Department of Social Systems and Management

Discussion Paper Series

No. 1186

A Critical Consideration of Harsanyi's Preference Utilitarianism:

The Double-Profile Approach

By
Mamoru KANEKO and Henri NZITAT

November 2007

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

A Critical Consideration of Harsanyi's Preference Utilitarianism: The Double-Profile Approach*

Mamoru Kaneko† and Henri Nzitat‡

10 November 2007

Abstract

This paper examines Harsanyi's preference utilitarianism by giving a double-profile approach. In this approach, each individual has two kinds of preference relations: personal private preferences and personal ethical preferences. Moreover, the social ethnical observer has its own social preferences. We assume that those preference relations all satisfy the von Neumann-Morgenstern axioms. Using Harsanyi's axioms, personal private preferences are aggregated into personal ethical preferences, and then personal ethical preferences are aggregated into the social preferences. We give certain axioms on those aggregation steps to explain the difference between the two steps. We construct this approach so that we can study how interpersonal utility comparisons are involved in those steps.

1. Introduction

Harsanyi's utilitarianism, proposed in Harsanyi [10], [11], [12], [13] and some other papers, has been controversial in various respects. Roughly speaking, it claims that a welfare judgement should be based on the welfare function defined as the weighted sum of individual utility values:

$$w(p) = \sum_i \alpha_i u_i(p). \tag{1.1}$$

In this paper, we will provide a double-profile approach in order to facilitate critical examinations of some controversial respects. In this section, we explain our approach and results, while looking at some discussions in the literature of Harsanyi's utilitarianism.

*The authors thank Adrian Miroiu, Ryuichiro Ishikawa and Tai-wei Hu for comments on earlier drafts of this paper. The first author is partially supported by Grant-in-Aids for Scientific Research No.18330034, Ministry of Education, Science and Culture.

†Institute of Policy and Planning Sciences, University of Tsukuba, Ibaraki 305-8573, Japan (kaneko@shako.sk.tsukuba.ac.jp)

‡Ecole Nationale Supérieure d'Ingenieurs de Caen & Centre de Recherche, Caen, France (nzitat@aol.com)

The main mathematical theorem of Harsanyi's utilitarianism was given in [11]. In addition to the von Neumann-Morgenstern expected utility axioms, he gave one additional axiom, called the *Pareto Indifference*, and claimed that personal utilities expressed by u_i's and ethical (or social) utility expressed by w are connected in the form (1.1). This is often called the *single-profile approach.* His own proof had a flaw, but it was later corrected by Domotor [6].

The Harsanyi-Domotor theorem is simple and suggests a possibility of the existence of a welfare judgement (a welfare function), opposing to Arrow's impossibility theorem. In this sense, it has attracted some people's attention. Nevertheless, even though the proof was corrected, Harsanyi's utilitarianism itself has still remained conceptually unclear and its status in the present welfare economics is not well settled.

Besides the mathematical flaw corrected by Domotor, unclear points and difficulties are related to

(1) the relationship between an individual and society;

(2) interpersonal comparability of utilities.

Harsanyi gave certain arguments for these points. However, he himself had slightly changed his attitude for (1) in his later papers, but had stuck to one justification for (2) rather than examined carefully its possibility. In the Harsanyi-Domotor Theorem, it is unclear where interpersonal comparability of utilities is involved (Mongin [22], p.349). In the following, we explain our double-profile approach by reflecting on these points.

Harsanyi's single-profile approach is expressed as an $(n+1)$-tuple of preference relations $\langle \preceq; \preceq_1, ..., \preceq_n \rangle$. The Harsanyi-Domotor Theorem states that if those relations satisfy the von Neumann-Morgenstern utility axioms and the Axiom of Pareto indifference, then any utility representations $w, u_1, ..., u_n$ of $\preceq, \preceq_1, ..., \preceq_n$ are connected in the form (1.1).

In Harsanyi [10] and [11], $\preceq_1, ..., \preceq_n$ are regarded as personal preferences and $\preceq$ is social preferences. In those papers, $\preceq$ could be regarded also as personal ethical preferences, since in his position of those papers, every individual reached the same personal ethical preferences $\preceq$, which becomes social preferences. Here he used the famous "veil of ignorance" argument[1] as well as the argument for interpersonal comparability of utilities to support the same personal ethical preferences. Later, e.g., in Harsanyi [12] and [13], he seemed to retreat from claiming $\preceq$ as social preferences and to regard it only as personal ethical preferences. The reason might come from forcible criticisms given by several authors. Here, we refer to Pattanaik [26] and Diamond [5].

Pattanaik [26] suggested to separate these preferences into three levels:

(a): personal private preferences;

[1] The "veil of ignorance" argument is more famous with Rawls [28], but, more or less, the same form was already given in Harsanyi [10]. However, Rawls claimed to reach the maxmin welfare criterion, while Harsanyi did his utilitarianism.

(b): personal ethical preferences;

(c): social preferences.

Personal ethical preferences (b) are obtained by aggregating personal private preferences (a). Pattanaik suggested two other ways of aggregations: from (a) to (c) and from (b) to (c). In [12] and [13], Harsanyi admitted and adopted this separation, though he had stuck to the consideration of (a) and (b). With this separation, we need also to assume who aggregate preferences (a) to (c) and (b) to (c). In this paper, we will assume that the *hypothetical outside observer* aggregates (b) to (c). We assume that a hypothetical outside observer is intended to be "impersonal" and "impartial"[2]. We take these as meaning that the hypothetical outside observer has complete knowledge about personal ethical preferences (a). We will give a comment on the aggregation from (a) to (c) in Section 7.

Now, we meet one epistemological problem. If a personal individual is also impersonal and impartial in the sense that he knows (a) objectively as like the hypothetical outside observer, the above separation would become effectively nothing. Here, we assume that a personal individual aggregating (a) to (b) is an ordinary person in society, and have very limited and even false knowledge, or may take only some people into account. This sounds to deviate from an "ethical judgement". However, as soon as (b) is separated from (c), these aspects should be inevitably taken into account. We would like to provide a mathematical framework so as to study the difference between (b) and (c).

We provide the above three levels of preferences:

$$D = \left[\preceq^0 \mid \langle \preceq^i ; \preceq_{i1}, ..., \preceq_{in} \rangle : i = 1, ..., n \right], \tag{1.2}$$

where $\preceq_{ij}$ is personal private preferences of individual j believed by individual i, and $\preceq^i$ is i's ethical preferences. Finally, $\preceq^0$ is social preferences. We allow $\preceq_{ij}$ to be different from j's preferences $\preceq_{jj}$ in j's mind, even which may differ from the true preferences observed by the hypothetical observer. In this paper, we do not discuss this problem; but, instead, we consider aggregations from (a): $\preceq_{i1}, ..., \preceq_{in}$ to (b): $\preceq^i$ and from (b): $\preceq^1, ..., \preceq^n$ to (c): $\preceq^0$. In Section 7, we will comment on an aggregation from (a) to (c) as well as on the truthfulness of $\preceq_{jj}$.

The aggregation step from (a) to (b) is, more or less, the same as the Harsanyi-Domotor Theorem. However, to emphasize differences from the aggregation step from (b) to (c), we will discuss the step from (a) to (b) in Section 4. Here, we take the position that when $\preceq^i$ is assumed to be obtained from $\preceq_{i1}, ..., \preceq_{in}$, individual i makes

[2] Rabinowicz-Österberg [27] suggested the interpretation that personal ethical preferences are constructed by a personal individual and social preferences are made by the outside observer. They discussed this distinction from a philosophical point of view.

interpersonal comparisons of utilities in the mind of i. We treat this rather as a psychological, maybe, anthropological problem. From the objective ethical-moral point of view, this is rather arbitrary and depends upon individual i.

The hypothetical outside observer having all information about $\langle \preceq^1, ..., \preceq^n \rangle$ aggregates them into social preferences $\preceq^0$. In this case, we have the problem of interpersonal comparability of utilities in the objective and scientific point of view. Even though we assume the existence of a hypothetical outside observer knowing all about $\langle \preceq^1, ..., \preceq^n \rangle$, the possibility of such comparisons is a different problem (see Kaneko [16] and Broome [2] for arguments against such comparisons). We avoid this problem, but we see where such interpersonal comparisons of utilities are involved in the aggregation from (b) to (c) and consider how they are formulated in the double-profile approach.

The last comment is on the use of von Neumann-Morgenstern expected utility axioms for those three levels of preferences. In our approach, the levels (a) and (b) may not be problematic, since those preferences are individual preferences (cf., Karni [20]). However, following Diamond's [5], it may be problematic in assuming expected utility axioms for social preferences (c). Nevertheless, we take a research strategy to examine the consequences under the expected utility axioms even for (c). In Section 7, we will discuss Diamond's example and also look at the counter argument to it by Nakamura-Nakayama [23].[3]

Technically speaking, our results are based on Domotor [6] and Weymark [33], and some results are almost immediately obtained by combining results in these papers and some other papers. This paper is intended not to give new technical results but to offer a framework to reconsider the entire Harsanyi's preference utilitarianism.

The paper is written as follows: In Section 2.1, we give a quick review of expected utility theory. This is because of the observation that some difficulties in Harsanyi's utilitarianism are caused by starting with utility functions rather than with the very primitive concepts of preferences. In Section 2.2, we review Harsanyi's single-profile approach. In Section 3, we will give the double-profile approach. In Section 4, we consider the aggregation from personal preferences to personal ethical preferences. In Section 5, we consider the aggregation from personal ethical preferences to social preferences. In Section 7, we will reconsider our results comparing some papers in the literature.

2. Expected Utility Theory and Harsanyi's Single-Profile Approach

2.1. Expected Utility Theory

Let X be a finite nonempty set of pure alternatives, which will be interpreted as social pure alternatives in the main body of the paper. We sometimes write $X = \{x_1, ..., x_k\}$.

[3] In the theory of the Nash social welfare fuction given in Kaneko-Nakamura [17], expected utility theory is assumed at the individual level, but it is not at the social aggregated level.

Let $L(X)$ be the set of all probability distributions over X, often called *prospects*. The set $L(X)$ can be regarded as a subset of $R^{|X|}$; actually it is the simplex of $|X|-1$ dimensions. Restricting the linear structure of $R^{|X|}$ to $L(X)$, we regard $L(X)$ to be a convex set with respect to the convex operation transferred from $R^{|X|}$. We denote a convex combination by $\alpha p * (1-\alpha)q$, which is here interpreted as a *compound prospect* meaning that p occurs with probability α and q occurs with probability $1-\alpha$.

A von Neumann-Morgenstern [34] preference relation is a binary relation $\preceq$ on $L(X)$, and is assumed to satisfy the following three axioms of expected utility theory:

Axiom NM1 (Ordering): $\preceq$ is a complete preordering on $L(X)$.

The symmetric part and asymmetric part of $\preceq$ are now written as $\sim$ and $\prec$, respectively. That is, $p \sim q$ means that p and q are indifferent, and $p \prec q$ means that q is strictly preferred to p.

Axiom NM2 (Intermediate Value): If $p \preceq r \preceq q$, then there is an $\alpha \in [0,1]$ such that $r \sim \alpha p * (1-\alpha)q$.

Axiom NM3 (Independence): Let p, q be arbitrary prospects in $L(X)$.

(1): If $p \prec q$, then $\alpha p * (1-\alpha)r \prec \alpha q * (1-\alpha)r$ for all $\alpha \in (0,1)$ and $r \in L(X)$.

(2): If $p \sim q$, then $\alpha p * (1-\alpha)r \sim \alpha q * (1-\alpha)r$ for all $\alpha \in (0,1)$ and $r \in L(X)$.

The following is known as the main theorem of expected utility theory:

Theorem 2.1. (Expected Utility Theorem): A binary relation $\preceq$ satisfies Axioms NM1-NM3 if and only if there is a real-valued function u on $L(X)$ such that

(1): for all $p, q \in L(X)$, $u(p) \leq u(q)$ if and only if $p \preceq q$;

(2): for all $p, q \in L(X)$ and $\alpha \in [0,1]$, $u(\alpha p * (1-\alpha)q) = \alpha u(p) + (1-\alpha)u(q)$.

There are many axiomatizations of expected utility theory such as given in Herstein-Milnor [14] and Fishburn [7] (see Hammond [9] for a recent survey). The above is a variant of the axiomatizations given in [14] and [7]. A proof of the above theorem is found in Kaneko-Wooders [19]. For the results in the following sections, any axiomatization is fine as far as Theorem 2.1.(1) and (2) are guaranteed.

A real-valued function u satisfying (1) and (2) of Theorem 2.1 is called an *von Neumann-Morgenstern utility function* representing $\preceq$. Let $U(\preceq)$ be the set of all vN-M utility functions representing $\preceq$.

One consequence, relevant here, derived from (2) of Theorem 2.1 is that $u(p)$ can be written as

$$u(p) = \sum_{t=1}^{k} p_t u(x_t) \text{ for all } p \in L(X), \tag{2.1}$$

where we follow the convention that x_t is regarded as identical to the unit vector $\mathbf{e}^t = (0, ..., 0, 1, 0, ..., 0)$ with the t-th coordinate 1. Later, we may use the vector expression $\mathbf{u} = (u(x_1), ..., u(x_k))$. Then, (2.1) can be written as the inner product $p \cdot \mathbf{u}$.

We should mention another theorem in expected utility theory, which is known as the uniqueness theorem (cf., Herstein-Milnor [14] for its proof) and is crucial for Harsanyi's preference utilitarianism.

Theorem 2.2. (Uniqueness up to Positive Linear Transformations): vN-M utility functions u and v are in the same set $U(\preceq)$ if and only if there are scalars $\alpha > 0$ and β such that $u(p) = \alpha v(p) + \beta$ for all $p \in L(X)$.

2.2. Harsanyi's Single-Profile Approach.

Now we give a brief summary of Harsanyi's single-profile approach and mention the main theorem, which was originally given by Harsanyi [11] and was corrected by Domotor [6].

Let us consider a society consisting of individuals $1, ..., n$. The set of all individuals is denoted by $N = \{1, ..., n\}$. Let i be an individual. A *single-profile frame* is given as an $n+1$-tuple $\langle \preceq; \preceq_1, ..., \preceq_n \rangle$, each of which is assumed to satisfy Axioms NM1-NM3. In our double-profile approach, each $\preceq_i$ will be interpreted as either personal private preferences or personal ethical preferences, and correspondingly, $\preceq$ will be interpreted as either personal ethical preferences or social preferences.

Harsanyi's [11] Pareto Indifference is given as follows:

Axiom H (Pareto Indifferent): If $p \sim_i q$ for all $i \in N$, then $p \sim q$.

That is, if all individuals are indifferent between p and q, then p and q are also indifferent with respect to $\preceq$. Only under this axiom, we have the following theorem.

Theorem 2.3 (Harsanyi and Domotor): Assume Axiom H on $\langle \preceq; \preceq_1, ..., \preceq_n \rangle$. Take an arbitrary $u \in U(\preceq)$ and arbitrary $u_i \in U(\preceq_i)$ for $i = 1, ..., n$. Then, there are $\alpha_1, ..., \alpha_n$ and β such that

$$u(p) = \sum_i \alpha_i u_i(p) + \beta \text{ for all } p \in L(X). \tag{2.2}$$

In this theorem, $\alpha_1, ..., \alpha_n$ may not be uniquely determined, and also their signs may be positive, negative or even zero. However, the unique determination of and the positivities of $\alpha_1, ..., \alpha_n$ are important for our research as well as from the viewpoint of welfare economics.

Here we introduce one notion called a test-transition, which was implicitly used in Harsanyi [11] and later was more explicitly used (e.g., Fishburn [8], Weymark [33], d'Aspremont-Gevers [4]). Let $\langle \preceq; \preceq_1, ..., \preceq_n \rangle$ be a single-profile frame. We say that two prospects q and p form a *test-transition*, denoted by $\langle q \rhd_i p \rangle$, for individual i iff $q \prec_i p$ and $q \sim_j p$ for all $j \in N - \{i\}$. That is, in a test-transition $\langle q \rhd_i p \rangle$, only individual i's utility is improved but all others remain unchanged. This notion will be a key for our study.

It was shown in Fishburn [8] that there are test-transitions for all $i \in N$ if and only if the coefficients $\alpha_1, ..., \alpha_n$ in Theorem 2.3 are uniquely determined. The existence of test-transitions for all $i \in N$ is later called the *Independent Prospect condition* (see also Weymark [33] and d'Aspremont-Gevers [4] for more details).

3. Double-Profile Approach

As stated in Section 1, the double-profile frame is given as $D = \left[\preceq^0 | \, \{\langle \preceq^i ; \preceq_{i1}, ..., \preceq_{in} \rangle\}_{i \in N}\right]$ in (1.2). Preference relation $\preceq_{ij}$ is personal private preferences of individual j believed by individual i, and preference relation $\preceq^i$ is personal ethical preferences of individual i. Finally, $\preceq^0$ is social preferences. Throughout the paper, we assume that those preference relations all satisfy Axiom NM1-NM3.

We will consider the two steps of aggregations:

Step A▶B: $\preceq^i$ is obtained from $\preceq_{i1}, ..., \preceq_{in}$;

Step B▶C: $\preceq^0$ is obtained from $\preceq^1, ..., \preceq^n$.

These correspond to the aggregations from (a) to (b) and from (b) to (c) in Section 1. Step A▶B is taken by personal individual i, and step B▶C is taken by the hypothetical outside observer[4]. Thus, these two steps have different natures. Here, we discuss the difference between these steps before going to the mathematical analysis of them.

In step A▶B taken by personal individual i, personal preference relation $\preceq_{ij}$ of individual j is imagined in the mind of individual i. As stated as in Section 1, a personal individual is supposed to be an ordinary person in society without special knowledge about society or other people. This means that $\preceq_{ij}$ may contain false components or is simply false. To discuss such falsity, we need to assume the existence of the true personal preferences. It may be possible to assume that $\preceq_{ii}$ is the true one, though this assumption itself is already quite demanding[5]. When we consider the aggregation from (a) to (c), e.g., from $\langle \preceq_{11}, ..., \preceq_{nn} \rangle$ to $\preceq^0$, these preferences may be regarded as true ones known to the hypothetical observer. We postpone this problem to Section 7, and now we consider the aggregation steps A▶B and B▶C.

Since individual i is an ordinary person, it would be natural for him to take only his relatives or neighbors seriously into account and simply ignores other people. This ignorance of some individual j by i is treated as having constant relations $\preceq_{ij}$, i.e., all

[4] One interpretation of the hypothetical outside observer is the ideal (computation) machine to treat information. This idea was discussed in Kaneko [16] to argue that it could not make decision unless it is given rules for mechanical decision making. In this paper, we do not touch this problem.

[5] For an idividual, having his preferences entirely differs from knowing his own preferences. It is similar to that we have brains, but we do not know their structur and functionings.

prospects are indifferent for j in i's mind. This case will be expressed as coefficients $\alpha_{ij} = 0$ for utility functions in the theorem to be given in Section 4.

Thus, individual i develops his ethical preferences $\preceq^i$ in his individualistic manner, which may be arbitrary from the objective point of view. Nevertheless, assuming the existence of a well-defined $\preceq^i$ satisfying NM1-NM3 axioms already involves some type of interpersonal comparability of utilities (Broome [3]). Here, individual i makes such comparisons in his individualistic manner.

Interpersonal comparability of utilities is sometimes justified by claiming that real people are making such comparisons. However, this is entirely different from the existence of an objective and scientific procedure to measure interpersonal comparisons of utilities. The former is an anthropological observation, and the latter is to demand the existence of some impersonal and impartial procedure. This will be discussed in Sections 5 and 7.

Step B►C is conceptually very different from step A►B. This step is taken by the hypothetical outside observer who is assumed to be impersonal and impartial. We interpret these conditions as meaning that he knows all information about $\preceq^1, ..., \preceq^n$ and aggregates them those relations into $\preceq^0$. This is also related to the problem of interpersonal comparability of utility. In Sections 5 and 7., we will give some comments on these points.

Although, as pointed out in Section 1, it may be problematic to assume the expected utility axioms, specifically, Axiom NM3, on $\preceq^0$, we avoid to talk about this problem. Rather we will talk about how interpersonal comparisons of utilities are treated operationally. While such comparisons are arbitrary for personal ethical preferences, they cannot be arbitrary in step B►C. Some operational form will be given.

For the above argument, the existence of a test-transition plays an important role. When the cardinality of X is small relative to n or the variety of $\preceq^1, ..., \preceq^n$ is small, we cannot guarantee the existence of test-transitions for all individuals. On the other hand, when the cardinality of X and $\preceq^1, ..., \preceq^n$ have some large variety relative to n, we can have test-transitions. However, since $\preceq^1, ..., \preceq^n$ are already aggregated concepts and cannot be given arbitrary, they would have a smaller variety than $\preceq_{i1}, ..., \preceq_{in}$. Hence, we should examine a condition for the existence of test-transitions for individuals, which will be discussed in Section 6.

Let us see that $\preceq_{i1}, ..., \preceq_{in}$ have typically a great variety. Recall that preferences $\preceq_{ij}$ may be relevant j's private domain of behavior. To describe this idea, one possibility is to assume that the set of pure alternatives X is given as the product form $X = \prod_{j \in N} X_j$, and then that $\preceq_{ij}$ depends only upon the j-th coordinate X_j. In this case, it may be the case that the projected preference relations over X_j and $X_{j'}$ for j and j' may be identical, but the original $\preceq_{ij}$ and $\preceq_{ij'}$ are entirely different, since they have different relevant coordinates in X. Also the cardinality of X is much larger than the number of individuals n. For example, even when $|X_j| = 2$ for all $j \in N$, then $n < 2^n = |X|$. This observation about the variety found in $\preceq_{i1}, ..., \preceq_{in}$ will be relevant in the consideration of the existence of test-transitions in Section 6.

4. From Personal Private Preferences to Personal Ethical Preferences

In this section, we consider an aggregation from personal private preferences to person ethical preferences. Mathematically speaking, Theorem 2.3 may be regarded as this aggregation. However, we will give a reformulation to have a clear-cut statement and state one additional axiom.

We give the following additional axiom for $\langle \preceq^i; \preceq_{i1}, ..., \preceq_{in} \rangle$.

Axiom PC$_i$ (Positive Correlation): There exists a test-transition $\langle q \rhd_i p \rangle$ for individual i such that $q \prec^i p$.

It states that there is some transition from q to p such that only individual i is better off, all the others are indifferent between q and p and also individual i ethically prefers p to q (i.e., $q \prec^i p$). His ethical judgement is made in favor for himself.

Now, we have the following theorem.

Theorem 4.1 (Subjective utilitarianism): Assume Axioms H and PC$_i$ for $\langle \preceq^i; \preceq_{i1} , ..., \preceq_{in} \rangle$. Take an arbitrary $u^i \in U(\preceq^i)$. There are disjoint subsets N^+, N^- of N and $u_{ij} \in U(\preceq_{ij})$ for all $j \in N^+ \cup N^-$ such that $i \in N^+$ and

$$u^i(p) = \sum_{j \in N^+} u_{ij}(p) - \sum_{j \in N^-} u_{ij}(p) \text{ for all } p \in L(X), \tag{4.1}$$

where u_{ii} is uniquely determined.

Without Axiom PC$_i$, this theorem is simply a different representation of Theorem 2.3, and N^+ may be empty in addition to the possible emptiness of N^-. The reason for presenting this theorem is to emphasize that in the aggregation of personal private preferences to person ethical preferences, individual i makes his ethical judgement in his individualistic manner: He can make any judgement. First, N^+ consists of people whose preferences individual i assesses positively; and N^- consists of people whose preferences individual i assesses negatively. Moreover, individual i may ignore people in $N - (N^+ \cup N^-)$. It may be the case where $N^+ \cup N^-$ is a small portion including i himself in society. If we stick to our interpretation that individual i makes his ethical judgement in his individualistic manner, this interpretation would be natural.

In the above theorem, u_{ij}'s and N^+, N^- may not be uniquely determined, since they may be correlated. One sufficient condition for an individual j to belong to N^+ (or N^-) is that there is a test-transition $\langle q \rhd_j p \rangle$ for individual j in $\langle \preceq^i; \preceq_{i1}, ..., \preceq_{in} \rangle$ such that $q \prec^i p$ (or $p \prec^i q$, respectively), in which case, u_{ij} is uniquely determined.

Proof of Theorem 4.1. Take an arbitrary $u^i \in U(\preceq^i)$ and $v_{i1} \in U(\preceq_{i1}), ..., v_{in} \in U(\preceq_{in})$. By Theorem 2.3, there are $\alpha_{i1}, ..., \alpha_{in}$ and β_i such that

$$u^i(p) = \sum_j \alpha_{ij} v_{ij}(p) + \beta_i \text{ for all } p \in L(X).$$

By Axiom PC$_i$, there is a test-transition $\langle q \rhd_i p \rangle$ for individual i such that $q \prec^i p$. Using the above formula, we evaluate this test-transition $\langle q \rhd_i p \rangle$:

$$u^i(p) - u^i(q) = \alpha_{ij}(v_{ii}(p) - v_{ii}(q)).$$

Since $q \prec^i p$, α_{ii} is positive and is uniquely determined to be $(u^i(p) - u^i(q))/(v_{ii}(p) - v_{ii}(q))$.

Now, let $N^+ = \{j : \alpha_{ij} > 0\}$ and $N^- = \{j : \alpha_{ij} < 0\}$. Then, N^+ and N^- are disjoint and $i \in N^+$. Now, we define: for $j \in N^+$, $u_{ij}(p) = \alpha_{ij} v_{ij}(p) + \beta_i / |N^+|$ for all $p \in L(X)$. Then, $u_{ij} \in U(\preceq_{ij})$ for all $j \in N^+$ by Theorem 2.2. Also, we define: for $j \in N^-$, $u_{ij}(p) = -\alpha_{ij} v_{ij}(p)$ for all $p \in L(X)$. Then, $u_{ij} \in U(\preceq_{ij})$ for all $j \in N^-$ by Theorem 2.2. Using those u_{ij}'s, we have (4.1). ■

5. Aggregation of Personal Ethical Preferences into Social Preferences

In Theorem 4.1(subjective utilitarianism), interpersonal comparability of utilities (preferences) is already included in the assumption of the existence of the well-defined preferences $\preceq^i$ over $L(X)$. However, this ethical judgement is made by individual i in his individualistic manner. We did not give any concrete constraint on it. For the aggregation of personal ethical preferences to social preferences, we give a certain axiom on the aggregation reflecting interpersonal comparisons of utilities.

Now, we give two axioms which look quite different from Axiom H, but actually, these two axioms imply Axiom H. This is related to Weymark's [33] result, which will be used in the proof of the main theorem.

The first axiom is as follows:

Axiom WP (Weak Pareto): If $q \prec^i p$ for all $i \in N$, then $q \prec^0 p$.

This states that social preferences are positively associated with personal ethical preferences. This is not yet related to interpersonal comparability of utilities. The next axiom formulates a necessary condition for interpersonal comparability of utilities. For any two individuals, utility differences (preference differences) are expressed in terms of test-transitions, and they are compared. The next axiom claims that some utility differences give the same effect on social preferences, and hence the results obtained by the test-transitions are socially indifferent. This is formulated in terms of compound prospects.

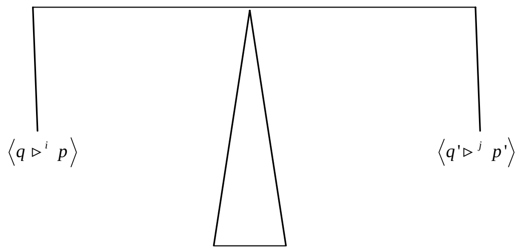

Figure 5.1: the ICU scale

We formulate interpersonal comparability of utilities in the following form for $\langle \preceq^0 : \preceq^1 , ..., \preceq^n \rangle$.

Axiom ICU (Interpersonal Comparability of Utilities): For any $i, j \in N$, there are test-transitions $\langle q \rhd^i p \rangle$ and $\langle q' \rhd^j p' \rangle$ for i and j, respectively, such that $\frac{1}{2}p * \frac{1}{2}q' \sim^0 \frac{1}{2}q * \frac{1}{2}p'$. Note that all test-transitions are in $\langle \preceq^1, ..., \preceq^n \rangle$.

$$\begin{tikzcd} & \frac{1}{2}q * \frac{1}{2}q' \arrow[dl] \arrow[dr] & \\ \frac{1}{2}p * \frac{1}{2}q' & \sim^0 & \frac{1}{2}q * \frac{1}{2}p' \end{tikzcd}$$

Diagram 5.1

In expected utility theory, utility differences or preferences over transitions are not defined[6]. However, using compound prospects, we can effectively use such differences. The idea is depicted in Diagram 5.1. The transition from $\frac{1}{2}q * \frac{1}{2}q'$ to $\frac{1}{2}p * \frac{1}{2}q'$ is regarded as test-transition $\langle q \rhd^i p \rangle$, and the transition from $\frac{1}{2}q * \frac{1}{2}q'$ to $\frac{1}{2}q * \frac{1}{2}p'$ is test-transition $\langle q' \rhd^j p' \rangle$. The intended social indifference between $\langle q \rhd^i p \rangle$ and $\langle q' \rhd^j p' \rangle$ is expressed by the social indifference between $\frac{1}{2}p * \frac{1}{2}q'$ and $\frac{1}{2}q * \frac{1}{2}p'$.

Using the utility function u^0, this indifference between the two transitions is written as

$$u^0(\frac{1}{2}p * \frac{1}{2}q') - u^0(\frac{1}{2}q * \frac{1}{2}q') = u^0(\frac{1}{2}q * \frac{1}{2}p') - u^0(\frac{1}{2}q * \frac{1}{2}q'),$$

from which we have, by Theorem 2.1.(2), $\frac{1}{2}u^0(p) - \frac{1}{2}u^0(q) = \frac{1}{2}u^0(p') - \frac{1}{2}u^0(q')$, i.e.,

$$u^0(p) - u^0(q) = u^0(p') - u^0(q').$$

[6] Kaneko [16] used an argument similar to Axiom ICU so as to introduce utility differences into (one-person) expected utility theory (see also Shubik [31], Appendix A). This was against Luce-Raiffa [21]'s fallary 3 (p.32) that expected utility theory does not include the notion of utility differences as its legitimate components and so utility differences should not be discussed.

If we have already $u^0(\cdot) = \sum_{l\in N} u^l(\cdot)$, then this equation becomes $u^i(p)-u^i(q) = u^j(p')-u^j(q')$. Thus, Axiom ICU asserts interpersonal comparability of utility differences, rather than of absolute utility levels.

It is the basic assumption that the hypothetical outside observer has some objective and scientific way to make comparisons of utility (preference) differences, which we call the *ICU scale* (Figure 5.1). By this scale the hypothetical outside observer makes comparisons of individual utility differences for any two individuals. After some comparisons, he can find two utility differences to balance the ICU scale, which means that these two utility differences are equal with respect to the ICU scale. In this case, Axiom ICU claims that when all the other individuals are fixed, the two utility differences are evaluated as the same with respect to social preferences.

Now, we state the main theorem of this section. After presenting the theorem, we will discuss how we interpret Axiom ICU with respect to interpersonal comparability of utilities.

Theorem 5.1 (From Personal Ethical to Social Preferences I): Assume Axioms WP and ICU for $\langle\preceq^0:\preceq^1,...,\preceq^n\rangle$. Take an arbitrary $u^0 \in U(\preceq^0)$. There are $(u^1,...,u^n) \in U(\preceq^1)\times ... \times U(\preceq^n)$ such that

(1): $u^0(p) = \sum_{i\in N} u^i(p)$ for all $p \in L(X)$;

(2): if $\langle q \rhd^i p\rangle$ and $\langle q' \rhd^j p'\rangle$ are test-transitions for individuals $i, j \in N$ given in Axiom ICU, then

$$u^i(p) - u^i(q) = u^j(p') - u^j(q'). \tag{5.1}$$

Before a proof of Theorem 5.1, we will give various comments on the theorem. In this theorem, Axiom H is not needed and is, actually, obtained from Axioms WP and ICU. This will be clearer when we give a proof of the theorem. The logical independence of Axioms WP and ICU will also become clear in our proof of the theorem.

We have alternative formulations of the above theorem. Here, we mention one alternative formulation[7]. In Axiom ICU, utility differences between any two people are compared, and test-transitions are allowed to depend upon two individuals. An alternative formulation of Axiom ICU is to require the existence of partially common test-transitions for all people.

Axiom ICU*: There are test-transitions $\{\langle q^i \rhd^i p^i\rangle : i \in N\}$ such that for all $i, j \in N$, $q^i = q^j = q$ and $\frac{n-1}{n}q * \frac{1}{n}p^i \sim^0 \frac{n-1}{n}q * \frac{1}{n}p^j$.

[7] Another alternative to Axiom ICU is given in d'Aspremont-Gevers [4] in the analysis of a social welfare function: Axiom ICU^0 : for all i, i', there are $p, q \in L(X)$ such that $p \prec^i q$, $q \prec^{i'} p$, $p \sim^j q$ for all $j \neq i, i'$ and $p \sim^0 q$. Theorem 5.1 remains valid if we keep Axiom WP, replace Axiom ICU by ICU^0 and additionally assume Minimum Agreement condition: for some $p, q \in L(X)$, $p \prec^i q$ for all $i \in N$. This can be proved using Weymark's [33] theorem (Lemma 5.1 in this paper).

We can replace Axiom ICU by this ICU* for Theorem 5.1. The proof, given below, of Theorem 5.1 can be modified slightly to obtain this version. Actually, Axiom ICU* is equivalent to Axiom ICU under the expected utility axioms NM1-NM3 and Axiom WP.

Theorem 5.1 is formulated in order to see in what way interpersonal comparability of utilities is involved. However, it is also possible to formulate the essentially same theorem by avoiding an explicit formulation of interpersonal comparability of utilities, which was given by Weymark [33]. We contrast Theorem 5.1 with the theorem without using an explicit formulation of interpersonal comparability of utilities.

First, we modify Axiom PC$_i$ $\langle \preceq^i ; \preceq_{i1}, ..., \preceq_{in} \rangle$ into the following:

Axiom PC (Positive Correlation): There exist test-transitions $\{\langle q^i \rhd^i p^i \rangle : i \in N\}$ in $\langle \preceq^0 : \preceq^1, ..., \preceq^n \rangle$ such that $q^i \prec^0 p^i$.

It states that every individual i has a transition $\langle q^i \rhd^i p^i \rangle$ so that the society prefers p^i to q^i. Here, interpersonal comparability of utilities is not explicit. However, we have the same consequence as Theorem 5.1, which was, more or less, included in Weymark [33] for the single-profile approach.

Theorem 5.2 (From Personal Ethical to Social Preferences II): Assume Axioms H and PC for $\langle \preceq^0 : \preceq^1, ..., \preceq^n \rangle$. Take an arbitrary $u^0 \in U(\preceq^0)$. There are $(u^1, ..., u^n) \in U(\preceq^1) \times ... \times U(\preceq^n)$ such that

(1): $u^0(p) = \sum_{i \in N} u^i(p)$ for all $p \in L(X)$;

(2): for any $i, j \in N$, there are test transitions $\langle q \rhd^i p \rangle$ and $\langle q' \rhd^j p' \rangle$ for $i, j \in N$ such that (5.1) holds.

Thus, interpersonal comparability of utility differences is not explicitly formulated, but (2) means that it is derived from the other axioms. We interpret this as meaning that interpersonal comparability of utility differences is already included in $\langle \preceq^0 : \preceq^1, ..., \preceq^n \rangle$. This is the view of Broome [3], which we mentioned in Section 3.

The first assertion of this theorem can be proved in the same as the proof of Theorem 4.1, but the second assertion will be proved in the end of this section.

Let us return to Theorem 5.1. Since it asserts a representation of $\preceq^0$, (1) can be replaced by

(1*): $u^0(p) = \dfrac{1}{n} \sum_{i \in N} u^i(p)$ for all $p \in L(X)$.

As far as we take Theorem 5.1 as a representation theorem, we have no conceptual difference between (1) and (1*). However, from the the viewpoint of Harsanyi's [10] "veil of ignorance", we should take (1*). We adopt (1) for the main theorem since there are no differences in their welfare implications and since (1) is simpler than (1*).

Besides the fact the single-profile approach includes implicitly interpersonal comparability as mentioned above, we may ask why Axiom ICU is not assumed in the aggregation from personal private preferences (a) to personal ethical preferences (b). As already stressed several times, a personal individual may be very individualistic and not impartial: He may evaluate other people's private preferences negatively or even may ignore their existence. For this reason, we impose Axiom ICU for only the aggregation by the hypothetical outside observer.

As far as $\preceq_{ij}$ and $\preceq_{ij'}$ are positively taken in the scope of individual i, we may assume Axiom ICU for between j and j' in the mind of individual i, like Theorem 5.2. That is, we can have Axiom ICU for j and j' in the mind of individual i under the assumption of the existence of test-transitions for j and j' in $\langle \preceq^i; \preceq_{i1}, ..., \preceq_{in}\rangle$. Thus, to this extent, we can have Axiom ICU for the aggregation from (a) to (b). When $\preceq_{ij}$ and $\preceq_{ij'}$ are negatively associated by individual i, we need to consider negative test-transitions for individuals j and j.

Among people close each other such as family members, some ways having the property described by Axiom ICU may be used even for a personal individual's comparisons. They follow social custom or convention taught by elders. If some have positive evaluations each other, Axiom ICU holds effectively for them.

The above paragraphs raise a question of what aspect of interpretational comparisons of utilities Axiom ICU captures. We have assumed the existence of a procedure to measure utility differences interpersonally, which is expressed by the ICU scale. Now, the very basic problem is whether or not we can define such a procedure in an objective and scientific (impersonal and impartial) manner. Axiom ICU describes only the necessary condition for such a scale, assuming its existence. This problem will be discussed again in Section 7.

For a proof of Theorem 5.1, we start with Weymark's [33], p.216, Theorem 2.

Lemma 5.1 (Weymark): Let $\langle \preceq; \preceq_1, ..., \preceq_n\rangle$ be a single-profile frame satisfying Axiom WP. Then, for any $u \in U(\preceq)$ and $u_i \in U(\preceq_i), i \in N$, there are $\eta \geq 0$, $\alpha_i \geq 0$, $i \in N$ and β such that $\alpha_j > 0$ for some $j \in N$ and

$$\eta u(p) = \sum_{i \in N} \alpha_i u_i(p) + \beta \text{ for all } p \in N. \tag{5.2}$$

In addition, if we have some $p, q \in L(X)$ with $q \prec_i p$ for all $i \in N$, then, η is positive.

Using Lemma 5.1, we can now prove Theorem 5.1.

Proof of Theorem 5.1. To prove this theorem, we need to show observation that

$$\text{there are } p, q \in L(X) \text{ such that } q \prec^i p \text{ for all } i \in N \text{ and } q \prec^0 p. \tag{5.3}$$

This is not guaranteed by Axiom WP, but follows from Axioms ICU and WP.[8] By Axiom ICU, we have test-transitions $\{\langle q^i \rhd^i p^i\rangle : i \in N\}$. We let $q = (\frac{1}{n}q^1 * ... * \frac{1}{n}q^n)$ and $p = (\frac{1}{n}p^1 * ... * \frac{1}{n}p^n)$. Then, for any $i \in N$ and any $v^i \in U(\preceq^i)$, we have $v^i(q^j) = v^i(p^j)$ for $j \neq i$ and $v^i(q^i) < v^i(p^i)$ since $\{\langle q^i \rhd^i p^i\rangle : i \in N\}$ are test-transitions for all individuals. This implies

$$v^i(q) = \frac{1}{n}(\sum_{j \neq i} v^i(q^j) + v^i(q^i)) < \frac{1}{n}(\sum_{j \neq i} v^i(p^j) + v^i(p^i)) = v^i(p).$$

Since this strict inequality holds for all $i \in N$, we have $q \prec^0 p$ by Axiom WP[9]. Thus, we have the minimal agreement condition for $\langle \preceq^0; \preceq^1, ..., \preceq^n\rangle$.

Now, take an arbitrary $u^0 \in U(\preceq^0)$ and $v^1 \in U(\preceq^1), ..., v^n \in U(\preceq^n)$. By Lemma 5.1, we have $\alpha_1, ..., \alpha_n$ and β_0 such that

$$u^0(p) = \sum_j \alpha_j v^j(p) + \beta_0 \text{ for all } p \in L(X). \tag{5.4}$$

Here, we can assume $\eta = 1$ in (5.2) by a transformation since $\eta > 0$ by the additional statement of Lemma 5.1. Now, we show that $\alpha_j > 0$ for all $j \in N$. Once this is proved, we can define $u^j(\cdot) = \alpha_j v^j(\cdot) + \beta_0/n$ for all $j \in N$. Then we have the first assertion.

By (5.3), there are p, q such that $0 < u^0(p) - u^0(q)$ and $0 < v^j(p) - v^j(q)$ for all $j \in N$. Thus, $\alpha_i > 0$ for at least one individual i.

Now, we take any $i' \neq i$. Then, by Axiom ICU, we find test-transitions $\langle q \rhd^i p\rangle$ and $\langle q' \rhd^{i'} p'\rangle$ for i and i' with $\frac{1}{2}p * \frac{1}{2}q' \sim^0 \frac{1}{2}q * \frac{1}{2}p'$. This together with (5.4) implies $\sum_j \alpha_j(\frac{1}{2}v^j(p) + \frac{1}{2}v^j(q')) = \sum_j \alpha_j(\frac{1}{2}v^j(q) + \frac{1}{2}v^j(p'))$, which is equivalent to

$$\sum_j \alpha_j(v^j(p) - v^j(q)) = \sum_j \alpha_j(v^j(p') - v^j(q')). \tag{5.5}$$

Since $\langle q \rhd^i p\rangle$ and $\langle q' \rhd^{i'} p'\rangle$ are test-transitions for i and i', respectively, (5.5) implies

$$\alpha_i(v^i(p) - v^i(q)) = \alpha_{i'}(v^{i'}(p') - v^{i'}(q')). \tag{5.6}$$

Now, $0 < \alpha_i$ by the previous result and $0 < v^i(p) - v^i(q)$ because $\langle q \rhd^i p\rangle$ is a test-transition. Thus, we have $0 < \alpha_{i'}(v^{i'}(p') - v^{i'}(q'))$. Since $0 < v^{i'}(p') - v^{i'}(q')$ because $\langle q' \rhd^{i'} p'\rangle$ is a test-transition, we have $\alpha_{i'} > 0$. Now, we have shown that $\alpha_j > 0$ for all $j \in N$.

When we set $u^j(\cdot) = \alpha_j v^j(\cdot) + \beta_0/n$ for all $j \in N$, we have assertions (1) by (5.4) and (2) by (5.6). ■

[8] This assertion was given as Proposition 3 in Weymark [33]. But for the reader's convenience sake, we give a proof of it.

[9] This argument for the existence of p, q was suggested by Mongin in his personal communication to Weymark (see Weymark [33].)

Proof of Theorem 5.2.(2): By Axiom PC, we have test-transitions $\langle q^i \rhd^i p^i \rangle$ and $\langle q^j \rhd^j p^j \rangle$ for any individuals i, j. By (1) of this theorem, we have $u^i(p^i) - u^i(q^i) > 0$ and $u^j(p^j) - u^j(q^j) > 0$. Suppose $u^i(p^i) - u^i(q^i) \geq u^j(p^j) - u^j(q^j)$. Then, there is an $\alpha > 0$ such that $\alpha(u^i(p^i) - u^i(q^i)) = u^j(p^j) - u^j(q^j)$. Then, let $p^* = \alpha p^i + (1 - \alpha) q^i$. Then, $u^i(p^*) - u^i(q^i) = \alpha(u^i(p^i) - u^i(q^i)) = u^j(p^j) - u^j(q^j)$. ■

6. Remarks on the Existence of Test-Transitions

Each personal ethical preference relation $\preceq^i$ is obtained by aggregating personal private preferences $\preceq_{i1}, ..., \preceq_{in}$. Therefore, $\preceq^1, ..., \preceq^n$ may have some similarity to prevent from having test-transitions. We start with one example.

Example 6.1: We suppose that the personal ethical preferences are all the same, i.e., $\preceq^1 = ... = \preceq^n$. Then, no individual has any test-transition, i.e., if i prefers p to q, all the others prefer the same. This is an extreme case, but it may come from the following situation: Suppose that every individual i has the same beliefs over the private preference relations of others, that is, $\preceq_{ij} = \preceq_{i'j}$ for all $i, i', j \in N$. It is still allowed that $\preceq_{ij}$ varies with j. If all individuals i have an identical aggregation principles from (a) to (b), then we could have $\preceq^1 = ... = \preceq^n$. However, if $\preceq_{ij}$ and $\preceq_{i'j}$ are different and/or if their aggregation principles are different, we would expect some variety in $\preceq^1$, ..., $\preceq^n$. Thus, it is not immediate to see whether or not $\langle \preceq^1, ..., \preceq^n \rangle$ have test-transitions.

Here, we consider conditions for the existence of test-transitions. The first result for the existence of a test-transition is about a fixed individual i. For the following theorem, recall the vector expression $\mathbf{u} = (u(x_1), ..., u(x_k))$ introduced in Section 2.

Theorem 6.1 (Existence of a Test-Transition for One): There is a test-transition $\langle q \rhd^i p \rangle$ for individual i if and only if for any $u^1 \in U(\preceq^1), .., u^n \in U(\preceq^n)$, $\mathbf{u}^i$ cannot be expressed as a linear combination of $(\mathbf{u}^j)_{j \neq i}$ and $\mathbf{1} = (1, ..., 1)$, i.e., there are no $(\alpha_j)_{j \neq i}$ and β satisfying

$$\mathbf{u}^i = \sum_{j \neq i} \alpha_j \mathbf{u}^j + \beta \mathbf{1}. \tag{6.1}$$

Proof. (*If*): We prove the contrapositive of the assertion. Suppose that there is no test-transition $\langle q \rhd^i p \rangle$ for individual i. It means that for any $p, q \in L(X)$, if $p \sim^j q$ for any $j \neq i$, then $p \sim^i q$. This is regarded as Axiom H (Pareto Indifference) for $\langle \preceq^i; (\preceq^j)_{j \neq i} \rangle$. Hence, we can apply Theorem 2.3 (Harsanyi-Domotor) to this, and then for any $u^i \in U(\preceq^i)$, $u^j \in U(\preceq^j)$ $(j \neq i)$, we have coefficients $(\alpha_j)_{j \neq i}$ and β satisfying

$$u^i(p) = \sum_{j \neq i} \alpha_j u^j(p) + \beta \text{ for all } p \in L(X).$$

Letting $p = x_t$ for each $t = 1, ..., k$, we have (6.1). This means the negation of the latter statement of the theorem.

(*Only-If*): We prove also the contrapositive of the assertion. Suppose that for any $u^i \in U(\preceq^i)$ and any $u^j \in U(\preceq^j)$ $(j \neq i)$, there are coefficients $(\alpha_j)_{j\neq i}$ and β satisfying (6.1). From this, we obtain, for all $p \in L(X)$,

$$u^i(p) = p \cdot \mathbf{u}^i = p \cdot (\sum_{j\neq i} \alpha_j \mathbf{u}^j + \beta \mathbf{1}) = \sum_{j\neq i} \alpha_j \sum_{t=1}^{k} p_t(u^j(x_t) + \beta) = \sum_{j\neq i} \alpha_j u^j(p) + \beta. \quad (6.2)$$

Now, let p, q be any prospects. Suppose $q \prec^i p$. Then, $u^i(q) < u^i(p)$. By (6.2), $u^j(q) \neq u^j(p)$ for some $j \neq i$. This means that there is no test-transition for individual i. ■

Combining the above theorem for all individuals, we have the following theorem due to Weymark [33], p.213, Proposition 2.

Theorem 6.2 (Existence of Test-Transitions for All): For all $i \in N$ there is a test-transition $\langle q \rhd^i p \rangle$ for i if and only if for any $u^1 \in U(\preceq^1), ..., u^n \in U(\preceq^n)$, the vectors $\mathbf{u}^1, ..., \mathbf{u}^n, \mathbf{1}$ are linearly independent in the sense of linear algebra.

This theorem states that to have test-transitions for all $i \in N$, vectors $\mathbf{u}^1, ..., \mathbf{u}^n, \mathbf{1}$ are linear independent, which implies that the attainable utility set, i.e., the convex hull of $\mathbf{u}^1, ..., \mathbf{u}^n, \mathbf{1}$ is n dimensions. Recall that each of $\mathbf{u}^1, ..., \mathbf{u}^n, \mathbf{1}$ is a k-vector. When $k = |X|$ is larger than n, and when $\langle \preceq_{11}, ..., \preceq_{1n} \rangle, ..., \langle \preceq_{n1}, ..., \preceq_{nn} \rangle$ are diversified enough, the linear independence of $\mathbf{u}^1, ..., \mathbf{u}^n, \mathbf{1}$ is typically expected. As stated in the last paragraph of Section 3, $|X|$ is expected to be much larger than n. Unless the personal aggregation procedure from $\langle \preceq_{11}, ..., \preceq_{1n} \rangle, ..., \langle \preceq_{n1}, ..., \preceq_{nn} \rangle$ to $\preceq^1, ..., \preceq^n$ are perfectly uniform, we can expect enough varieties in $\preceq^1, ..., \preceq^n$. In sum, it would be expected to have test-transitions for all individuals.

7. Discussions

We provided the double-profile approach to Harsanyi's preference utilitarianism, and considered the two steps of aggregations. In the first step from personal private preferences $\langle \preceq_{i1}, ..., \preceq_{in} \rangle$ to personal ethical preferences $\preceq^i$, the aggregation theorem we gave was, more or less, the same as the Harsanyi-Domotor Theorem. In the second step from personal ethical preferences $\langle \preceq^1, ..., \preceq^n \rangle$ to social preferences $\preceq^0$, Axiom ICM was the main axiom for aggregation. The axiomatizations of those steps were given so as to explain the conceptual difference between these two steps.

In the first step, we gave Axiom PD$_i$ to determine only the positive association between individual i's private preferences and his ethical preferences. The other individuals' personal preferences are associated with i's ethical preferences in an arbitrary way. We interpret this as meaning that individual i could make ethical judgement in his individualistic manner.

Contrary to the first step, in the second step, we required two axioms, one of which describes a way of interpersonal comparisons of utility differences. We simply formulated this axiom as the form as if the hypothetical outside observer has a well-defined scale to measure utility differences. Although interpersonal comparability of utilities is one central issue, before the consideration of it, perhaps, we should give some comments on the aggregation from personal private preferences (a) to social preferences (c), and also on the expected utility axioms on social preferences.

There are two ways for the aggregation from (a) to (c). One is simply to concatenate the aggregation from (a) to (b) to the aggregation from (b) to (c). This means to combine Theorems 4.1 and 5.1. For the double-profile frame $D = \left[\preceq^0 | \ \{\langle \preceq^i; \preceq_{i1}, ..., \preceq_{in}\rangle\}_{i \in N}\right]$, we assume Axioms H and PC$_i$ for each $\langle \preceq^i; \preceq_{i1}, ..., \preceq_{in}\rangle$ and Axioms WP, ICU for $\langle \preceq^0; \preceq^1, ..., \preceq^n\rangle$, in addition to Axioms NM1-NM3 for all the preference relations. Then, we have

$$u^0(p) = \sum_{i \in N} (\sum_{j \in N_i^+} u_{ij}(p) - \sum_{j \in N_i^-} u_{ij}(p)) \text{ for all } p \in L(X),$$

where we put the subscript i to the set N^+ and N^- given by Theorems 4.1 for $\langle \preceq^i; \preceq_{i1}, ..., \preceq_{in}\rangle$. In this aggregation, personal private preferences are counted several times, which was the criticism given by Ng [25].

The above aggregation is a straightforward concatenation of the two steps. Here, it would be more important to remark about the aggregation from (a) to (c) from the viewpoint of the hypothetical outside observer. In this case, it does not make sense to take all subjective preferences $\preceq_{i1}, ..., \preceq_{in}, i \in N$. Instead, the observer should take the true preferences if they ever exist. If $\preceq_{11}, ..., \preceq_{nn}$ are true preferences, he aggregates them directly to $\preceq^0$, which takes the same form as Theorem 5.1. The existence of true preferences may be problematic, but this criticism can be applied to the existence of other preferences. Another possible criticism is that individual i's personal preferences include ethical sentiments as well as private ones. It means that the separation between (a) and (b) may be difficult. This was suggested by Sobel's [32], p.259. Nevertheless, if personal private preferences can be defined, his suggestion is interpreted as meaning that (b) is obtained from (a). At any rate, we should admit that there are a lot of subtle conceptual problems around these interpretations.

Now, let us consider the adoption of Axioms NM1-NM3 for all preferences; in particular, Axiom NM3 may be problematic to social preferences. We may say that it is not problematic for personal private preferences (a) and personal ethical preferences (b), but is problematic certainly for social preferences (c). One reason is the counterexample given by Diamond [5] for social preferences[10]. Here, we look at his example and a counter argument given by Nakamura-Nakayama [23], since by examining them, we

[10] The example given by Diamond [5] was already suggested by Hobbes [15], Part I, Chap.15, the 13th Law of Nature.

can see some consistency and closedness of the world of Harsanyi.

In the example, an indivisible unit M is distributed to two people 1 and 2. Consider two social alternatives: $(M, 0)$ and $(0, M)$, i.e., the indivisible unit M is given to individual 1 or 2 respectively. Under the assumption that these two individuals are symmetrically identical, we can assume that these two social alternatives are socially indifferent, i.e., $(M, 0) \sim^0 (0, M)$. By Axiom NM3.(2), we have

$$\frac{1}{2}(M, 0) * \frac{1}{2}(0, M) \sim^0 \frac{1}{2}(0, M) * \frac{1}{2}(0, M) = (0, M).$$

Similarly, $\frac{1}{2}(M, 0) * \frac{1}{2}(0, M) \sim^0 (M, 0)$. Thus, the prospect $\frac{1}{2}(M, 0) * \frac{1}{2}(0, M)$ is socially indifferent to simply giving M to one individual, though the prospect (lottery) gives each individual a fair chance to get M. Diamond [5] criticized that the prospect giving a fair chance should be socially preferred to simply giving M to one person.

Nakamura-Nakayama [23] gave a counter argument based on Harsanyi [10] to the criticism by Diamond. Their counter argument starts with the assumption that the social (or ethical) evaluations of $(M, 0)$ and $(0, M)$ should be made in the original position before the veil of ignorance. The evaluation of $(M, 0)$ in the original position is described as $\frac{1}{2}[(1, 2) : (M, 0)] * \frac{1}{2}[(2, 1) : (M, 0)]$, meaning that with probability $\frac{1}{2}$, individual in question would take position 1, and with the remaining probability, he would take the position 2. Hence, neither $(M, 0)$ nor $(0, M)$ could appear in the premise of Axiom NM3.(2). Thus, even if individual 2 is shown to the present situation $(M, 0)$, he should think of a chance to get M behind $(M, 0)$. Nakamura-Nakayama concluded that in the scope of Harsanyi's [10] view, Diamond's example is not really a counter example.

Reflecting carefully on the above argument, however, we would notice that Axiom NM3 is trivialized by falsifying its premise. If we follow this argument, we would not adopt the whole expected utility theory since only some specific probability distribution is allowed. Even though we forget this difficulty, it would be unclear from the beginning why we should think about Diamond's example in the scope of [10] view, since the original position given by Harsanyi [10] is a purely hypothetical concept constructed by himself. Diamond's [5] example sounds still natural and important for us to take it seriously. This is compatible with the suggestion by Sen [29] that the social welfare function should be nonlinear with respect to probability.

To a certain extent, those arguments are applied to personal ethical preferences if the individual ethical judgement shares the same principle or its part with the hypothetical outside observer. Actually we can avoid NM3 to construct a social preference relation, which is the theory of the Nash social welfare function given in Kaneko-Nakamura [17]. However, to discuss whether or not we should adopt Axiom NM3 for social preferences, we need some models of an individual being and the hypothetical outside observer, as given in Kaneko [16]. Without giving such models, only arguments based on our "intuition" are possible.

This last comment is also related to the problem of interpersonal comparability of utilities. Axiom ICU describes one necessary function of an objective and scientific scale to allow interpersonal comparisons of utility differences, under the assumption of the existence of such a scale. However, the possibility of such existence is difficult to be supported. Kaneko [16] and Broome [2] criticized such a possibility examining the very basis of individual preferences.

Nevertheless, there is another route to define interpersonal comparisons of utilities. Typically in the literature, the route from interpersonal comparisons to social preferences (social welfare) has been discussed. Another possibility is to reverse this order, i.e., social preferences are first defined and then interpersonal comparisons are defined based on the social preferences. In Theorem 5.2, we avoid the explicit use of interpersonal comparability of utilities. Nevertheless, its second assertion implies that interpersonal comparability of utility differences are already involved in the axiomatization. A different possibility of defining social welfare differences was suggested and discussed in Kaneko-Nakamura [18] along the line of Alt [1] and Shapley [30]. Kaneko-Nakamura adopted an axiom to define social welfare differences for the Nash social welfare function proposed in [17][11]. This consideration is possible only for the Nash social welfare function, but is not applied to Harsanyi's preference utilitarianism.

References

[1] Alt, F., (1936), Über die Messbarkeit des Nutzens, *Z. Nationalökonomie* 7:161-169. English translation, (1971). In: Chipmann, F., et al, eds, *Preferences, Utility and Demand*, Chap.20. Hourcour Brace Jovanovich, New York.

[2] Broome, J., (1993), A cause of preference is not an object of preference, *Social Choice Welfare* 10, 57-68.

[3] Broome, J., (1999), Can there be a preference-based utilitarianism?. In Maurice Salles and John Weymark (eds), *Justice, Political Liberalism and Utilitarianism*, Cambridge: Cambridge University Press.

[11] The theory of the Nash social welfare function needs the reference point called the *origin*, which corresponds, technically speaking, to the disagreement point Nash's [24] bargaining theory. However, the origin should not be arbitrary like as Nash's theory but should be defined from the viewpoint of social welfare. To this extent, interpersonal comparability of utilities is already involved in the theory of the Nash social welfare function.

[4] d'Aspremont, C. and L. Geverse, (2002), Social welfare functionals and interpersonal comparability, *Handbook of Social Choice and Welfare* Vol.1, 461-540. Elsvier Science B.V., London.

[5] Diamond, P., (1967), Cardinal welfare, individualistic ethics and interpersonal comparisons of utility: a comment, *Journal of Political Economy* 75, 765-766.

[6] Domotor, Z., (1979), Ordered sum and tensor product of linear utility structures, *Theory and Decision* 11, 375-399.

[7] Fishburn, P., (1970), *Utility Theory for Decision Making*, John Wiley & Sons, New York.

[8] Fishburn, P., (1984), On Harsanyi's utilitarian cardinal welfare theorem, *Theory and Decision* 17, 21-28.

[9] Hammond, P.J., (1998), Objective utility theory, *Hand of Utility Theory* 1, Kluwer Academic Press, Dordrecht, 143-211.

[10] Harsanyi, J. C., (1953), Cardinal utility in welfare economics and in the theory of risk-taking, *Journal of Political Economy* 61, 434-435.

[11] Harsanyi, J. C., (1955), Cardinal welfare, individualists ethics, and interpersonal comparisons of utility, *Journal of Political Economy* 63, 309-321.

[12] Harsanyi, J. C., (1975), Nonlinear social welfare functions, *Theory and Decision* 6, 311-332.

[13] Harsanyi, J. C., (1977), Morality and the theory of rational behavior, *Social Research* 44.

[14] Herstein, I. N. and J. Milnor, (1953), An axiomatic approach to measurable utility, *Econometrica* 21, 291-297.

[15] Hobbes, T., (1651), *Leviathan*, (1651): The Clarendon Press (1909), Oxford.

[16] Kaneko, M., (1984), On interpersonal utility comparisons, *Social Choice and Welfare* 1, 165-175.

[17] Kaneko, M. and K. Nakamura, (1979a), The Nash social welfare functions, *Econometrica* 47, 423-435.

[18] Kaneko, M. and K. Nakamura, (1979b), Cardinalization of the Nash social welfare functions, *Economic Study Quarterly* 30, 236-242.

[19] Kaneko, M., and M. H. Wooders, (2004), Utility theories in cooperative games, *Handbook of Utility Theory Vol.2.* Chapter 19, 1065-1098. Kluwer Academic Press.

[20] Karni, E. (1996), Social welfare functions and fairness, *Social Choice and Welfare* 13, 487-496.

[21] Luce, R. D., and H. Raiffa, (1957), *Games and Decisions*, John Wiley and Sons Inc., Boston.

[22] Mongin, P. (1994). Harsanyi's aggregation theorem: multiprofile version and unsettled questions, *Social Choice and Welfare* 11, 331-354.

[23] Nakamura, K., and M. Nakayama, (1978), Discrepancy between Harsanyi and Diamond on Social Preferences, *Journal of Economic Studies* 24, 233-237: included in *Game Theory and Social Choice, Selected papers of Kunjiro Nakamura.* ed. Mitsuo Suzuki, (1981), 95-99, Keiso Shobo.

[24] Nash, J. F. (1950), The Bargaining problem, *Econometrica* 18, 155-162.

[25] Ng, Y.-K., (1999), Utility, informed preference or happiness: Following Harsanyi's argument to its logical conclusion, *Social Choice and Welfare* 16, 197-216.

[26] Pattanaik, P. K., (1968), Risk, impersonality, and the social welfare function, *Journal of Political Economy* 76, 1153-1169.

[27] Rabinowicz, W. and Österberg, J., (1996), Value based on preferences: On two interpretations of preference utilitarianism, *Economics and Philosophy* 12, 1-27.

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

[29] Sen, A., (1973), On Economic Inequality, Clarendon Press, Oxford.

[30] Shapley, L. S., (1975), Cardinal utility from intensity comparisons, R-1683-PR, Rand Corporation.

[31] Shubik, M., (1984), *Game Theory in the Social Sciences: Concepts and Solutions*, MIT Press, Cambridge MA.

[32] Sobel, D., (1998), Well-being as the object of moral consideration, *Economics and Philosophy* 14, 249-281.

[33] Weymark, J. A. (1993), Harsanyi's social aggregation theorem and weak Pareto principle, *Social Choice and Welfare* 10, 209-221.

[34] von Neumann, J., and O. Morgenstern, (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.