi's, requires one's identifying with the others, and leads to the puzzling difficulties of personal identity. These difficulties may or may not be insuperable, but the version under discussion cannot ensure the identity of extended preferences. In the other account, which is Harsanyi's, and which we called the causal account, one must be careful to decide whether extended preference comparisons are preference comparisons in the precise sense, or whether they are just a name for direct comparisons of utility amounts by an observer. In the literal interpretation in terms of preference, Harsanyi's causal account is worthless as

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

regards delivering the conclusion that extended comparisons are uniform. In the non-literal interpretation in terms of utility, the causal account may lead to a claim that extended comparisons are uniform, but comparisons are now of utility amounts, and this is not what the Impartial Observer Theorem is concerned with.

The probabilistic discussion of the Impartial Observer Theorem reinforces the negative conclusion that a genuinely utilitarian formula, that is, with equal weights and no observer-dependence, is out of reach, given the chosen primitives and premisses. In our main probabilistic model, observer-dependence follows from the observer's implicit reliance on a subjective probability and the weights are unequal because they are determined by this subjective probability. In another probabilistic model which was only sketched, observer-dependence is still threatening on the utility side, and the weights are unequal, this time, for objective reasons. The first model is that of standard Bayesianism; the second model corresponds to an unelaborated theoretical possibility, namely, 'objective Bayesianism'. All and all, the probabilistic analysis confirms Rawls's initial intuition that 'there seems to be no objective grounds in the initial situation for assuming that one has an equal chance of turning out to be anybody. That is, this assumption is not founded upon known features of one's society' (1971, p. 168).

Given this list of negative results, the argument can take two different directions – and, it would seem, only two. The first line consists in taking the 'theorem' to be about what it is overtly, namely, preferences, and attempts to live with the fact that preference judgements are by their very nature subjective, hence non-uniform. Then comes an interesting argument due to Pattanaik (1968). By a reasoning of his own,[27] he comes to the conclusion that the 'theorem' could only lead to observer-dependent social rules. Pattanaik's version appears to coincide with the 'Many Impartial Observers' Theorem (Section 5). At this juncture, he suggests applying Harsanyi's *other* theorem, that is, the 1955 Aggregation Theorem, to the many moral utility functions, in order to construct the society's moral utility function. With the benefit of hindsight, Pattanaik's suggestion strikes one as unconvincing. Since the Pareto principle is the driving force in Harsanyi's other theorem, it means that the comparison between diverse moral rules within the society is settled by unanimity considerations. There are philosophical reasons for doubting that the Pareto principle should apply to moral judgements in this highly direct way. Besides, there are technical difficulties with Pattanaik's resolution as soon as subjective probability enters the framework of the Impartial

[27] As mentioned in Section 3, Pattanaik (1968) stresses the diversity of risk attitudes, hence of VNM extended preferences, across individual observers. So his reasoning is identical neither with Broome's (1993) nor with ours.

Observer Theorem. In the Subjective Probability Version (Section 6), we have obtained moral expected utility functions which can differ from one observer to another *both* in terms of these observers' utility functions and probability measures. But it has been demonstrated that Harsanyi's Aggregation Theorem does not extend to the case of differing subjective probabilities, even for weak forms of the Pareto principle.[28] So Pattanaik's solution is blocked in this relevant version of the 'theorem' – a difficulty that he could not foreshadow at the time of his comment.

The lesson of all this is that if the primitives are taken to be preferences, there is no way out of observer-dependence. This is a disappointing conclusion, but we should emphasize that along the present lines, at least something can be *proved*. The Subjective Probability Version of the Impartial Observer Theorem is neither entirely trivial nor void of ethical content.

The second line of analysis is to draw the ultimate consequences of Harsanyi's causal conception of interpersonal comparisons. We said that the causal conception provided him with a reason to claim that extended utility assessments are uniform across observers. Harsanyi's mistake was to confound this statement with the non-equivalent one that extended preferences are uniform. If the utility numbers are indicative of *well-being*, one gets a roughly plausible idea of both extended judgements and why they are uniform. One may wonder, however, if the present line of analysis can deliver a *theorem* in the technical sense of a formal proof. In the formalism, the common numerical assessments, rather than the observers' extended preference relations, would be the primitives. Then, the 'theorem' could consist only in applying the expected utility formula directly to these utility assessements, given some conceptually defensible choice of probability values; objective frequencies would typically be resorted to. The final result would be an observer-independent generalized utilitarian rule (i.e., with unequal weights). However, the argument leading to it can only be philosophical, not mathematical. We now have an observer who tries to assess social states in terms of the chances of well-being that each of these states implies for him. Granting that well-being can be quantified in the first place, a case must be made for using objective frequencies as the relevant probability concept to model this observer's ignorance. And a further argument is called for to defend one's use of the expected utility formula, since it is not possible to invoke the (preference-based!) VNM representation theorem. The second line of analysis does not deliver an Impartial Observer Theorem in the genuine sense of the word 'theorem'.

[28] For the finite state space of this paper, see Seidenfeld, Schervish and Kadane (1989) and Mongin (1998).

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

APPENDIX

Vickrey's version of the Impartial Observer Theorem is mathematically trivial, and so is Harsanyi's, but we give a formal proof of the latter to highlight the role of the various assumptions.

Proof of Harsanyi's Impartial Observer Theorem. Take a VNM representation v_1 of $\mathrm{E} \geq^i$. For each i, CO implies that the restriction $v_1(., i)$ to X is a utility representation of $\geq^i$ on X. We put $u'_i(.) = v_1(., i)$ for $i = 1, \ldots, n$. From UEP v_1 can also serve as a VNM representation of $\mathrm{E} \geq^i$ for $i = 2, \ldots, n$. That is to say:

(#) for $i = 1, \ldots, n,\ v_1(\mathrm{L}_x) \geq v_i(\mathrm{L}_y)$ iff $\mathrm{L}_x \,\mathrm{E} \geq^i \mathrm{L}_y$

hence from EC:

for $i = 1, \ldots, n,\ v_1(\mathrm{L}_x) \geq v_1(\mathrm{L}_y)$ iff $x \,^*\!\geq^i y$.

Putting $v_1(\mathrm{L}_x) = w(x)$ for all x, we obtain the common representation of moral preferences w. By the VNM property of v_1, it satisfies that:

$$v_1(\mathrm{L}_x) = w(x) = 1/n\, v_1(x, 1) + \ldots + 1/n\, v_1(x, n) = 1/n\, u'_1(x) + \ldots + 1/n\, u'_n(x),$$

as required.

If UEP is not part of the assumptions, step (#) is not permissible, and the proof must be changed as follows. For each i and each j, CO implies that the restriction $v_i(., j)$ to X is a utility representation of $\geq^j$ on X. We put $u'_{ij}(.) = v_i(., j)$ for $i, j = 1, \ldots, n$. Applying EC to each v_i separately we get an i-dependent representation of moral preferences:

$$v_i(\mathrm{L}_x) = w_i(x) = 1/n\, u'_{i1}(x) + \ldots + 1/n\, u'_{in}(x).$$

This is the 'Many Impartial Observers' Theorem in Section 5.

In Weymark (1991, Theorem 9) and Karni and Weymark (1998), a more assertive conclusion follows from stronger premisses. There, X is defined to be a *lottery* set, and the individual preference relations are assumed to be VNM, with the result that the u'_i in the additive representation above are VNM functions. This ingredient is unnecessary to the formalization as long as the latter does not aim at highlighting the cardinality issues underlying the choice of the u'_i. Notice that the 'theorem' could receive an even more economical formalization than the present one. By the VNM the extended preferences are defined and satisfy the VNM axioms on the set of *all* extended lotteries. This is too strong an assumption given the actual use made of the VNM representation theorem in the proof. Restricted domains may do, as Karni and Weymark (1998) have made clear in their own framework.

It is clear that the proof above involves an element of arbitrariness, that is, the choice of the VNM representation $v_1(., i)$ to represent all the

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

relevant preference relations. Instead of putting $u'_i(.) = v_1(., i)$ for $i = 1, \ldots, n$, we could put $u'_i(.) = f_i \; o \; v_1(., i)$ for $i = 1, \ldots, n$, for any choice of strictly increasing functions, and the axioms would still be satisfied. The result would not be an additive representation, but only the following *additively separable* representation:

$$w(x) = (f_1)^{-1}(u'_1) = (x)) + \ldots + (f_n)^{-1}(u'_n(x)).$$

The arbitrariness involved here is the essence of what we called the Sen–Weymark objection in Section 4.

Here is a sketch of a proof for the Subjective Probability Variant of Section 6. As a preliminary fact, observe that the set L of social states is convex because of the usual definitions of the sum of functions and the multiplication of a function by a scalar, in terms of point values.

Sketch of the proof. Applying a variant of the VNM theorem due to Fishburn (1970, Theorem 13.1), $\overset{S}{\geq}{}^{i}$ on L can be represented by:

(1) $\Sigma_{j \in N} \; u^*_{ij}(f(j)) = \Sigma_{j \in N} \Sigma_{x_j \in X_j} \; u^*_{ij}(x_j) f(x_j, j)$,

where the u^*_{ij} are VNM representations unique up to positive affine transformations involving a *common* multiplicative factor.

It follows that for all j, $u^*_{ij}(f(j)$ represents $(\overset{S}{\geq}{}^{i})_j$ and that this function is constant if and only if $(\overset{S}{\geq}{}^{i})_j$ is trivial. From PC, CO′ and NT this possibility is excluded.

A direct application of the VNM theorem to $\overset{E}{\geq}{}^{i}$ on Δ' entails that this preference relation can be represented by:

(2) $\Sigma_{j \in N} \Sigma_{x_j \in X_j} \; u_i(x_j, j) \; l(x_j, j)$,

where u_i is unique up to a positive affine transformation.

This implies that for each j, $\Sigma_{x_j \in X_j} \; u_i(x_j, j) \; l(x_j, j)$ is a VNM representation of $(\overset{E}{\geq}{}^{i})_j$.

Now, apply PC to the VNM representations of $(\overset{S}{\geq}{}^{i})_j$ and $(\overset{E}{\geq}{}^{i})_j$, respectively, that have just been obtained. It follows that for each j, there are numbers $a_{ij} > 0$ and b_{ij} such that for all $x_j \in X_j$

$$u^*_{ij}(x_j) = a_{ij} \; u_i(x_j, j) + b_{ij}.$$

Renormalize the $u^*_{ij}(x_j)$ to set $b_{ij} = 0$ for all j. Put $p_i(j) = a_{ij} / \Sigma_{j \in N} \; a_{ij}$ and $u_{ij}(x_j) = u_i(x_j, j) \; (\Sigma_{j \in N} \; a_{ij})$. Then, given the uniqueness conditions for (1),

(#) $w_i(f) = \Sigma_{j \in N} \Sigma_{x_j \in X_j} \; p_i(j) \; u_{ij}(x_j) \, f(x_j, j)$

represents $\overset{S}{\geq}{}^{i}$, and given the uniqueness condition for (2),

(##) $v_i(l) = \Sigma_{j \in N} \Sigma_{x_j \in X_j} \; u_{ij}(x_j) \; l(x_j, j)$

represents $\overset{E}{\geq}{}^{i}$.

https://doi.org/10.1017/S0266267101000219 Published online by Cambridge University Press

As to the uniqueness part of the Subjective Probability variant, it follows from the uniqueness conditions for (1) and (2).[29]

REFERENCES

Anscombe, F. J. and R. J. Aumann. 1963. 'A definition of subjective probability'. *Annals of Mathematical Statistics*, 34:199–205

Arrow, K. J. [1951] 1963. *Social Choice and Individual Values*. Yale University Press

Arrow, K. J. 1977. 'Extended sympathy and the possibility of social choice'. *American Economic Review, Papers and Proceedings*, 67:219–29. Reprinted in K. J. Arrow, *Collected Works, 1: Social Choice and Justice*, Harvard University Press, ch. 11

d'Aspremont, C. et L. A. Gérard-Varet. 1991. 'Utilitarian fundamentalism and limited information'. In *Interpersonal Comparisons of Well-Being*, pp. 371–85. J. Elster and J. E. Roemer (eds.). Cambridge University Press

Broome, J. 1993. 'A cause of preference is not an object of preference'. *Social Choice and Welfare*, 10:57–68

Fishburn, P. C. 1970. *Utility Theory for Decision Making*. Wiley

Fontaine, P. 1997. 'Identification and economic behavior. Sympathy and empathy in historical perspective'. *Economics and Philosophy*, 13:261–80

Hampton, J. 1980. 'Contracts and choices: does Rawls have a social contract theory?' *Journal of Philosophy*, 77:315–38

Harsanyi, J. C. 1953. 'Cardinal utility in welfare economics and in the theory of risk-taking'. *Journal of Political Economy*, 61:434–35. Reprinted in J. C. Harsanyi (1976)

Harsanyi, J. C. 1955. 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility'. *Journal of Political Economy*, 63:309–21. Reprinted in J. C. Harsanyi (1976)

Harsanyi, J. C. 1958. 'Ethics in terms of hypothetical imperatives'. *Mind*, 67:305–16. Reprinted in J. C. Harsanyi (1976)

Harsanyi, J. C. 1967–1968. 'Games with incomplete information played by ''Bayesian'' players', I–II. *Management Science*, 14:159–82, 320–34 and 486–502

Harsanyi, J. C. 1976. *Essays on Ethics, Social Behavior, and Scientific Explanation*. P. Reidel

Harsanyi, J. C. 1977a. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press

Harvey, C. M. 1999. 'Aggregation of individuals' preference intensities into social preference intensity'. *Social Choice and Welfare*, 16:65–79

Jeffrey, R. 1983. *The Logic of Decision*. University of Chicago Press, 2nd edn

Kaneko, M. 1984. 'On interpersonal utility comparisons'. *Social Choice and Welfare*, 1:165–75

Karni, E. 1998. 'Impartiality: definition and representation'. *Econometrica*, 66:1405–15

Karni, E. and P. Mongin 2000. 'On the determination of subjective probability by choices'. *Management Science*, 46:233–48

Karni, E. and D. Schmeidler 1981. 'An expected-utility theory for state-dependent preferences'. Working Paper 48–80. The Foerder Institute of Economic Research, Tel Aviv University

Karni, E. and J. Weymark 1998. 'An informationally parsimonious impartial observer theorem'. *Social Choice and Welfare*, 15:321–32

[29] It is instructive to compare the present variant with Karni's (1998) subjective probability construction. Like most of Harsanyi's followers, Karni considers a single impartial observer. However, the major difference is that Karni's variant includes a special axiom to the effect that the observer will give equal probability to each individual, so that the social rule will have equal weights. Karni's own interpretation is on the normative rather than the epistemic side, but one could also see his axiom as a preference analogue in the subjective probability framework of Laplace's principle.

Kolm, S. C. 1972. *Justice et équité*. Editions du Centre National de la Recherche Scientifique. English trans. *Justice and Equity*. M.I.T. Press

Kolm, S. C. 1994. 'The meaning of fundamental preferences'. *Social Choice and Welfare*, 11:193–98

Levi, I. 1977. 'Four types of ignorance'. *Social Research*, 44:745–56

McKay, A. F. 1986. 'Extended sympathy and interpersonal utility comparisons'. *Journal of Philosophy*, 83:305–22

Mongin, P. 1994. 'Harsanyi's aggregation theorem: multi-profile version and unsettled questions'. *Social Choice and Welfare*, 11:331–54

Mongin, P. 1998. 'The paradox of the Bayesian experts and state-dependent utility theory'. *Journal of Mathematical Economics*, 29:331–61

Mongin, P. and C. d'Aspremont. 1998. 'Utility theory and ethics'. In *Handbook of Utility Theory*. S. Barberà, P. Hammond and C. Seidl (eds.). Kluwer

Pattanaik, P. K. 1968. 'Risk, impersonality, and the social welfare function'. *Journal of Political Economy*, 76:1152–69

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press

Rothenberg, J. 1961. *The Measurement of Social Welfare*. Prentice-Hall

Savage, L. J. 1954. *The Foundations of Statistics*. Dover

Schervish, M. J., T. Seidenfeld and J. B. Kadane. 1991. 'Shared preferences and state-dependent utilities'. *Management Science*, 37:1575–89

Seidenfeld, T., M. J. Schervish. and J. B. Kadane. 1989. 'On the shared preferences of two decision makers'. *Journal of Philosophy*, 86:225–44

Sen, A. 1970. *Collective Choice and Social Welfare*. North Holland

Sen, A. 1986. 'Social choice theory'. In *Handbook of Mathematical Economics*, III, pp. 1073–81. K. J. Arrow and M. D. Intriltigator (eds.). North Holland

Suppes, P. 1966. 'Some formal models of grading principles'. *Synthese*, 6:284–306

Suzumura, K. 1983. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge University Press

Vickrey, W. S. 1945. 'Measuring marginal utility by reaction to risk'. *Econometrica*, 13:319–33

Vickrey, W. S. 1960. 'Utility, strategy, and social decision rules'. *Quarterly Journal of Economics*, 74:507–35

Vickrey, W. S. 1994. *Public Economics. Selected Papers by William Vickrey*. R. Arnott, K. Arrow, A. Atkinson and J. Drèze (eds.). Cambridge University Press

Weymark, J. 1991. 'A reconsideration of the Harsanyi–Sen debate on utilitarianism'. In *Interpersonal Comparisons of Well-Being*, pp. 255–320. J. Elster and J. E. Roemer (eds). Cambridge University Press