DOI: 10.1111/phpr.12783

ORIGINAL ARTICLE

Philosophy and
Phenomenological Research

Aggregation Without Interpersonal Comparisons of Well-Being

Jacob M. Nebel

University of Southern California

Correspondence
Jacob M. Nebel
Email: jnebel@usc.edu

Abstract

This paper is about the role of interpersonal comparisons in Harsanyi's aggregation theorem. Harsanyi interpreted his theorem to show that a broadly utilitarian theory of distribution must be true even if there are no interpersonal comparisons of well-being. How is this possible? The orthodox view is that it is not. Some argue that the interpersonal comparability of well-being is hidden in Harsanyi's premises. Others argue that it is a surprising conclusion of Harsanyi's theorem, which is not presupposed by any one of the premises. I argue instead that Harsanyi was right: his theorem and its weighted-utilitarian conclusion do not require interpersonal comparisons of well-being. The key to making sense of this possibility is to treat Harsanyi's weights as dimensional constants rather than dimensionless numbers.

1 | INTRODUCTION

Harsanyi (1955) claimed to derive "an additive cardinal social welfare function" from principles of individual and social rationality and respect for rational preferences. According to Harsanyi's aggregation theorem, if both individual and social preferences satisfy the axioms of expected utility theory, and if society prefers anything that is preferred by all individuals, then society's preferences can be represented as maximizing a weighted sum of individual utilities.

Harsanyi doesn't call this social welfare function "utilitarian"—at least, not in that paper. But he went on to interpret it that way, claiming his theorem "to show that the Bayesian rationality postulates, together with a very natural Pareto optimality requirement, logically entail a utilitarian ethic" (Harsanyi, 1978, p. 226)—surprisingly, "even if interpersonal utility comparisons are not admitted" (228).

This is puzzling, not least because utilitarianism would typically be taken to require "everybody to count for one, nobody for more than one" (Mill, 1863). The conclusion of Harsanyi's theorem is compatible with the assignment of different weights to different people. But let us set this worry aside. What puzzles me is how any form of broadly utilitarian aggregation, weighted or otherwise, could be possible without interpersonal comparisons of well-being. If we cannot even compare the well-beings of different individuals, how can we sensibly add them up, even after weighting them? The orthodox answer of social choice theory is that this is impossible: "weighted utilitarianism requires interpersonal comparisons of utility gains and losses" (Blackorby et al., 2008, p. 138).

The puzzle deepens when we compare Harsanyi's theorem to an even more foundational result of social choice theory: Arrow's impossibility theorem. Arrow (1950) showed that a social choice procedure that delivers an ordering for all possible arrangements of individual preferences, respects the unanimous preferences of individuals, and ranks any two alternatives only by considering individuals' preferences between those alternatives must be a dictatorship, in which one person's preferences always prevail. Arrow worked in a purely ordinal framework in which utilitarianism could not even be formulated. But Sen (1970) and others have extended the result to informational settings that include cardinally measurable utilities. And the standard lesson of these results is that "admitting cardinality of utilities *without* interpersonal comparisons does not change Arrow's impossibility theorem at all" (Sen, 1999, p. 357).

Harsanyi admits that the "use of cardinal utilities is insufficient to enable us to avoid Arrow's Impossibility Theorem" (Harsanyi, 1979, p. 303). And he seems to accept Arrow's conditions when formulated in terms of individual utility functions. How, then, could he have possibly derived a weighted utilitarian social welfare function in a way that "does not depend on the possibility of interpersonal utility comparisons" (Harsanyi, 1979, p. 294)? This is the question I want to explore in this paper.

The general consensus appears to be that, at least when Harsanyi's conditions are supplemented in such a way that they support weighted utilitarianism, his result does require interpersonal comparisons. Some argue that the possibility of interpersonal comparisons is presupposed by one of Harsanyi's premises (Broome, 1991). Others argue that it is a surprising conclusion of Harsanyi's theorem (Jeffrey, 1971; Mongin, 1994). I will argue instead that Harsanyi was right: his theorem and its weighted-utilitarian conclusion do not require interpersonal comparisons of well-being. The key idea is to understand Harsanyi's weights not as real numbers but rather as *dimensional constants*. Defending this thesis will require us to rethink some core ideas in the theory of social choice and welfare. In doing so, I will argue that the standard lesson drawn from Sen's extension of Arrow's theorem is mistaken. But, first, we need to get Harsanyi's theorem on the table.

2 | HARSANYI'S THEOREM

Consider a set of outcomes $X = \{x_1, \ldots, x_m\}$. A *lottery* over these outcomes assigns an ("objective") probability to each outcome in such a way that the probabilities sum to one. We can harmlessly treat outcomes as if they were lotteries since, for each outcome, there is a degenerate lottery that guarantees that outcome. Where p is a lottery and x_j is an outcome, p_j is the probability assigned by p to x_j.

We have a fixed population of individuals, numbered $1, \ldots, n$. For each individual i, there is a relation $\succ_i$ of being better for i—i's *betterness relation*—over the set of all lotteries. Let $\succ$ (no subscript) denote the "overall" betterness relation. (Harsanyi interprets $\succ_i$ as i's preference relation and $\succ$ as a social preference relation. I follow Broome 1991, Dreier 2004, and others in reinterpreting these as betterness relations for the sake of generality. But the question of this paper is especially important on Harsanyi's preference-theoretic interpretation, because it is especially controversial whether interpersonal comparisons are possible in such a framework; see, e.g., Greaves and Lederman, 2018; Hausman, 1995.)

Harsanyi's theorem has three premises. The first is that each individual's betterness relation has an expected utility representation. A real-valued function $u_i(\cdot)$ *represents* i's betterness relation just in case it assigns higher numbers to lotteries that are better for i—i.e., for any lotteries p and q : $p \succ_i q$ iff $u_i(p) > u_i(q)$. We call $u_i(\cdot)$ a *utility function* for i's betterness relation. An *expectational* utility function is one with the following property: the number it assigns to any lottery p is the expected value of $u_i(\cdot)$ over p's outcomes (which, recall, we are treating as degenerate lotteries). That is,

$$u_i(p) = p_1 u_i(x_1) + \ldots + p_m u_i(x_m) \tag{1}$$

An expectational utility function is unique up to positive affine transformation—i.e., multiplication by a positive number and addition of a constant. If $u_i(\cdot)$ is an expectational utility function that represents i's betterness relation, then for any positive α and any β, i's betterness relation can also be represented by $v_i(\cdot) = \alpha u_i(\cdot) + \beta$, which will also be expectational.[1]

Harsanyi's second premise is that the overall betterness relation has an expected utility representation. This means that there is a *social welfare function* $W(\cdot)$, unique up to positive affine transformation, that assigns higher numbers to better lotteries—i.e., $p \succ q$ iff $W(p) > W(q)$—and is expectational:

$$W(p) = p_1 W(x_1) + \ldots + p_m W(x_m) \tag{2}$$

Harsanyi's third premise is

Strong Pareto If one lottery is at least as good for each person as another, then it is at least as good overall. If, in addition, it is better for someone, then it is better overall.

Harsanyi's theorem is that, if these premises are true, there are positive real numbers (weights) $k_1, \ldots, k_n$ such that, for any lottery p:

$$W(p) = k_1 u_1(p) + \ldots + k_n u_n(p) \tag{3}$$

This means that the overall betterness relation can be represented as maximizing the expectation of a weighted sum of individual utilities, where those utilities expectationally represent the individuals' betterness relations.

We now have Harsanyi's aggregation theorem on the table. Notice that none of the premises seems to require interpersonal comparisons of well-being. For reasons that will be explained in the next few sections, however, equation (3) does not necessarily express a *utilitarian* principle of aggregation—and not just because the weights can differ by person. To get closer to a utilitarian conclusion, we need some additional assumptions.

[1] Harsanyi uses Marschak (1950)'s variation on the axioms of von Neumann and Morgenstern (1947). It does not matter for his theorem which axiomatization is used, so long as it provides an expected utility representation.

3 | GOODNESS AND UTILITY

A (weighted) utilitarian is typically thought to believe that there is some quantity—goodness for a person, or well-being—the (weighted) sum of which ought to be maximized. What is the relation between this quantity and an expected utility representation of an individual's betterness relation?

To mark this distinction more clearly, let "$\mathbf{g_i}(p)$" denote the goodness of lottery p for person i—for brevity, i-goodness. I use boldface to indicate that this symbol does not designate a *number*. It is important to distinguish between a dimensioned quantity and the number used by a particular scale to represent that quantity.[2] My mass is 75 kg. This quantity is not a number. What is a number is the ratio of this quantity to the mass of the standard kilogram—i.e., the number assigned by the kilogram scale to my mass (75). Something's goodness for a person is not a number. The question is whether, and in what sense, it can be represented by a number, such as the value of an expectational utility function.

We can assume that, for any lotteries p and q, p's goodness for i is greater than q's just in case p is better for i than q: $\mathbf{g_i}(p) > \mathbf{g_i}(q)$ iff $p \succ_i q$. Harsanyi's first premise was that $\succ_i$ has an expected utility representation. So there is a utility function $u_i(\cdot)$ that represents $\succ_i$, in the sense that $p \succ_i q$ iff $u_i(p) > u_i(q)$. Thus, $u_i(\cdot)$ assigns a higher number to one of two lotteries just in case its goodness for i is greater: $\mathbf{g_i}(p) > \mathbf{g_i}(q)$ iff $u_i(p) > u_i(q)$. In this sense, $u_i(\cdot)$ provides an *ordinal* scale of i-goodness.

Intuitively, we can make comparisons not only of i-goodness, but also of *differences* in i-goodness. One thing might be much better for you than another, whereas a third thing might be only slightly better. Let "$\mathbf{g_i}(p) - \mathbf{g_i}(q)$" denote the difference in i-goodness between p and q—i.e., *how much* better p is for i than q. (On Harsanyi's preference-theoretic interpretation, this can be thought of as the strength of i's preference between p and q.) Harsanyi's premises do not imply that there are such differences, but let's assume that there are. What is the relation between differences in i-goodness and differences in u_i-utility?

We might think that differences in u_i-utility represent differences in i-goodness, in at least the following sense:

$$\mathbf{g_i}(p) - \mathbf{g_i}(q) > \mathbf{g_i}(r) - \mathbf{g_i}(s) \Leftrightarrow u_i(p) - u_i(q) > u_i(r) - u_i(s) \tag{4}$$

But this does not follow from the assumptions made so far. We have assumed, with Harsanyi, that $\succ_i$ has an expected utility representation. But this doesn't mean, and the axioms of expected utility theory do not imply, that $\succ_i$ can *only* be represented by an expectational utility function. To see this, suppose that we apply an order-preserving but nonaffine transformation to each $u_i(\cdot)$. For example, let $v_i(\cdot)$ be the square of $u_i(\cdot)$: $v_i(p) = [u_i(p)]^2$ for each lottery p. (Assume that $u_i(\cdot)$ only assigns nonnegative numbers.) $v_i(\cdot)$ still represents the ordering $\succ_i$, in the sense that $p \succ_i q$ iff $v_i(p) > v_i(q)$. But this representation is not expectational. It instead has the less convenient form:

$$v_i(p) = \left(p_1\sqrt{v_i(x_1)} + \ldots + p_m\sqrt{v_i(x_m)}\right)^2 \tag{5}$$

None of our assumptions so far rule out the possibility that $v_i(\cdot)$ represents differences in i-goodness, in the following sense:

$$\mathbf{g_i}(p) - \mathbf{g_i}(q) > \mathbf{g_i}(r) - \mathbf{g_i}(s) \Leftrightarrow v_i(p) - v_i(q) > v_i(r) - v_i(s) \tag{6}$$

[2] I try to remain neutral about controversial issues in the metaphysics of quantities (see, e.g., Dasgupta, 2013; Eddon, 2013; Sider, 2020; Wolff, 2020). I assume that any plausible theory of quantities should be compatible with the claims I want to make about them here (or suitable reinterpretations of those claims).

Suppose that (6) is true. Then (4) cannot also be true: u_i-differences and v_i-differences cannot both represent differences in i-goodness. Consider the following utility assignments:

	u_i	v_i
w	1	1
x	3	9
y	4	16
z	5	25

The difference between $u_i(x)$ and $u_i(w)$ is greater than the difference between $u_i(z)$ and $u_i(y)$. But the difference between $v_i(x)$ and $v_i(w)$ is less than the difference between $v_i(z)$ and $v_i(y)$. The two scales cannot both represent differences in i-goodness.

To see why this matters for Harsanyi's purposes, suppose that (6) is true, as is compatible with Harsanyi's premises. Suppose we accept the conclusion of Harsanyi's theorem, that overall betterness can be represented by an expectational social welfare function of the form stated in equation (3): $W(p) = k_1u_1(p) + \ldots + k_nu_n(p)$. We can rewrite this equation in terms of v_i-utilities:

$$W(p) = k_1\sqrt{v_1(p)} + \ldots + k_n\sqrt{v_n(p)} \tag{7}$$

In terms of v_i-utilities, this is not a weighted utilitarian representation but a weighted *prioritarian* one: each person's v_i-utility has diminishing marginal value with respect to social welfare. If v_i-differences represent differences in goodness for i, then Harsanyi's theorem would imply that a difference in goodness for i matters more the worse off i is. To rule out this prioritarian interpretation, Harsanyi needs some reason to privilege the family of expectational representations rather than nonaffine transformations thereof. This is the influential Sen (1977)–Weymark (1991) critique of Harsanyi's theorem (for responses, see Fleurbaey and Mongin 2016; Greaves 2017; Risse 2002).

To address this problem, I will attribute another assumption to Harsanyi, which Broome (1991) calls

Bernoulli's Hypothesis For any lotteries p and q, p is better for i than q just in case p's expectation of goodness for i is greater than q's.

This principle strengthens Harsanyi's first premise, that each person's betterness relation has an expected utility representation. Bernoulli's Hypothesis implies that $u_i(\cdot)$ represents differences in i-goodness, so that (4) is true and (6) is false. More than this, it implies that there are meaningful *ratios* of goodness differences, for any given person. These ratios are represented by expectational utility functions, in the sense that

$$\frac{\mathbf{g_i}(p) - \mathbf{g_i}(q)}{\mathbf{g_i}(r) - \mathbf{g_i}(s)} = \frac{u_i(p) - u_i(q)}{u_i(r) - u_i(s)} \tag{8}$$

Such ratios are preserved by any positive affine transformation of $u_i(\cdot)$. In this sense, $u_i(\cdot)$ is a *cardinal* scale of i-goodness. For example, the Celsius scale is a cardinal scale of temperature. Let $\mathbf{T}(x)$ denote the temperature of an object x, and let $T_C(x)$ denote the number that represents x's temperature on the Celsius scale. The ratio of two differences in degrees Celsius is the ratio of two differences in temperature:

$$\frac{\mathbf{T}(a) - \mathbf{T}(b)}{\mathbf{T}(c) - \mathbf{T}(d)} = \frac{T_C(a) - T_C(b)}{T_C(c) - T_C(d)} \tag{9}$$

And this ratio is preserved by any positive affine transformation of the Celsius scale (e.g., the Farenheit scale).

The addition of Bernoulli's Hypothesis rules out the prioritarian interpretation of Harsanyi's conclusion. But how close does it take us to utilitarian aggregation?

4 | SUMMATION

Broome (1991) claims that, given Bernoulli's Hypothesis, the conclusion of Harsanyi's theorem implies

Summation For any lotteries p and q, p is better than q just in case the sum of the differences in goodness for each person between p and q is positive:

$$\left[\mathbf{g_1}(p) - \mathbf{g_1}(q)\right] + \ldots + \left[\mathbf{g_n}(p) - \mathbf{g_n}(q)\right] > 0 \tag{10}$$

There are two reasons to doubt this. One is the removal of Harsanyi's weights. The other is that differences in goodness *for different people* are being summed. Let's start with the second.

We have assumed that intrapersonal comparisons of goodness, differences in goodness, and ratios of differences in goodness are meaningful—indeed, more than this, that ratios of differences in goodness for a person are real numbers. But we have not assumed the meaningfulness of *interpersonal* comparisons of any kind. And Summation could hardly be true unless we can compare differences in goodness for different people. We cannot add up quantities that cannot even be compared. There is no quantity that is the sum of my mass and your height because these are quantities of different dimensions. We can of course add the number that represents my mass on the kilogram scale and the number that represents your height on the inches scale; this operation is not the addition of mass and height, but of relatively arbitrary dimensionless numbers. Summation presupposes that we can add *differences in goodness* across people; goodness is not a number, nor are differences in goodness. Such quantities must be interpersonally comparable in order for this operation to make sense.

Summation presupposes that interpersonal comparisons of differences in goodness are meaningful. But we have not explicitly assumed that such comparisons are meaningful. So how could Broome have obtained Summation from Harsanyi's conclusion and Bernoulli's Hypothesis?

Broome suggests that the possibility of interpersonal comparisons is hidden in Harsanyi's premises. Harsanyi's second premise was that the overall betterness relation has an expected utility representation. This implies the *completeness* of the overall betterness relation: that, for any lotteries p and q, either p is better than q, q is better than p, or p and q are equally good. In other words, there are no "gaps" in overall betterness. But, according to Broome (1991, p. 220), "If one person's good cannot be compared with another's, then the general betterness relation will simply not be complete." So, Broome claims, Harsanyi has assumed the possibility of interpersonal comparisons after all.

If Broome means that the absence of interpersonal comparisons would *by itself* rule out completeness, that is clearly false. Without further assumptions, the overall betterness relation could have nothing to do with individual betterness relations. It could rank lotteries in terms of expected quantities of cheese; the absence of interpersonal comparisons of goodness would be no barrier to the completeness of the resulting ranking. Even if we require overall betterness to be a function of individual betterness relations, by imposing a Pareto condition, it is still wrong. One counterexample is a *majoritarian* betterness relation: one alternative is better than another just in case the first is better for at least as many people as the second is; two alternatives are equally good just in case each is better than the other for the same number of people. Or consider a *serial dictatorship*: person 1's betterness relation determines which of any two alternatives is better overall unless they are equally good for her, in which case person 2's betterness relation determines it unless they are equally good for her, and so on; two alternatives are equally good just in case they are equally good for each person. These rules violate other conditions required for an expected utility representation, but they satisfy completeness and are paradigm cases of rules that do not require interpersonal comparisons of well-being. They do involve comparisons of the *moral significance* of different people's well-beings: the majoritarian treats every gain or loss in well-being as having equal moral significance, no matter whose well-being it is or how much is gained or lost; a dictatorship treats gains and losses for the dictator as having greater moral significance than gains and losses for everyone else. But these are not interpersonal comparisons *of well-being*: the majoritarian need not insist that every gain or loss to each person is of the same size; dictatorship does not entail that the dictator gains more well-being from any change that is better for her than anyone else could possibly gain or lose. Since these rules satisfy completeness but do not require interpersonal comparisons of well-being, it is not clear why completeness in particular should be picked out as smuggling in the possibility of interpersonal comparisons. Nor does this possibility seem to be implied by any of Harsanyi's other premises. So it is hard to see how Harsanyi's premises and Bernoulli's Hypothesis could together imply Summation.[3]

We will return to this issue soon. For now, let us turn to our other question about Broome's purported derivation of Summation: the disappearance of Harsanyi's weights.

Here is what happens to Harsanyi's weights in Broome's attempt to derive Summation. Start with the conclusion of Harsanyi's theorem: overall betterness can be represented as maximizing a weighted sum of utilities representing individual betterness relations, as in equation (3): $W(p) = k_1u_1(p) + \ldots + k_nu_n(p)$. Broome's strategy proceeds in two steps. The first is to normalize each person's utility function so that the weights are equal. Recall that an expected utility representation is unique up to positive affine transformation. And observe that, for each individual i, the function $v_i(\cdot) = k_iu_i(\cdot)$ is a positive affine transformation of $u_i(\cdot)$. We can rewrite equation (3) in terms of an *unweighted* sum:

$$W(p) = v_1(p) + \ldots + v_n(p) \tag{11}$$

The second step is to move from (11) to Summation via Bernoulli's Hypothesis. Bernoulli's Hypothesis implies that each person's expectational utility function represents her good on a cardinal scale. So "the total of people's utilities will measure the total of people's good" (Broome, 1991, p. 218). Since the sum of these utilities represents the overall betterness relation, equation (11) seems to imply Summation.

These two steps, however, rely on inconsistent assumptions about interpersonal comparisons. Suppose first that interpersonal comparisons of goodness differences are not meaningful. Then the second step of the argument is not valid. Equation (11) entails that there is some scale of goodness for person 1, some scale of goodness for person 2, and so on, such that the overall betterness relation can be represented by the sum of the numbers on these scales. But if interpersonal comparisons are not meaningful, then we cannot infer from this representation that overall betterness is represented by the sum of goodness for different people. Consider an analogy. There is some scale of height, some scale of mass, and some scale of temporal duration such that the number that represents your height on that scale plus the number that represents your mass on that scale equals the number that represents

[3] I think the later Broome (2004) would agree, since he distinguishes differences in goodness for a person from differences in overall good, and recognizes a need to justify the possibility of interpersonal comparisons.

your age on that scale. We cannot conclude from this that your age is, or is even represented by, the sum of your height and your mass, since there is no such quantity as the sum of your height and your mass. Moving from equation (11) to Summation is like that. It takes a sum of numbers, each of which represents a magnitude of some distinct quantity on independent scales, to represent the sum of the various quantities represented, which is not even well-defined.

Suppose next that interpersonal comparisons of goodness differences are meaningful, and in particular that they are represented by the expectational utility functions $u_1(\cdot), \ldots, u_n(\cdot)$. Then the first step of the argument is not valid. For if the weights in equation (3) are distinct—i.e., if $k_i \neq k_j$ for some i and j—then the transformations used to cancel out the weights will use different scale factors. So the interpersonal comparisons of goodness differences represented by $u_1(\cdot), \ldots, u_n(\cdot)$ will change when those representations are normalized to yield an unweighted sum in terms of $v_1(\cdot), \ldots, v_n(\cdot)$. Consider another analogy. Suppose we have a collection of two apples and ten oranges, where each apple has the same mass and each orange has the same mass. Let $m_g(a)$, $m_g(o)$, and $m_g(c)$ denote the mass in grams of a single apple, a single orange, and the collection of apples and oranges respectively:

$$m_g(c) = 2m_g(a) + 10m_g(o)$$

Suppose we cancel out the "weights" by converting the scale of apple-mass to half-grams and the scale of orange-mass to decigrams:

$$m_g(c) = m_{g/2}(a) + m_{g/10}(o)$$

We obviously cannot infer from this that the mass of the collection is the sum of the mass of one apple and the mass of one orange, because we are no longer representing apple-mass and orange-mass on the same scale. Broome's strategy is something like this, in moving from equation (3) to (11). It takes a bunch of numbers, each of which represents a magnitude of some quantity (goodness for a person) on the same scale, and converts those numbers independently onto values of different scales. But the new utility scales will not in general preserve the interpersonal comparisons represented by the original ones, so we cannot take the new unweighted representation to represent the same thing as the original weighted representation.

As far as I can see, there is no way to get Summation from Bernoulli's Hypothesis and the conclusion of Harsanyi's theorem alone. For if the expectational utility functions $u_1(\cdot), \ldots, u_n(\cdot)$ convey interpersonal information, that information will not be preserved by individual-specific rescalings that deliver an unweighted representation. And if they don't convey such information, then we cannot infer that the sum of utilities represents a sum of goodness across people, since the latter will not even be meaningful.

5 | INDEPENDENCE OF IRRELEVANT ALTERNATIVES

What does Bernoulli's Hypothesis add to Harsanyi's conclusion, then, if it does not get us to Summation? A natural thought would be that it implies

Number-Weighted Summation There are positive real numbers $k_1, \ldots, k_n$ such that, for any lotteries p and q: p is better than q just in case the weighted sum of the differences in goodness for each person between p and q is positive:

$$k_1\left[\mathbf{g_1}(p) - \mathbf{g_1}(q)\right] + \ldots + k_n\left[\mathbf{g_n}(p) - \mathbf{g_n}(q)\right] > 0 \quad (12)$$

There are two reasons to doubt this.

First, each term in this summation—$k_i\left[\mathbf{g_i}(p) - \mathbf{g_i}(q)\right]$—is a difference in goodness for i. So this doctrine, too, is meaningful only if differences in goodness for different people are comparable. But, again, we have not assumed that they are.

Second, none of our assumptions so far imply that the weight of each person is fixed, in the following sense. Harsanyi's conclusion is that, given an expected utility representation of person 1's betterness relation, an expected utility representation for person 2's betterness relation, and so on, there are weights $k_1, \ldots, k_n$ such that the overall betterness relation can be represented as maximizing the sum of utilities so weighted. But suppose we swap some individual i's utility function with a positive affine transformation that shrinks or expands the differences in i's utility between outcomes, leaving everyone else's utility function unchanged. Then if we used the same set of weights, the original betterness ordering may fail to maximize the sum of utilities so weighted. To preserve the overall betterness ordering, we'd need to change the social welfare function by adjusting the weight on i's utilities. This seems strange, because there is supposed to be nothing special about i's original utility function, and—if interpersonal comparisons of well-being are not assumed—no important connections between the utility functions of different individuals. As Weymark (1991, p. 281) puts it, "Simply because new utility representations have been adopted, society behaves as if it has changed the way it aggregates individual utilities."

Both problems are addressed by an extension of Harsanyi's theorem due to Mongin (1994). Mongin proves a version of Harsanyi's theorem for what Sen (1970) calls social welfare function*als*. A social welfare functional assigns an overall betterness ordering to each possible profile of utility functions. A *profile* U is a list of utility functions, one for each individual: $U = \left(u_1(\cdot), \ldots, u_n(\cdot)\right)$. We are concerned here only with profiles of expectational utility functions. If we specify each person's utility function, the social welfare functional will deliver a ranking of lotteries. If we feed it a different profile of utility functions, it will give us a (possibly different) ranking. The social welfare functional does not itself, however, change from profile to profile.[4]

There are two reasons to want a social welfare functional rather than a mere social welfare function such as Harsanyi's. The first is that we want a way to compare alternatives that does not depend on how the individual betterness facts are represented. Harsanyi's conclusion does not provide that, as we have seen.

The second reason has to do with *non*-representational changes in the individual betterness facts. This reason is most compelling in the original preference-theoretic framework of Harsanyi. People can change their preferences, and it would be nice to have a way to compare alternatives that is invariant with respect to such changes. Harsanyi's theorem does not provide that, since the weights can change when new utility functions are used. This consideration is harder to translate when we are concerned with individual betterness relations, since we might think the individual betterness facts to be necessary truths. But, on some theories of prudential value, what is good for a person can depend on some contingent or temporary matters—e.g., a person's values, projects, or relationships. It seems desirable to have a way of comparing alternatives that is compatible with such views and is robust to changes in such matters. We might also want to consider multiple profiles given our uncertainty about the individual betterness facts.

Harsanyi stated his theorem in a single-profile framework. So the weights in his conclusion depend on the particular utility functions we use to represent each person's betterness ordering, and on what that ordering happens to be. To extend Harsanyi's theorem to the multi-profile setting of social welfare functionals, Mongin adds an *independence of irrelevant alternatives* condition. To introduce

[4] For helpful overviews of the social welfare functional literature, see Adler (2019), Bossert and Weymark (2004), d'Aspremont and Gevers (2002), and Weymark (2016).

this condition, take any two profiles $U = \left(u_1(\cdot), \ldots, u_n(\cdot)\right)$ and $V = \left(v_1(\cdot), \ldots, v_n(\cdot)\right)$. Say that U and V *coincide* on two alternatives just in case each person's utility function assigns the same value to those alternatives in both profiles. That is, U and V coincide on p and q just in case, for every individual i: $u_i(p) = v_i(p)$ and $u_i(q) = v_i(q)$. Say that the overall betterness relations assigned to U and V *agree* on two alternatives just in case they rank those alternatives the same way—that is, $p \succ^U q$ iff $p \succ^V q$, where $\succ^U$ and $\succ^V$ denote the overall betterness relations assigned by the social welfare functional to U and V respectively. According to

Independence of Irrelevant Alternatives If two profiles coincide on two alternatives, then the overall betterness relations assigned to those two profiles must agree on those two alternatives.

This means that changes to people's utility functions that preserve the values assigned to particular alternatives cannot change—because they are irrelevant to—the social ranking of those particular alternatives. In the context of Harsanyi's theorem, Independence of Irrelevant Alternatives prohibits the weights from depending on the particular profile of utility functions to be evaluated. For if the weights differed by profile, then one could find a pair of alternatives on which two profiles coincide, with the weighted sums of those pairs differing between the two profiles. So the overall betterness relations assigned to those two profiles would disagree on those alternatives, since Harsanyi's other conditions require overall betterness to be representable as maximizing the weighted sum of utilities. Independence of Irrelevant Alternatives rules out this possibility.

Let us, for now, treat Independence of Irrelevant Alternatives as if it were one of Harsanyi's official premises. He seems to accept it in Harsanyi (1979, sec. 6). We will revisit this addition in section 9. Until then, by "Harsanyi's premises" I mean his original premises plus Bernoulli's Hypothesis and Independence of Irrelevant Alternatives.

When added to Harsanyi's other conditions, Independence of Irrelevant Alternatives allows Mongin to derive a social welfare functional with the following property. There are positive real numbers $k_1, \ldots, k_n$ such that, for any profile U, and any lotteries p and q:

$$p \succ^U q \Leftrightarrow k_1 \left[u_1(p) - u_1(q)\right] + \ldots + k_n \left[u_n(p) - u_n(q)\right] > 0 \tag{13}$$

The weights, importantly, do not vary from profile to profile.

If we assume Bernoulli's Hypothesis, then the u_i-difference for each person represents the difference in goodness for each person. And the weights on these differences do not depend on the particular utility functions we use. This seems to entail Number-Weighted Summation. This is puzzling, though. Number-Weighted Summation requires interpersonal comparisons of differences in goodness. And we have not assumed such comparisons to be possible.

Mongin and d'Aspremont (1998) suggest that the possibility of interpersonal comparisons is a surprising *conclusion* of the theorem, not (as Broome claimed) an assumption of it. To explain why they think this, we need to undertand how interpersonal comparability is typically understood in the framework of social welfare functionals.

6 | INVARIANCE CONDITIONS

What does it mean for interpersonal comparisons of differences in goodness to be *meaningful*? An important tradition in the theory of measurement understands meaningfulness in terms of *invariance*. As

Roberts (1984, p. 71) puts it, “A statement involving numerical scales is meaningful if and only if its truth (or falsity) remains unchanged under all admissible transformations of all the scales involved.” Suppose, for example, that the temperatures of a and b are such that the Celsius scale assigns a number to a that is twice the number assigned to b: $T_C(a) = 2T_C(b)$. Can we infer that the temperature of a is twice the temperature of b: $\mathbf{T}(a) = 2\mathbf{T}(b)$? No, because the Farenheit scale will assign different values to a and b so that the statement is not true on that scale: $T_F(a) \neq 2T_F(b)$. Since the Farenheit scale is an admissible transformation of the Celsius scale, we are supposed to conclude that statements about temperature ratios are not meaningful. Statements about ratios of temperature *differences* are, though, because the numerical representations of such ratios are invariant under all admissible transformations. This is what makes Celsius and Farenheit cardinal scales.

The idea of invariance as a criterion for meaningfulness is typically made precise, in social choice theory, by conditions that require invariance of the social welfare functional to certain kinds of transformations on utility profiles. Suppose we have two profiles of utility functions $U = (u_1(\cdot), \ldots, u_n(\cdot))$ and $V = (v_1(\cdot), \ldots, v_n(\cdot))$. The idea is that, if U and V are related in a certain way, then they contain the same meaningful information about the good of individuals, in much the same way that the Farenheit and Celsius scales represent the same facts about temperature. If U and V are so related, then any differences between them are mere artefacts of the utility representation, much like ratios between numbers assigned by the Celsius scale. The social welfare functional should therefore assign the same overall betterness relation to U and V.

For example, suppose that each person’s good is only ordinally measurable and that we cannot make interpersonal comparisons of any kind. If i-goodness is only ordinally measurable, the only significant feature of $u_i(\cdot)$ is the order in which it ranks alternatives. This feature is preserved by any order-preserving transformation of $u_i(\cdot)$. And if we cannot compare the good of different individuals, then admissible transformations of different people’s utility functions need not preserve any comparisons between individuals, since such comparisons are not supposed to be meaningful. So we should be able to apply different transformations for different people without changing the overall betterness ordering.

To make this precise, consider a profile $U = (u_1(\cdot), \ldots, u_n(\cdot))$. And consider a list of order-preserving transformations $\phi = (\phi_1(\cdot), \ldots, \phi_n(\cdot))$. Each ϕ_i takes a utility function and returns another utility function that preserves the ordering of alternatives for i. Let $\phi(U) = (\phi_1(u_1(\cdot)), \ldots, \phi_n(u_n(\cdot))$. If utility is only ordinally measurable and not interpersonally comparable, then U and $\phi(U)$ are *informationally equivalent*. If the social welfare functional must assign the same betterness relation to informationally equivalent profiles, then we have

Intrapersonal Ordinal Invariance For any profiles U and V, if there is some list of order-preserving transformations $\phi = (\phi_1, \ldots, \phi_n)$ such that $V = \phi(U)$, then $\succ^U$ and $\succ^V$ must agree on all alternatives.

This means that, if two profiles are such that each person’s utility function in one profile is some (possibly distinct for each person) order-preserving transformation of her utility function in the other profile, then those two profiles must be assigned the same overall betterness relation.

Intrapersonal Ordinal Invariance is extremely restrictive. It yields the informational framework of Arrow (1950), who worked with profiles of preference orderings rather than utility functions. In the terminology of social welfare functionals and individual betterness relations, Arrow’s theorem can be stated as follows. Consider a social welfare functional that assigns an overall betterness ordering to every possible profile of utility functions. Suppose that this social welfare functional satisfies Intrapersonal Ordinal Invariance, Independence of Irrelevant Alternatives, and

Weak Pareto If one alternative is better for each person than another, then the first must be better than the second.

Then, according to Arrow's theorem, the social welfare functional must be dictatorial: there must be some individual such that, whenever an alternative is better for her, it is better overall.

Suppose instead that goodness for each person is *cardinally* measurable, but still not interpersonally comparable. A cardinal scale is unique up to positive affine transformation. Consider again profile $U = (u_1(\cdot), \ldots, u_n(\cdot))$. And consider a list of functions $\phi = (\phi_1(\cdot), \ldots, \phi_n(\cdot))$, where this time each ϕ_i is a positive affine transformation: for each i, there is some positive α_i and some β_i—each of which may be different for each person—such that $\phi_i(u_i) = \alpha_i u_i + \beta_i$. If each person's good is only cardinally measurable and not interpersonally comparable, then U and $\phi(U)$ would seem to contain the same meaningful information. This yields

Intrapersonal Cardinal Invariance For any profiles U and V, if there is some list of positive affine transformations $\phi = (\alpha_1 u_1 + \beta_1, \ldots, \alpha_n u_n + \beta_n)$ such that $V = \phi(U)$, then $\succ^U$ and $\succ^V$ must agree on all alternatives.

Can we avoid Arrow's impossibility by weakening the invariance requirement from Intrapersonal Ordinal Invariance to Intrapersonal Cardinal Invariance? Sen (1970) showed that the answer is no: Intrapersonal Cardinal Invariance and Independence of Irrelevant Alternatives together imply Intrapersonal Ordinal Invariance, so the addition of cardinal information makes no difference in avoiding dictatorship.

We are now in a position to understand Mongin and d'Aspremont's suggestion with which we ended section 5. A weighted utilitarian social welfare functional violates Intrapersonal Cardinal Invariance. This can be seen by picking a set of weights and then blowing up one person's utility function, leaving others' unchanged, so that maximizing the weighted sum leads to a different ordering. But if interpersonal comparisons of goodness differences were not meaningful, then Intrapersonal Cardinal Invariance would have to be true. So Harsanyi's premises, when translated to the multi-profile setting, together seem to imply that such comparisons must be possible.[5]

I now want to argue, though, that this impression is misleading: Intrapersonal Cardinal Invariance does not follow from the view that well-being is cardinally measurable but not interpersonally comparable.

7 | LUCE'S PRINCIPLE

We have seen that, when translated to the multi-profile setting of social welfare functionals, Harsanyi's conditions violate Intrapersonal Cardinal Invariance. My question is whether Harsanyi's conditions

[5] Here is another way to see the point. Suppose we deny the possibility of interpersonal comparisons and therefore try to add Intrapersonal Cardinal Invariance to Harsanyi's other conditions. These include Independence of Irrelevant Alternatives and Strong Pareto, which entails Weak Pareto. A simple extension of Arrow's theorem shows that the only social welfare functional compatible with Intrapersonal Cardinal Invariance, Independence of Irrelevant Alternatives, and Strong Pareto is a serial dictatorship (Gevers 1979; Luce and Raiffa 1957). But a serial dictatorship is incompatible with expected utility theory for the overall betterness relation (specifically, with the continuity axiom). Since we cannot add Intrapersonal Cardinal Invariance to Harsanyi's other conditions, this seems to show that Harsanyi's conditions entail interpersonal comparability.

must therefore imply, as Mongin and others have assumed, that interpersonal comparisons of differences in goodness are possible. I argue that they don't.

This may seem downright confused. If each person's utility function represents her betterness relation on a cardinal scale, then any positive affine transformation of that scale should represent her betterness relation just as well; and if differences in goodness for different people cannot be compared, then it should hardly matter whether we apply the same or different admissible transformations to different people's utility functions. To insist otherwise would be like insisting that, upon switching from grams to kilograms when measuring mass, one must also switch from meters to kilometers when measuring distance. And that would be absurd. This is why Intrapersonal Cardinal Invariance seems to follow from the absence of interpersonal comparisons, and therefore why Harsanyi's violation of Intrapersonal Cardinal Invariance seems to imply that interpersonal comparisons must be meaningful.

More generally, it seems that if our numerical representation of some relation is unique up to certain kinds of transformations, then it should not matter which numerical representation within that family of transformations we use. This is because all representations within that family preserve the same meaningful statements about that relation. If we explain the meaningfulness of various comparisons in terms of the admissible class of transformations in this way, it seems to follow that the social welfare functional must be invariant to that admissible class of transformations. Otherwise, the overall betterness relation delivered by the social welfare functional would seem to objectionably depend on an arbitrary (and indeed meaningless) artefact of how individuals' betterness relations are represented. So, if Harsanyi's premises entail that the social welfare functional is not invariant to independent affine transformations of different people's utility functions, then this must be because they require interpersonal comparisons to be meaningful.

This reasoning implicitly appeals a more general principle due to Luce (1959, p. 85). According to

Luce's Principle "Admissible transformations of one or more of the independent variables shall lead … only to admissible transformations of the dependent variables."

To see what this means, consider a law that reports the value of some dependent variable y in terms of independent variables $x_1, \ldots, x_n$:

$$y = f(x_1, \ldots, x_n) \tag{14}$$

For example, y might be the value of a social welfare function that represents overall betterness, and the x's might be the values of individual utility functions that represent goodness for each person. Now suppose, for each i, that ϕ_i is an admissible transformation of x_i—in our case, a positive affine transformation. Then, according to Luce's Principle, there must be some admissible transformation ψ of the dependent variable such that

$$\psi(y) = f(\phi_1(x_1), \ldots, \phi_n(x_n)) \tag{15}$$

And so, by equation (14):

$$\psi(f(x_1, \ldots, x_n)) = f(\phi_1(x_1), \ldots, \phi_n(x_n)) \tag{16}$$

When the independent variables are values of utility functions that represent interpersonally noncomparable differences in goodness, and the dependent variable is the value of a social welfare function that represents the overall betterness relation, this implies Intrapersonal Cardinal Invariance, since an admissible transformation of the social welfare function must at least be order-preserving.

Luce's Principle, however, is false, at least in this unqualified form—as Luce (1962) came to agree. Consider some radioactive material whose mass decays exponentially with time (Rozeboom, 1962). Suppose that the material weighs 10 kg at some initial time, and that its mass in kilograms m at any time t minutes later is the value of the following function:

$$m(t) = 10\mathrm{e}^{-t/2} \tag{17}$$

Since mass and temporal duration are both ratio-scale measurable, Luce's principle implies that a similarity transformation applied to the measure of time should yield some similarity transformation in the measure of mass. That is, for any positive α, there must be some positive β such that, for any value of t:

$$\beta m(t) = 10\mathrm{e}^{-\alpha t/2} \tag{18}$$

And so, by equation (17):

$$\beta 10\mathrm{e}^{-t/2} = 10e^{-\alpha t/2} \tag{19}$$

or more simply:

$$\beta = \mathrm{e}^{(1-\alpha)t/2} \tag{20}$$

But the only α for which there is such a fixed β is α=1—i.e., the identity transformation. In all other cases, the value of $\mathrm{e}^{(1-\alpha)t/2}$ varies with the time variable t. So, if we fix a transformation $\alpha \neq 1$—say, 1/60, as when converting from minutes to seconds—there will be no similarity transformation from mass-in-kilograms to mass-in-some-other-unit that will preserve the ratios of mass numbers assigned for all values of t.

The problem posed by this sort of example is that Luce's Principle need not hold for laws that contain *dimensional constants*. The law represented by equation (17) contains two dimensional constants: the initial mass of the material and the rate of decay. The latter is responsible for the violation of Luce's Principle. Their role and dimensionality is clearer if we express the law as a relation between the underlying dimensioned quantities rather than numbers on a scale representing those quantities. Using boldface letters to represent dimensioned variables:

$$\mathbf{m}(\mathbf{t}) = (10\,\mathrm{kg})\mathrm{e}^{-\mathbf{t}/(2\text{ minutes})} \tag{21}$$

where **t** is an amount of time and **m**(**t**) is the mass of the material that amount of time later than the initial time. Now suppose we want to represent equation (21) with numbers rather than dimensioned quantities. Our choice of scale for mass does not constrain our choice of scale for time, since these quantities are of independent dimensions. But, when selecting a numerical representation of (21), our choice of scale for a dimensioned variable does constrain our choice of scale for any constant of that dimension. Our original representation (17) obeys this constraint when we interpret the variable t as numbers of minutes and m as numbers of kilograms, since the dimensional constants 2 minutes and 10 kg are represented by the numbers 2 and 10 respectively. When we change the representation by applying an admissible transformation to the independent variable—i.e., multiplying t by α—we must apply the same transformation to the constant 2, since 2 is supposed to represent a quantity of the same dimension as t. And, similarly, we should expect the resulting transformation of the dependent variable to be applied also to the constant of the same dimension. Thus, what we should expect is that, for any positive α, there is some positive β such that, for any number of minutes t:

$$\beta m(t) = \beta 10\mathrm{e}^{-\alpha t/(2\alpha)} \tag{22}$$

And this is equivalent to equation (17). When we take care to apply the same transformations to the numbers representing dimensional constants as we apply to the numerical variables representing quantities of those dimensions, then the resulting requirement is trivially satisfied. This is consistent with Roberts's criterion for meaningfulness, which requires invariance "under all admissible transformations of *all* the scales involved" (emphasis added), including scales for dimensional constants.

Luce (1959, p. 93)'s interpretation of this situation is that dimensional constants "cancel out" any change in scales. Rozeboom (1962, p. 545) suggests that this "amounts to nothing but a more or less arbitrary selection of one of the admissible scalings of that variable, and then working up an 'absolute' interpretation for that scale." This may give the impression that the underlying quantities are themselves representable on an absolute scale. But this would be misleading. It is true that dividing a duration of time **t** by the constant 2 minutes yields a value that does not depend on the scale on which time is measured. But the ratio **t**/(2 minutes) is not a time; it is a dimensionless number. The exponent of a decay law is indeed measurable on an absolute scale, because exponents are numbers and cannot have dimension. But, for that very reason, the exponent of a decay law is not the value of a scale *for* time or other dimensioned quantity. Similarly, the ratio of **m**(**t**) to 10 kg does not depend on the unit of measurement, and the only admissible transformation of this ratio is the identity transformation, so we have an absolute scale of *something*. But it's not an absolute scale of *mass*. It's an absolute scale of the ratio of some mass to 10 kg; that ratio is not itself a mass, but a dimensionless number.

8 | HARSANYI'S WEIGHTS AS DIMENSIONAL CONSTANTS

We have seen that the inference from cardinal noncomparability of well-being to Intrapersonal Cardinal Invariance is an instance of Luce's Principle, with individual utilities as the independent variables and social welfare as the dependent variable (which is assumed to be at least ordinally measurable). And we have seen that Luce's Principle is not valid for laws that contain dimensional constants. In Nebel (2021) I argue that many theories of social welfare should be taken to appeal to such constants (though my discussion there is restricted to ratio-scale measurability with full interpersonal comparability). My suggestion here is that Harsanyi's weights should be understood as dimensional constants.

Recall that we are assuming Bernoulli's Hypothesis. So there is more than mere ordinal structure to betterness for an individual. There are differences in goodness for an individual. We can treat goodness for an individual, then, as a dimensioned quantity, much as we treat temperature. Ratios of differences in goodness, for a given individual, are real numbers.

To reflect the absence of interpersonal comparisons, we will treat goodness for one person and goodness for another person as quantities of different dimensions. Much as lengths and masses cannot be added together, differences in goodness for different people cannot be added together. This reflects the familiar point that, in the absence of interpersonal comparisons of goodness, the sum of different people's utilities does not represent a meaningful quantity.

The *weighted* sum of differences in goodness, however, may be a meaningful quantity, *if* the weights are dimensional constants. To illustrate this, consider an example involving more familiar dimensions that we antecedently know to be independent. Suppose that the price of a car is an additive function of its mass and its volume, where the price is linear with respect to each variable. A simple model of this would be to assign a price to each unit of mass and a price to each unit of volume and to add the price of a car's total mass and the price of its total volume to obtain its overall price. Let $\mathbf{p}(x)$ be the price of car x, $\mathbf{m}(x)$ its mass, and $\mathbf{V}(x)$ volume. Let $\lambda_{\mathbf{p/m}}$ be a constant of dimension $[\mathbf{p}]/[\mathbf{m}]$ (e.g., \$$n$ per kilogram) and $\lambda_{\mathbf{p/v}}$ a constant of dimension $[\mathbf{p}]/[\mathbf{V}]$ (e.g., \$$m$ per cubic meter). Then, on this model we have

$$\mathbf{p}(x) = \lambda_{\mathbf{p/m}}\mathbf{m}(x) + \lambda_{\mathbf{p/v}}\mathbf{V}(x) \tag{23}$$

Equation (23) does not involve the meaningless operation of adding mass to volume. It converts mass to money and volume to money, via the dimensional constants, and then adds two quantities of money—which are all perfectly meaningful operations. This does not treat quantities of mass as comparable to quantities of volume: there are no comparisons of the form $\mathbf{m}(x) > \mathbf{V}(x)$.

We can now return to the conclusion of Harsanyi's theorem. As stated in section 5, Number-Weighted Summation requires interpersonal comparisons of goodness differences, which we have assumed not to be meaningful. But it can be easily reformulated not to require such comparisons.

We have assumed, via Bernoulli's Hypothesis, that there are such things as ratios of differences in goodness for a person. We have made no analogous assumption for overall goodness. For this reason, we will not treat social welfare as a dimensioned quantity, and overall betterness will be treated as an ordinal notion. But it would be easy to extend the following remarks if there is such a quantity.

In the ideology of dimensioned quantities, we can state Harsanyi's desired conclusion as follows:

Quantity-Weighted Summation There are positive differences in goodness $\mathbf{k_1}, \ldots, \mathbf{k_n}$ such that, for any lotteries p and q: p is better than q just in case

$$\left[1/\mathbf{k_1}\right]\left[\mathbf{g_1}(p) - \mathbf{g_1}(q)\right] + \ldots + \left[1/\mathbf{k_n}\right]\left[\mathbf{g_n}(p) - \mathbf{g_n}(q)\right] > 0 \tag{24}$$

Each person's difference in goodness is weighted by the reciprocal of some constant difference in goodness for her. Each term in the summation is a real number: the ratio of each person's difference in goodness to that constant. One lottery is better than another just in case the sum of those ratios is positive. (If we accept Bernoulli's Hypothesis for overall goodness and want a "cardinal" version of Quantity-Weighted Summation, each weight should be the ratio of some difference in overall goodness to some difference in goodness for each person—i.e., the "rate" by which increments of i-goodness increase overall goodness, analogous to the dimensional constants in equation (23)—so that the weighted sum is a total difference in overall goodness.)

Quantity-Weighted Summation requires no interpersonal comparisons of goodness differences. It only requires intrapersonal ratio comparisons of goodness differences, which are meaningful if Bernoulli's Hypothesis is correct, and the uncontroversially meaningful operation of adding real numbers. By way of analogy, equation (23) for the price of a car does not require interdimensional comparisons of masses and volumes. It only requires ratios of money to mass, of money to volume, and sums of money.

We can think of Quantity-Weighted Summation as a generalization of utilitarianism. If interpersonal comparisons are meaningful, then we can treat each person's good as a quantity of the same dimension, making the constants $\mathbf{k_1}, \ldots, \mathbf{k_n}$ comparable. Utilitarians accept a version of Quantity-Weighted Summation in which $\mathbf{k_i} = \mathbf{k_j}$ for all individuals i and j. In that case, Quantity-Weighted Summation compares alternatives in the same way as Summation. But, when interpersonal comparisons are ruled out and each person's good is treated as a quantity of a distinct dimension, it is not even meaningful to say that two people's weights are equal, or that one person's weight is greater than another's. It would be like comparing the gravitational constant to the speed of light.

When Harsanyi's weights are understood as constants of distinct dimensions, it is clear why we should not expect Intrapersonal Cardinal Invariance to hold. Suppose that we fix a utility scale $u_i(\cdot)$ for each individual and represent overall betterness with an expectational social welfare function $W(\cdot)$ that satisfies equation (3): $W(p) = k_1 u_i(p) + \ldots + k_n u_n(p)$. A scale for each individual determines a particular number to represent the weight $\mathbf{k_i}$. For example, if $\mathbf{k_i} = \mathbf{g_i}(p) - \mathbf{g_i}(q)$, then let $k_i = 1/\left[u_i(p) - u_i(q)\right]$ (though this should not be taken to suggest that each person's weight must be the difference in goodness for her between some alternatives that are actually in the domain under consideration). Intrapersonal Cardinal Invariance would require, in accordance with Luce's Principle, that any combination of admissible transformations of each individual's utility scale result in at least an order-preserving transformation of $W(\cdot)$—that is, for any positive $\alpha_1, \ldots, \alpha_n$ and any $\beta_1, \ldots, \beta_n$,

$$W(p) > W(q) \Leftrightarrow \sum_{i=1}^{n} k_i \alpha_i u_i(p) + \beta_i > \sum_{i=1}^{n} k_i \alpha_i u_i(q) + \beta_i \tag{25}$$

which, by equation (3) and since the β_i's cancel out, implies that for any positive $\alpha_1, \ldots, \alpha_n$:

$$\sum_{i=1}^{n} k_i u_i(p) > \sum_{i=1}^{n} k_i u_i(q) \Leftrightarrow \sum_{i=1}^{n} k_i \alpha_i u_i(p) > \sum_{i=1}^{n} k_i \alpha_i u_i(q) \tag{26}$$

But this is not in general possible if all the k_i's must be positive, as Strong Pareto requires. (For example, let p be better for person 1, and let q be better for everyone else. Then we can find some sufficiently large α_1 and sufficiently small $\alpha_2, \ldots, \alpha_n$ such that p comes out better overall. But we can also find some sufficiently small α'_1 and sufficiently large $\alpha'_2, \ldots, \alpha'_n$ such that p comes out worse.)

As we have seen, however, Luce's Principle is false precisely because it fails to account for the role of dimensional constants. If the weights $\mathbf{k_1}, \ldots, \mathbf{k_n}$ are dimensional constants rather than dimensionless numbers, then we should only expect an admissible transformation of the utility scales to yield an admissible transformation of the social welfare function *given* a corresponding admissible transformation of the weights. Remember that a choice of scale for i-goodness determines a number to represent $\mathbf{k_i}$. If k_i represents person i's weight when u_i represents i's betterness relation, then k_i/α_i will represent i's weight when $\alpha_i u_i + \beta_i$ represents i's betterness relation. So we have:

$$W(p) > W(q) \Leftrightarrow \sum_{i=1}^{n} (k_i/\alpha_i)\alpha_i u_i(p) + \beta_i > \sum_{i=1}^{n} (k_i/\alpha_i)\alpha_i u_i(q) + \beta_i \tag{27}$$

and the transformations cancel out, leaving us with the original ordering:

$$W(p) > W(q) \Leftrightarrow \sum_{i=1}^{n} k_i u_i(p) > \sum_{i=1}^{n} k_i u_i(q) \tag{28}$$

As in the example of radioactive decay from section 7, when we take care to apply the same transformations to the numbers representing constants of i-goodness as we apply to the numerical variables representing *i*-goodness, then the social welfare function meets Roberts's criterion of invariance "under all admissible transformations of all the scales involved." Note that the dimensional constants do not themselves change when we use a different scale for each person's good; only their numerical representations change.

What does this imply for the standard utilitarian practice of representing a person's good by a utility function and comparing alternatives by their sums of utilities? If interpersonal comparisons of well-being are meaningful, this is no problem. But if interpersonal comparisons of well-being are not meaningful, then this is an error. It is like adding up a person's mass-in-grams and her height-in-inches and taking the resulting sum to represent a significant quantity. This error is not avoided by multiplying people's utilities by a fixed set of numerical weights, because the appropriate weights depend on the scale used to measure each person's well-being. If we measure *i*-goodness in *i*-shmutils instead of *i*-utils, we need to convert the number used to represent *i*'s weight from *i*-utils^{-1} to *i*-shmutils^{-1}, in order to preserve the overall betterness ranking of alternatives.

I have suggested that Harsanyi's weights be understood as dimensional constants and differences in goodness for each person as dimensioned quantities. This makes Quantity-Weighted Summation possible even without interpersonal comparisons of well-being. This violates Intrapersonal Cardinal Invariance, but we have seen that Intrapersonal Cardinal Invariance does not follow from the cardinal measurability and interpersonal noncomparability of well-being when the social welfare function can contain dimensional constants. Since Harsanyi's weights can plausibly be interpreted as dimensional constants, we should not expect it to satisfy Intrapersonal Cardinal Invariance and cannot conclude that Harsanyi's theorem requires interpersonal comparisons of well-being.

One might object that Quantity-Weighted Summation does require such comparisons, since the dimensional constants convert each person's difference in goodness onto the same (dimensionless) scale. But each person's weighted difference in goodness is not, on my interpretation, a difference in goodness for a person. It is the ratio of two differences in goodness, which (given Bernoulli's Hypothesis) is a real number. If comparing such ratios were all it took to make interpersonal comparisons of goodness, then such comparisons would be guaranteed so long as each person's good is measurable on a cardinal scale. Indeed, if that were all it took, then we could make comparisons of arbitrary quantities of different dimensions. Of course we can compare the ratio of my velocity to the speed of light and the ratio of my mass to yours, but this does not mean that velocities and masses are comparable.

I do not claim that Quantity-Weighted Summation is the view that utilitarians have meant by "weighted utilitarianism," or that there is anything incoherent about Summation or Number-Weighted Summation. They would be incoherent if interpersonal comparisons of well-being were impossible. But I do not believe they are, and neither did Harsanyi. My claim, rather, is that theorists who accept Harsanyi's premises while rejecting interpersonal comparisons of well-being should accept Quantity-Weighted Summation. Harsanyi wanted to show that even a skeptic about the possibility of interpersonal comparisons could be led, by his theorem, to a broadly utilitarian principle of aggregation. Quantity-Weighted Summation is the concusion to which they are led.

Of course, Quantity-Weighted Summation is not really utilitarian: it does not compare alternatives by their sums of well-being. But neither is Number-Weighted Summation. They both have formal features that are characteristically utilitarian: social welfare is an additively separable function of individual well-beings, and each person's well-being has constant, positive marginal value. It is a merely verbal question whether we classify Quantity-Weighted Summation as sufficiently close to utilitarianism to count as what Harsanyi called "a utilitarian ethic." But Quantity-Weighted Summation is certainly closer to utilitarianism than critics have thought Harsanyi could get—and perhaps as close as he could possibly get—without interpersonal comparisons of well-being.

9 | INDEPENDENCE OF IRRELEVANT ALTERNATIVES REVISITED

I have suggested that Harsanyi's desired conclusion can be formulated in a way that does not require interpersonal comparisons of well-being: namely, as Quantity-Weighted Summation. There is a problem, however. In section 5, we added Irrelevant Alternatives to the stock of Harsanyi's premises. But Quantity-Weighted Summation violates Independence of Irrelevant Alternatives. More generally, if we appeal to dimensional constants in a way that leads to violations of Luce's Principle, then Independence of Irrelevant Alternatives will not be satisfied. This is because the comparison of alternatives depends on more than the *numbers* assigned by each person's utility scale to each alternative: it also depends on the numbers assigned to dimensional constants. These numbers can change depending on the utility scales, though the constants they represent do not. This is an instance of a point due to Morreau and Weymark (2016): the Independence condition requires the comparison of two alternatives to depend only on the assignment of utilities to those alternatives, even when those utilities represent different quantities of well-being.

To see this, suppose that there are only two people. And suppose that each person's weight $\mathbf{k_i}$ is fixed at the value 1 i-util, where an i-util is some arbitrary difference in goodness for i. It is the difference in i-goodness between lotteries to which some utility function $u_i(\cdot)$—the i-util scale—assigns numbers that differ by 1. Now consider a utility profile $U = (u_1(\cdot), u_2(\cdot))$ and a pair of lotteries p and q such that:

$$\begin{array}{ccc} & u_1 & u_2 \\ p & 1 & 0 \\ q & 0 & 1 \end{array}$$

Since we have stipulated that each person's weight is the difference in goodness between alternatives to which $u_i(\cdot)$ assigns consecutive integers, we can conclude that $\mathbf{g_1}(p) - \mathbf{g_1}(q) = \mathbf{k_1}$ and $\mathbf{g_2}(q) - \mathbf{g_2}(p) = \mathbf{k_2}$, so the weighted sum of differences in goodness for each person between p and q will be zero. Quantity-Weighted Summation implies, then, that p and q are equally good. Now suppose we swap out person 1's utility function for another that assigns the same *numbers* to p and q, while changing the meaning of these numbers. Imagine that the difference in goodness for person 1 between p and q is increased by a factor of a thousand, but let us represent this difference using a "kilo-util" scale $v_1(\cdot)$ that assigns the number 1/1000 to $\mathbf{k_1}$ (leave person 2's utility function as is):

$$\begin{array}{ccc} & v_1 & u_2 \\ p & 1 & 0 \\ q & 0 & 1 \end{array}$$

Independence of Irrelevant Alternatives implies that this new profile must be assigned the same overall betterness ordering as the original, so p and q must still be equally good in this new profile. But Quantity-Weighted Summation implies that p is now better than q, since the difference in goodness for person 1 between p and q is now a thousand times the weight $\mathbf{k_1}$, and the difference in goodness for person 2 between q and p has stayed equal to $\mathbf{k_2}$, so the weighted difference for person 1 is greater than for person 2. Thus we have a violation of Independence of Irrelevant Alternatives.

However, this violation of Independence of Irrelevant Alternatives seems unobjectionable. The intuition that motivates Independence of Irrelevant Alternatives is that the comparison of two alternatives should depend only on how good those alternatives are for each person. Other alternatives are irrelevant, as are other features of the alternatives beyond their goodness for each person, in the sense that the goodness of each alternative for each person is the only variable on which the overall betterness relation depends. The violation due to dimensional constants is not like that. It's not as if the comparison of p and q depends on how good some third alternative r is for each person, or on some feature of the alternatives beyond their goodness for each person.

One might insist that it does, since I am interpreting Harsanyi's weights as differences in goodness for each person. But the weights are constant and need not be interpreted as the difference in goodness between some special pair of alternatives.[6] Perhaps there is some independent reason to think that overall betterness should not depend on any dimensional constant. This would rule out much more than just Quantity-Weighted Summation (Nebel 2021; see also Skow 2012). But, even if we are convinced that there are no such constants, we should distinguish that consideration from the sorts of consideration that motivate Independence of Irrelevant Alternatives, and have them reflected in independent principles.

We introduced Independence of Irrelevant Alternatives to prevent Harsanyi's weights from varying by profile. It was important to consider multiple profiles for two reasons. One was to ensure we could compare alternatives in a way that did not depend on the particular utility representation of individual betterness relations. The other was to let us compare alternatives in a way that did not depend on the particular *betterness facts*. This second reason was more controversial, since it is not clear that the individual betterness facts are contingent or temporary matters. But they might be on certain views of well-being, and on Harsanyi's preference-theoretic interpretation this consideration seems especially pressing.

For the sake of generality, let's assume that the individual betterness facts can change, not just our numerical representations of those facts. We can incorporate this idea by considering multiple profiles of *individual goodness* functions rather than utility functions. So far we have been interpreting $\mathbf{g_i}(p)$ as the goodness for person i of lottery p. Let's instead interpret it as the goodness for i of p *according to* the dimensioned profile $\mathbf{G} = (\mathbf{g_1}(\cdot), \ldots, \mathbf{g_n}(\cdot))$. We can think of each dimensioned profile as a possible assignment of distributions of goodness for each person. One profile might say that p's goodness for i is 1 i-util while another says that it is 2 i-utils.

Now think of a social welfare functional as a function from dimensioned profiles of individual goodness functions to overall betterness relations. We can then simply reinterpret Independence of Irrelevant Alternatives accordingly. Two dimensioned profiles $\mathbf{G}$ and $\mathbf{G'}$ coincide on p and q just in case, for each person i, p is just as good for i according to $\mathbf{G}$ as it is for i according to $\mathbf{G'}$—i.e., $\mathbf{g_i}(p) = \mathbf{g'_i}(p)$—and likewise for q. When reinterpreted in terms of dimensioned quantities of goodness, Independence of Irrelevant Alternatives requires the betterness relations assigned to two profiles to agree on two alternatives whenever those profiles coincide on those alternatives. This principle captures the idea that the comparison of alternatives should depend only on how good those alternatives for each person. Given this principle, the dimensional constant reflecting each person's weight cannot change as the individual goodness facts change; only its numerical representation can change.

This is just a sketch of how the Independence condition can be made compatible with Quantity-Weighted Summation. A much more detailed alternative to the orthodox framework of social welfare functionals has been proposed by Morreau and Weymark (2016). They introduce a framework of "scale-dependent" social welfare functionals, in which each person's utility function is paired with a *scale* that tells us what each utility number represents. This allows them to reformulate the standard axioms in ways that take into account the units in which welfare is measured. They do not consider

[6] In contrast to, for example, the maneuver in Gauthier (1986) to divide each person's difference in utility by her maximum gain, since introducing a superior alternative can change the maximum gain.

the role of dimensional constants, so the relation between their framework and the approach suggested here is not entirely clear. I believe that my main claims can be stated in their framework, but I leave open the exact relationship between the two approaches as a question for further research.

On the approach suggested here, as in Morreau and Weymark's framework, invariance conditions such as Intrapersonal Cardinal Invariance do not follow immediately from the measurability/comparability possibilities with which they are typically associated. Unlike profiles of utility functions, dimensioned profiles of individual goodness functions can differ only if there is a difference in how good some alternative is for someone. Transformations of individual goodness profiles are real changes in how good things are for people (this is not to imply that any such change is possible, since the domain of possible profiles may be limited). Without the further assumption that there are no dimensional constants, we cannot infer that the overall betterness ordering should be invariant to any class of such changes merely from the fact that our numerical representation of well-being can be admissibly transformed in such ways. As Morreau and Weymark argue, the standard Independence of Irrelevant Alternatives condition fails to distinguish between real changes in well-being and merely representational changes. Since there plainly is such a distinction, it is no objection to Quantity-Weighted Summation that it violates Independence of Irrelevant Alternatives as standardly formulated. The reformulation suggested here, in terms of dimensioned profiles of individual goodness functions, respects that distinction while also meeting the purpose to which the Independence of Irrelevant Alternatives is put in Mongin's extension of Harsanyi's theorem.

We can now see why the standard lesson drawn from Sen's extension of Arrow's theorem—that we need interpersonal comparisons of well-being to satisfy the Arrovian desiderata—is mistaken. Quantity-Weighted Summation is a counterexample. It assigns an overall betterness ordering to every possible profile of individual goodness functions, in a way that satisfies Weak Pareto, is nondictatorial, and depends, for any pair of alternatives, only on how good those alternatives are for each individual. Even though it violates Intrapersonal Cardinal Invariance, it does not require interpersonal comparisons of well-being. In a way, the mistake I am attributing to Sen and others is much like the mistake that Sen finds in Harsanyi, which led us to introduce Bernoulli's Hypothesis in section 3: he conflates a person's well-being with a utility function that represents her well-being. Cardinal measurability is sufficient to escape Arrow's impossibility when we treat each person's well-being as a dimensioned quantity and allow for the possibility of dimensional constants.

10 | CONCLUSION

I have argued that, when Harsanyi's weights are interpreted as dimensional constants, his aggregation theorem does not require interpersonal comparisons of well-being. But where do these dimensional constants come from? Harsanyi (1955, p. 316) suggested that, in the absence of interpersonal comparisons, the choice of weights must be arbitrary and depend on our "personal value judgments." In later work, however, Harsanyi argued that his theorem supported the possibility of interpersonal comparisons via the process of selecting weights. He claimed that an evaluator cannot avoid such comparisons,

> as long as he wants to choose the coefficients … of his social-welfare function … in a rational manner. This is so because the only way that individual *i* can judge how much relative weight a given set of coefficients … actually assigns to each individual's interests is by converting all *n* individuals' utility functions … into the same utility unit—which, of course, involves making interpersonal utility comparisons. (Harsanyi, 1977, p. 81)

Harsanyi's idea is that a rational procedure for selecting the weights must look something like this: we decide whose well-being is more important, whose is less important, and come up with a set of weights that reflect those comparisons. But, in order to tell whether a set of weights assigns more or less importance to one person's good than to another's, we must use utility functions that represent their good on the same scale.

The argument of this paper casts doubt on Harsanyi's suggestion. The proponent of Quantity-Weighted Summation need not start out with, or ever make, judgments about whose well-being is more or less important. Compare the car pricing analogy: we do not start out with judgments about whether mass or volume should be more expensive; such a comparison would not even make sense. We can instead select differences in well-being for each person that seem, upon reflection, to be equally important from a moral or social perspective. We need not claim that these are well-being differences of the same size, any more than setting the price-per-unit-volume and price-per-unit-mass of a car requires us to say that some unit of volume *equals* some unit of mass. We should distinguish comparisons of different people's well-being from comparisons of the moral or social significance of their well-being. Harsanyi's theorem requires that the latter are meaningful, but so do dictatorial and majoritarian social welfare functionals. It is one thing to say that a benefit to me is more important than a benefit to someone else—quite another to say that the first is larger than the second. (This is especially clear in the original preference-theoretic version of Harsanyi's theorem, since there is no obvious conceptual connection between the strength of a person's preferences and the social or moral weight of a preference of that strength.)

My argument also bears on the debate over the relevance of Harsanyi's theorem to utilitarianism. When confronted with Harsanyi's conclusion, it is natural to wonder why all the weights *shouldn't* just be equal. Why should some people's well-being receive greater weight than others'? This, I suspect, is what makes Broome's leap to Summation from Harsanyi's conclusion and Bernoulli's Hypothesis seem tempting: we naively assume that the weights are dimensionless numbers and see no reason to assign different numbers to different people. In the absence of interpersonal comparisons, however, the weights *cannot* all be equal—not because some people's weights must be greater than others', but rather because they are quantities of distinct dimensions and therefore not even comparable.

The broader lesson of this paper is this. Some of the most important results in social choice theory appeal crucially to invariance conditions, such as Intrapersonal Cardinal Invariance, that are supposed to correspond to some natural possibility concerning the measurability and comparability of well-being. These invariance conditions, however, do not follow from the intended measurability/comparability conditions without the additional assumption that there are no relevant dimensional constants. This does not make these results any less significant, since the invariance conditions may be natural and interesting enough even apart from their intended interpretation as straightforward consequences of measurability/comparability conditions. But the additional assumption we have highlighted—that there are no relevant dimensional constants—warrants at least as much investigation as the standard questions about measurability and comparability that have occupied much of this literature since the work of Arrow and Harsanyi.

ACKNOWLEDGEMENTS

I am very grateful to Matthew Adler, Harvey Lederman, Ralph Wedgwood, John Weymark, and an audience at the 2021 Virtual Conference on Social Choice Theory and Applications for helpful comments and discussion.

REFERENCES

Adler, M. D. (2019). *Measuring social welfare: An introduction*. Oxford University Press. https://doi.org/10.1093/oso/9780190643027.001.0001

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, *58*, 328–346.

Blackorby, C., Donaldson, D., & Weymark, J. A. (2008). Social aggregation and the expected utility hypothesis. In M. Fleurbaey, M. Salles, & J. A. Weymark (Eds.), *Justice, political liberalism, and utilitarianism: Themes from Harsanyi and Rawls* (pp. 136–183). Cambridge University Press.

Bossert, W., & Weymark, J. A. (2004). Utility in social choice. In S. Barberà, P. J. Hammond, & C. Seidl (Eds.), *Handbook of utility theory: Volume 2 extensions* (pp. 1099–1177). Springer US. https://doi.org/10.1007/978-1-4020-7964-1_7

Broome, J. (1991). *Weighing goods: Equality, uncertainty and time*. Wiley-Blackwell.

Broome, J. (2004). *Weighing lives*. Oxford University Press.

Dasgupta, S. (2013). Absolutism vs comparativism about quantity. *Oxford Studies in Metaphysics*, *8*, 105–150.

d'Aspremont, C., & Gevers, L. (2002). Social welfare functionals and interpersonal comparability. In K. J. Arrow, A. K. Sen, & K. Suzumura (Eds.), *Handbook of social choice and welfare* (pp. 459–541). Elsevier. https://doi.org/10.1016/S1574-0110(02)80014-5

Dreier, J. (2004). Decision theory and morality. In P. Rawling & A. Mele (Eds.), *Oxford handbook of rationality* (pp. 156–181). Oxford University Press.

Eddon, M. (2013). Quantitative properties. *Philosophy Compass*, *8*, 633–645. https://doi.org/10.1111/phc3.12049 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12049

Fleurbaey, M., & Mongin, P. (2016). The utilitarian relevance of the aggregation theorem. *American Economic Journal: Microeconomics*, *8*, 289–306. https://doi.org/10.1257/mic.20140238

Gauthier, D. (1986). *Morals by agreement*. Oxford University Press. http://oxford.universitypressscholarship.com/view/10.1093/0198249926.001.0001/acprof-9780198249924

Gevers, L. (1979). On interpersonal comparability and social welfare orderings. *Econometrica*, *47*, 75–89. https://doi.org/10.2307/1912347

Greaves, H. (2017). A reconsideration of the Harsanyi–Sen–Weymark debate on utilitarianism. *Utilitas*, *29*, 175–213. https://doi.org/10.1017/S0953820816000169

Greaves, H., & Lederman, H. (2018). Extended preferences and interpersonal comparisons of well-being. *Philosophy and Phenomenological Research*, *96*, 636–667. https://doi.org/10.1111/phpr.12334 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phpr.12334

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

Harsanyi, J. C. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge University Press.

Harsanyi, J. C. (1978). Bayesian decision theory and utilitarian ethics. *The American Economic Review*, *68*, 223–228.

Harsanyi, J. C. (1979). Bayesian decision theory, rule utilitarianism, and Arrow's impossibility theorem. *Theory and Decision*, *11*, 289–317.

Hausman, D. M. (1995). The impossibility of interpersonal utility comparisons. *Mind*, *104*, 473–490.

Jeffrey, R. C. (1971). On interpersonal utility theory. *The Journal of Philosophy*, *68*, 647–656. https://doi.org/10.2307/2024935

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*, 81–95.

Luce, R. D. (1962). Comments on Rozeboom's criticisms of 'On the Possible Psychophysical Laws'. *Psychological Review*, *69*, 548–551.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. Wiley.

Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. *Econometrica*, *18*, 111–141. https://doi.org/10.2307/1907264

Mill, J. S. (1863). *Utilitarianism*. Cleveland: Cambridge University Press.

Mongin, P. (1994). Harsanyi's aggregation theorem: Multi-profile version and unsettled questions. *Social Choice and Welfare*, *11*, 331–354.

Mongin, P., & d'Aspremont, C. (1998). Utility theory and ethics. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory volume 1: Principles* (pp. 371–481). Dordrecht: Kluwer Academic Publishers.

Morreau, M. & Weymark, J. A. (2016). Measurement scales and welfarist social choice. *Journal of Mathematical Psychology*, *75*, 127–136. https://doi.org/10.1016/j.jmp.2016.04.001

Nebel, J. M. (2021). Utils and Shmutils. *Ethics*, *131*, 571–599. https://doi.org/10.1086/712578

Risse, M. (2002). Harsanyi's 'Utilitarian Theorem' and utilitarianism. *Noûs*, *36*, 550–577.

Roberts, F. S. (1984). *Measurement theory: With applications to decisionmaking, utility, and the social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511759871

Rozeboom, W. W. (1962). The untenability of Luce's principle. *Psychological Review*, *69*, 542–547. https://doi.org/10.1037/h0041419

Sen, A. (1970). *Collective choice and social welfare*. Holden-Day San Francisco.

Sen, A. (1977). Non-linear social welfare functions: A reply to Professor Harsanyi. *Foundational problems in the special sciences* (pp. 297–302). Dordrecht: Springer. https://doi.org/10.1007/978-94-010-1141-9_19

Sen, A. (1999). The possibility of social choice. *The American Economic Review*, *89*, 349–378.

Sider, T. (2020). *The tools of metaphysics and the metaphysics of science*. Oxford University Press. http://oxford.universitypressscholarship.com/view/10.1093/oso/9780198811565.001.0001/oso-9780198811565

Skow, B. (2012). How to adjust utility for desert. *Australasian Journal of Philosophy*, *90*, 235–257. https://doi.org/10.1080/00048402.2011.572079

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*, 2nd edn. Princeton University Press.

Weymark, J. A. (1991). A reconsideration of the Harsanyi–Sen debate on utilitarianism. In J. Elster & J. E. Roemer (Eds.), *Interpersonal comparisons of well-being* (1st edn., pp. 255–320). Cambridge University Press. https://doi.org/10.1017/CBO9781139172387.009

Weymark, J. A. (2016). Social welfare functions. In M. D. Adler & M. Fleurbaey (Eds.), *The Oxford handbook of well-being and public policy*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199325818.013.5

Wolff, J. E. (2020). *The metaphysics of quantities*. Oxford University Press.

How to cite this article: Jacob M. Nebel Aggregation Without Interpersonal Comparisons of Well-Being. *Philosophy and Phenomenological Research*. 2022;105:18–41. https://doi.org/10.1111/phpr.12783