Harsanyi's Utilitarian Theorems without Tears

Johan E. Gustafsson and Kacper Kowalczyk*

Draft
January 29, 2025

According to classical utilitarianism, well-being consists in pleasure or happiness, the good consists in the sum of well-being, and moral rightness consists in maximizing the good. Leibniz was perhaps the first to formulate this doctrine. Bentham made it widely known.[1] For a long time, however, the second, summing part lacked any clear foundation. John Stuart Mill, Henry Sidgwick, and Richard Hare all gave arguments for utilitarianism, but they took this summing part for granted.[2] It was John Harsanyi who finally presented compelling arguments for this controversial part of the utilitarian doctrine.

This is a strong candidate for the most important contribution to ethical theory of the last century. Alas, ethicists have largely neglected it — likely because its presentation usually involves unobvious mathematics. John Broome, who has done much to popularize Harsanyi's work, comments that the additive structure of utilitarianism 'arises from the mathematics, in a not very intuitive fashion.'[3] The purpose of this piece is to explain Harsanyi's contribution, give an intuitive account of his arguments, and discuss their ethical significance. In our presentation, we favour intuition over generality and try to minimize any use of mathematics.

1. Introduction

Harsanyi gave three arguments for the summing part of the utilitarian doctrine: the social-aggregation theorem, the impartial-observer theorem, and the separability theorem. Remarkably, all three are contained in the seminal early paper from 1955,

* We would be grateful for any thoughts or comments on this paper. They can be sent to *johan.eric.gustafsson@gmail.com*.

[1] Leibniz 1700, p. 378 and Bentham 1789.

[2] See Mill 1969, pp. 234–239, Sidgwick 1907, pp. 380–382, and Hare 1982, pp. 26–27.

[3] Broome 2015, p. 252.

merely re-asserted and elaborated in Harsanyi's later work. In this piece, we focus on the social-aggregation theorem, but we explain the other two as well.[4]

The social-aggregation theorem crucially involves the idea of uncertainty about the outcomes of our choices. Typically, we do not choose between final *outcomes*; rather, we choose between *prospects*, which give chances to possible final outcomes. For example, if you think that sunshine and rain are equally likely, then taking an umbrella to work amounts to the prospect of staying dry for sure, whereas travelling light amounts to the prospect of equal chances of getting wet and staying dry.

Expected Utility Theory is the standard theory for choice under uncertainty. It has been developed for preference but can be applied to moral and individual evaluation as well. The theory claims that we can assign utility numbers to each possible final outcome in such a way that one prospect is at least as good — or as preferred — as another if and only if it has an at least as high *expected utility*, which is the sum of each final outcome's utility multiplied by its probability.

Expected Utility Theory rests on four axioms, that is, four basic principles from which the theory's main claim can be derived. These axioms will remain in the background of our discussion. We will appeal instead to intuitive claims which can be, more or less directly, derived from them. Yet the argument is, fundamentally, nothing more than repeated application of these axioms.[5]

[4] The impartial-observer theorem appears in Harsanyi 1953 — Harsanyi's first publication; see Fontaine 2010, p. 152 — and is subsequently discussed in 1975b, pp. 313–314; 1975a, p. 598; 1977b, pp. 48–51; 1977a, pp. 631–636; 1978, pp. 227-228; 1979, pp. 294–295; 1982, pp. 44-48; 1992, pp. 675–677. The social-aggregation theorem is discussed in Harsanyi 1955, pp. 312–314; 1975b, p. 313; 1977a, pp. 636–638; 1977b, pp. 64–69; 1978, pp. 226-227; 1979, pp. 292–294; 1982, pp. 48–49; 1992, pp. 677–679. The separability theorem is first established in Fleming 1952. It is endorsed by Harsanyi 1955 and elaborated in Harsanyi 1977b, pp. 69–81. Harsanyi (1977b, p. 293) emphasises that his case for utilitarianism can only be refuted if all three of these theorems fail to support utilitarianism. The most fully developed treatment of all three theorems can be found in Harsanyi 1977b, 48–83. In this piece, we follow Broome's reinterpretation of Harsanyi's theorems in terms of moral and individual *value* rather than moral and individual *preference*; see Broome 1987, pp. 410–412; 1991, pp. 151–164.

[5] The four axioms can be stated, following Jensen 1967, p. 173, as follows:

> *Completeness* For any pair of prospects, either the first is at least as good as the second, or the second is at least as good as the first.
>
> *Transitivity* If a first prospect is at least as good as a second, and the second is at least as good as a third, then the first is at least as good as the third.
>
> *Independence* If one prospect is better than another, then adding a chance of a third prospect to both of the original prospects does not change which one is better.
>
> *Continuity* If one prospect is better than a second, which is better than a third, then there are probabilities p and q between zero and one such that getting the first prospect with probability p and the third prospect otherwise is better than the second prospect, which is in turn better than getting the first prospect with probability q and the third

Harsanyi's social-aggregation theorem applies Expected Utility Theory in order to set up utility scales for moral and individual evaluations which it then connects by means of the following principle:

> *Ex-Ante Pareto* If one prospect is at least as good as another prospect for everyone, then the first is at least as good as the second. And, if, in addition, the first prospect is better than the second prospect for someone, then the first is better than the second.

This principle is an expression of ethical individualism, making moral evaluation depend solely on individual evaluations, without appeal to what Harsanyi calls 'the separate interests of a superindividual state or of impersonal cultural values.'[6]

The theorem says that, if this principle is true and both moral and individual evaluations of prospects satisfy the axioms of Expected Utility Theory, then a prospect is better if and only if the sum of utility numbers assigned to it by each individual is higher.

This theorem thus shows, in the first instance, that moral evaluation must track totals of utility numbers assigned by each individual. But what do these numbers really mean? Unless they are connected to well-being, we might wonder whether Harsanyi's theorem supports utilitarianism at all. Indeed, this is what Broome describes as 'the standard objection' to Harsanyi's case for utilitarianism.[7]

In what follows, we will examine Harsanyi's contribution in detail. We start with Expected Utility Theory, the foundation of Harsanyi's social-aggregation theorem (§2), which we then explain in an intuitive but rigorous way (§3). We then discuss the philosophical significance of this theorem (§4). Lastly, we present the impartial-

---

otherwise.

Also needed are additional assumptions about reducing complex prospects; see Jensen 1967, p. 170. Many different axiomatisations exist: Harsanyi (1955) uses Marschak 1950's axioms, while Harsanyi (1977b) uses those of Herstein and Milnor 1953. An accessible proof appears in Harsanyi 1977b, pp. 22–47, though the first such result is due to von Neumann and Morgenstern (1944, pp. 15–29).

[6] Harsanyi 1955, p. 311, 313. Harsanyi (1975b, p. 313) even calls it the 'individualism postulate.' Harsanyi (1955, pp. 313) used an 'indifference' version of Ex-Ante Pareto, which, given our reinterpretation, would say that if one prospect is equally good as another for everyone, then the two are equally good. He suggested adding a 'strong' version already in Harsanyi 1955, p. 314, fn. 12.

[7] See Broome 2008, p. 231. This objection targets the significance of the theorem's conclusion while accepting its premises. But there are also objections to the premises themselves. For example, Ex-Ante Pareto can be denied for prospects under uncertainty while retained for outcomes under certainty. This objection is pursued in Rabinowicz 2002, pp. 11–12, Fleurbaey and Voorhoeve 2013, pp. 117–121, and Adler and Sanchirico 2006, pp. 347–350, as well as Adler 2012, pp. 506–518; 2019, pp. 136–138. Another objection involves denying the axioms of Expected Utility Theory, particularly Independence; see Diamond 1967 and Epstein and Segal 1992.

observer theorem (§5) as well as the separability theorem (§6), which represents an attempt to justify utilitarianism without using probabilities.

2. Expected Utility

Let's suppose that you like bananas the most, coconuts the least, with strawberries in between. But how much do you like each one? We can try to use probabilities to figure this out.[8]

Let's say you are indifferent between getting a strawberry and a 50–50 chance of getting a banana or a coconut. This suggests that the chance of switching a strawberry for a banana is worth the same to you as the risk of getting a coconut instead of a strawberry. So strawberries seem to be halfway between bananas and coconuts in terms of your preference. But, if you do not mind giving up a strawberry for a slim chance of getting a banana and the risk of otherwise getting a coconut, this suggests that you like bananas much more than strawberries. This is because a slim chance of a banana is enough for you to outweigh the certainty of a strawberry. So strawberries seem to be nearer coconuts in terms of your preference.

In this way, we can use probabilities to assign *utility numbers* to prospects which you consider to be intermediate between coconuts and bananas. Coconuts and bananas thus can serve as the endpoints of the utility scale. For example, the utility of getting a strawberry can be defined as the number s for which the following holds:

(1) $\begin{matrix} 100\% \\ 🍓 \end{matrix} \sim \begin{matrix} s & 1-s \\ 🍌 & 🥥 \end{matrix}$

In this notation, outcomes are listed on the bottom line with their probabilities listed on the top line and '∼' denotes indifference. It is possible to define utility numbers in this way because Expected Utility Theory implies that every prospect intermediate in preference between two others is indifferent to some unique probability of getting the more preferred prospect and otherwise getting the less preferred one.[9]

The resulting utility scale has two useful features. First, it *represents* your preferences, in the sense that a prospect is assigned a higher utility number if and only if you prefer it. To see this, suppose that you prefer strawberries to apples:

(2) $\begin{matrix} 100\% \\ 🍓 \end{matrix} \succ \begin{matrix} 100\% \\ 🍏 \end{matrix}$

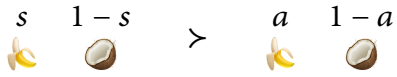
[8] The presentation below roughly follows Harsanyi 1977b, pp. 22–47.

[9] The derivation of this claim mainly relies on the Continuity axiom; see the argument for Lemma (5.6)(b) in Kreps 1988, pp. 46–48. But the claim itself is often presented as an axiom — sometimes known as 'solvability' — as in Marschak's (1950, p. 117) axiomatisation.

Here, '≻' denotes preference. Now, the utility of getting an apple is defined as the number a for which the following holds:

(3) 

Expected Utility Theory allows us to treat indifferent prospects interchangeably.[10] So, given (1) and (3), we find that (2) is equivalent to

(4) 

Lastly, Expected Utility Theory implies that, whenever only two outcomes are possible, the more preferred prospect is the one which makes the more preferred outcome more likely.[11] So (4) is equivalent to

(5) $$s > a$$

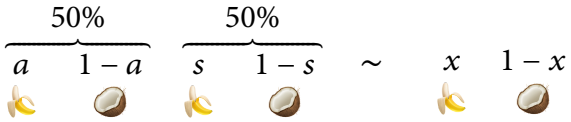
This shows that strawberries are assigned a higher utility number than apples if and only if you prefer strawberries to apples. So the utility scale represents your preferences.

The second useful feature of the utility scale is that it is *expectational*, in the sense that the utility assigned to a prospect equals the expected utility of its possible outcomes, that is, the sum of the utility of each outcome multiplied by its probability.

To see this, consider a prospect with a 50% chance of getting a strawberry and a 50% chance of getting an apple. The utility of this prospect is defined as the number x for which the following holds:

(6) 

Now, remember that Expected Utility Theory implies that indifferent prospects can be treated interchangeably. So, substituting (1) and (3) into (6), we have

(7) 

[10] The derivation of this claim mainly relies on the Independence axiom. See the argument for Lemma (5.6)(c) in Kreps 1988, pp. 48. But the claim itself is often presented as an axiom, as in Marschak's (1950, p. 120) axiomatization.

[11] See the argument for Lemma (5.6)(a) in Kreps 1988, p. 47.

The result is a complex prospect that assigns probabilities to other prospects rather than directly to outcomes. But note that this prospect assigns probability 0.5 to getting a banana with probability a, while also assigning probability 0.5 to getting a banana with probability s. So, overall, it assigns probability $0.5a + 0.5s$ to getting a banana. Expected Utility Theory thus allows us to simplify (7) as follows:

(8) $0.5(a+s)$ 🍌 $\quad 1-0.5(a+s)$ 🥥 $\quad\sim\quad x$ 🍌 $\quad 1-x$ 🥥

Now, remember that Expected Utility Theory imply that, whenever only two outcomes are possible, the more preferred prospect is the one which makes the more preferred outcome more likely. So (8) implies

(9) $$x = 0.5a + 0.5s$$

If x was greater than $0.5a+0.5s$, the prospect on the right would be preferred; if x was lesser than that, the prospect on the left would be preferred. But the two prospects are equally good, so x must equal $0.5a + 0.5s$. This is the utility of the prospect with a 50% chance of getting an apple (an outcome assigned utility a) and a 50% chance of getting a strawberry (an outcome assigned utility s). We found that it is equal to its expected utility. This shows that the utility assigned to the prospect is equal to the expected utility of its outcomes. So the utility scale is expectational.

The expectational character of our utility scale is therefore no mystery: it is essentially the result of repeatedly substituting equivalent prospects and simplifying probabilities in accordance with the laws of probability.[12]

But what if we set up our scale in terms of endpoints other than bananas and coconuts? Then for any fruit — let's use apples as an example — we would obtain new utility numbers related to the old ones in the following way:

(10) $\text{utility}_{\text{new}}(🍏) = \Big(\text{utility}_{\text{new}}(🍌) - \text{utility}_{\text{new}}(🥥)\Big) \cdot \text{utility}_{\text{old}}(🍏) + \text{utility}_{\text{new}}(🥥)$

This means the new utility number is the result of multiplying the old utility number by a positive constant (the new utility distance between the old endpoints) and adding a constant (the new utility of the old zero).[13] If two utility scales are related in

[12] Harsanyi (1977b, p. 38) notes that the Expected Utility Theory (in cases where there is a best and a worst outcome) 'is a direct consequence of the Multiplication and Addition Laws of the probability calculus.'

[13] To obtain this formula, note that banana and coconut — our old endpoints — are assigned some utilities on the new scale. Recall that, on the old scale, we defined the utility of an apple in terms of the

this way, we say that they are *linear transformations* of each other. As a result, items from the old scale would be spaced out in the same proportions on the new scale. Both scales would agree on which prospects are preferred and have the property that the utility of a prospect is its expected utility. The new scale would simply cover a larger range of items than the old one.

We can also invert the relationship in (10) to express old utilities in terms of new ones:

$$\text{utility}_{\text{old}}(\text{🍏}) = \frac{\text{utility}_{\text{new}}(\text{🍏}) - \text{utility}_{\text{new}}(\text{🥥})}{\text{utility}_{\text{new}}(\text{🍌}) - \text{utility}_{\text{new}}(\text{🥥})} \tag{11}$$

This inverse relationship allows us to extend our utility scale beyond its initial endpoints. We can assign utilities to items outside our original range by considering how they would be valued on a hypothetical, more extensive scale.

For items beyond the top endpoint, consider an item preferred to our original top (for example, grapes preferred to banana). We imagine a new scale where the original top (banana) has utility v. On this new scale, the preferred item (grapes) would have utility 1. Using the inverse relationship, we can assign grapes a utility of $1/v$ on our original scale. For items below the bottom endpoint, consider an item dispreferred to our original bottom (for example, lemon dispreferred to coconut). We imagine a new scale where the original bottom (coconut) has utility w. On this new scale, the dispreferred item (lemon) would have utility 0. Using the inverse relationship, we can assign lemon a utility of $-w/(1-w)$ on our original scale.

This method allows us to map utilities from hypothetical extended scales back to our original scale. As a result, we can construct a single, comprehensive utility scheme that represents an unlimited range of outcomes, even if there are no best or worst outcomes.

3. Social Aggregation

Let's now consider dividing fruit between two individuals, Ann and Bob. Ann likes apples while Bob likes strawberries. How to evaluate the outcome of giving Ann an apple and Bob a strawberry?

---

following indifference: an apple is indifferent to the prospect of getting a banana with probability a and getting a coconut otherwise. Now, the new scale, like the old one, represents the same preference relation while also being expectational. It follows that the new utility of getting an apple is equal to a times the new utility of getting a banana plus $(1-a)$ times the new utility of getting a coconut. Rearranging, we get the formula in the main text.

To answer this question, we first need to set up individual utility scales for Ann and Bob. For simplicity, let's suppose that both like bananas most and coconuts least. Coconuts and bananas can thus serve as endpoints of their respective scales.

Next, we need to set up a moral utility scale. We need to choose endpoints for this scale. The top endpoint can be the outcome of giving both a banana and the bottom endpoint the outcome of giving both a coconut. Given what Ann and Bob like, Ex-Ante Pareto implies that every possible fruit allocations is intermediate in moral value between these two outcomes.

Now, we can define the moral utility of giving Ann an apple and Bob a strawberry as the number x for which the following holds:

(12)

	100%		x	$1-x$
Ann	🍏	~ Ann	🍌	🥥
Bob	🍓	Bob	🍌	🥥

In this notation, rows specify outcomes for Ann and Bob, while '~' now denotes equal goodness. Now the task is to figure out what number x is. The first step uses Ex-Ante Pareto to relate our target outcome — giving Ann an apple and Bob a strawberry — to other outcomes whose utility might be easier to establish:

(13)

	50%	50%		50%	50%
Ann	🍏	🥥	~ Ann	🍏	🥥
Bob	🍓	🥥	Bob	🥥	🍓

This means that two prospects are equally good. On the left, Ann and Bob's outcomes are correlated: it is 50-50 whether both get their preferred fruits (apple and strawberry) or both get coconuts. On the right, their outcomes are anti-correlated: it is certain that one gets their preferred fruit (apple or strawberry) and the other gets a coconut but 50-50 who gets which.

Next, Ex-Ante Pareto helps us again by connecting the moral utility of the outcomes on the right-hand side of (13) to Ann's and Bob's individual utilities for apples and strawberries. To see this, note that Ann's utility for getting an apple is defined as the number a for which the following holds:

(14)

100%		a	$1-a$
🍏	$\sim_{\text{Ann}}$	🍌	🥥

Here, '$\sim_{\text{Ann}}$' denotes equal-goodness-for-Ann. Now, it follows from Ex-Ante Pareto that moral evaluation must coincide with Ann's evaluation if Bob is unaffected. So (14) implies

(15)

	100%			a	$1-a$
Ann	🍏	~	Ann	🍌	🥥
Bob	🥥		Bob	🥥	🥥

The outcome where Ann receives a banana and Bob receives a coconut is one where Ann gets the top endpoint of her scale and Bob the bottom of his. This outcome must be assigned some number on the moral utility scale. It is the number w for which the following holds:

(16)

	100%			w	$1-w$
Ann	🍌	~	Ann	🍌	🥥
Bob	🥥		Bob	🍌	🥥

So, substituting (16) into (15) and then simplifying, we get

(17)

	100%			aw	$1-aw$
Ann	🍏	~	Ann	🍌	🥥
Bob	🥥		Bob	🍌	🥥

This means that the utility of giving Ann an apple and Bob a coconut is equal to Ann's utility of getting an apple multiplied by a weight determined by the moral utility of giving Bob a coconut while giving Ann a banana, that is, giving Ann the top endpoint of her scale and Bob the bottom of his.

Now, let's consider Bob's case. Suppose that Bob's own utility for getting a strawberry is s. Let's also assume that giving Bob a banana (the top endpoint of his scale) while giving Ann a coconut (the bottom endpoint of hers) is assigned moral utility v. Applying a similar argument as we did for Ann, we can conclude

(18)

	100%			sv	$1-sv$
Ann	🥥	~	Ann	🍌	🥥
Bob	🍓		Bob	🍌	🥥

Now, we collect everything we have established so far. We substitute (12), (17), (18) into (13). We then get the following complex-looking prospect:

(19)

	50%				50%		50%	
	x	$1-x$	50%	~	aw	$1-aw$	sv	$1-sv$
Ann	🍌	🥥	🥥		Ann 🍌	🥥	🍌	🥥
Bob	🍌	🥥	🥥		Bob 🍌	🥥	🍌	🥥

Now, simplifying all this in accordance with the laws of probability, we get

		$0.5x$	$1 - 0.5x$			$0.5(aw + sv)$	$1 - 0.5(aw + sv)$
(20)	Ann	🍌	🥥	~	Ann	🍌	🥥
	Bob	🍌	🥥		Bob	🍌	🥥

This implies

$$x = aw + sv \tag{21}$$

If x was greater than $aw + sv$, the prospect on the left would be better; if x was lesser than that, the prospect on the right would be better. But the two prospects are equally good, so x must equal $aw + sv$. This the moral utility of giving Ann an apple and Bob a strawberry. We found that it is equal to the weighted sum of Ann's utility of getting an apple and Bob's utility of getting a strawberry. That is:

		100%			$aw + sv$	$1 - (aw + sv)$
(22)	Ann	🍏	~	Ann	🍌	🥥
	Bob	🍓		Bob	🍌	🥥

The moral utility scale represents moral evaluation. So, to compare two outcomes in terms of moral value, we can simply compare their moral utility. We can do this by adding up the utilities which individuals assign on their scales, weighting them based on where the tops of their scales are placed on the moral utility scale.

Put more generally, if we start with utility scales for Ann and Bob, there are weights such that moral evaluation can be represented by

$$\text{Ann's weight} \cdot \text{Ann's utility} + \text{Bob's weight} \cdot \text{Bob's utility} \tag{23}$$

But, since the unit of a utility scale is arbitrary, we can select individual utility scales with the weights built in. So there exist *some* utility scales for Ann and Bob such that moral evaluation can be represented by

$$\text{Ann's utility} + \text{Bob's utility} \tag{24}$$

This argument can be extended to cases with more numerous societies and to cases where individuals have different best and worst outcomes, or even where no best or worst outcomes exist. This proceeds roughly as with extending the utility scale in the individual case which we covered in §2.

This way of presenting Harsanyi's social-aggregation theorem should remove much of the mystery surrounding it. The theorem shows that moral value must have

additive structure. This comes primarily from Expected Utility Theory, which is inherited from the multiplication and addition laws of the probability calculus. In this way, the mathematics of probability imprints itself on moral evaluation. Ex-Ante Pareto, on the other hand, links situations where outcomes are received simultaneously to ones where they are received separately with equal probability, and it ensures that the moral evaluation of these individualized outcomes tracks their individual evaluation.[14]

4. Implications for Ethics

Utilitarianism claims that the good consists in the sum of individual well-being. For this to make sense, well-being has to be both *quantitatively* measurable and interpersonally *comparable*. That is, it must make sense both to say whether one individual is better off than another and by how much.

Harsanyi's social-aggregation theorem does not assume anything about the nature of individual well-being. Instead, it shows that utility numbers can be assigned to both individual and collective outcomes such that their sum represents the moral evaluation of prospects. How, then, can Harsanyi's theorem support utilitarianism? The answer depends on the relationship between utility and well-being. We can distinguish several possibilities here, based on whether well-being is interpersonally comparable and whether it is quantitatively measurable.

[14] Harsanyi's (1955) version of the theorem assumes the 'indifference' version of Ex-Ante Pareto. His conclusion is that moral value is represented by a weighted total of individual utility numbers, with weights not necessarily positive. The Pareto principle in Harsanyi 1977b, p. 65 is intermediate between that in Harsanyi 1955, p. 313 and our Ex-Ante Pareto which is only used in Harsanyi 1992, p. 678. Harsanyi's proof is, however, essentially the same in 1955, pp. 313–314 and 1977b, pp. 65–68. It relies on a lemma to the effect that, if the utility of one prospect is a fixed multiple of the utility assigned to another on everyone's scale, then the moral utility of the first is the same multiple of the moral utility of the second. The meaning of this lemma is, however, not intuitively obvious. The same lemma is used in Resnik's (1983; 1987, pp. 197–20) popular presentation of the theorem. Moreover, Harsanyi's proof crucially depends on an unstated assumption — which he only made explicit in Harsanyi 1992, p. 678 — roughly to the effect that individual outcomes are freely recombinable. This was first pointed out by Resnik 1983; see also Resnik 1987, pp. 200–204. But since ethical theory can consider merely logically possible situations, this simplifying assumption is unobjectionable even if in practice different people's outcomes cannot be freely recombined. Hence our presentation also relies on this assumption. The original version of Harsanyi's theorem can also be proved without the unstated assumption, albeit at the cost of unintuitive linear algebra and non-uniqueness of the resulting weights; see Fishburn 1984, Border 1985, Coulhon and Mongin 1989, but especially Selinger 1986. This aspect literature is helpfully surveyed by Weymark 1991, pp. 264–282; 1994. A different kind of approach to establishing Harsanyi's theorem is presented by Fleurbaey 2018, p. 9; see also Mongin and d'Aspremont 1998, p. 426–427 and Coulhon and Mongin 1989, pp. 183–187. This approach relies, however, on results about functional equations whose meaning is not intuitively obvious.

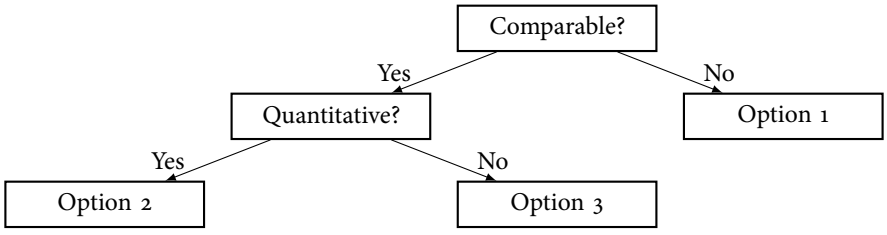


Figure 1: Nature of Well-being

OPTION 1: NON-COMPARABLE WELL-BEING

The first possibility is that well-being is not interpersonally comparable. In this case, utilitarianism makes no sense: we cannot meaningfully add up well-being across different individuals. Yet Harsanyi argued that his theorem still supports a form of utilitarianism. He drew an analogy to decision-making under uncertainty: just as rational agents must act as if they assigned subjective probabilities to hypotheses, moral agents must act as if they made interpersonal comparisons of well-being. The premises of Harsanyi's theorem thus imply we must behave as if interpersonal comparisons made sense, even if they do not.[15]

[15] See Harsanyi 1955, p. 321: 'There is here an interesting analogy with the theory of statistical decisions (and, in general, the theory of choosing among alternative hypotheses). In the same way as in the latter, it has been shown that a rational man (whose choices satisfy certain simple postulates of rationality) must act *as if* he ascribed numerical subjective probabilities to all alternative hypotheses, even if his factual information is insufficient to do this on an objective basis — so in welfare economics we have also found that a rational man (whose choices satisfy certain simple postulates of rationality and impartiality) must likewise act *as if* he made quantitative interpersonal comparisons of utility, even if his factual information is insufficient to do this on an objective basis.' See also Harsanyi 1975b, p. 325; 1977b, p. 60.

There is some debate about whether Harsanyi's social-aggregation theorem involves interpersonal comparisons. Broome (1987, p. 420) argued that it *presupposes* such comparisons: 'it is simply an assumption of the theorem that …interpersonal comparisons are possible'. See also Broome 1991, pp. 219–220. Conversely, Jeffrey (1971, p. 653) claimed that interpersonal comparisons *follow* from the theorem: 'Harsanyi's work allows us to derive a complete set of interpersonal comparisons of preferences from a social preference ranking which is judged to be fair, together with the personal preference rankings of the constitutive individuals'. See also Mongin and d'Aspremont 1998, p. 432. Recently, Nebel (2022) has argued, agreeing with Harsanyi himself, that interpersonal comparisons are neither presupposed nor implied by the theorem. Our presentation supports this view: the utilities added in the theorem are primarily moral utilities, correlated with individual utility scales through weights. The theorem does not assert that goods of different individuals can be directly added, which would be as nonsensical as adding lengths and masses.

OPTION 2: COMPARABLE AND QUANTITATIVE WELL-BEING

The second possibility is that well-being is both interpersonally comparable and quantitatively measurable. This is the only possibility in which utilitarianism makes sense.

Because well-being is comparable, the following plausible principle also makes sense:

> *The Principle of Impartiality* If two outcomes differ only in who has which level of well-being, then they are equally good.[16]

Given the rest of Harsanyi's theorem, this principle implies that moral value can be represented by the sum of individual utilities, all measured on the same utility scale. To see this, recall that Harsanyi's theorem allows us to define utility numbers for Ann and Bob such that moral value is represented by their sum. Let's suppose that well-being w is assigned zero on Ann's utility scale. By the Principle of Impartiality, the outcome with Ann at w and Bob at v is as good as the outcome with Ann at v and Bob at w. Harsanyi's theorem then implies:

$$\text{utility}_{\text{Ann}}(w) + \text{utility}_{\text{Bob}}(v) = \text{utility}_{\text{Ann}}(v) + \text{utility}_{\text{Bob}}(w) \tag{25}$$

Bob's utility of being at well-being level w is simply some number. So, substituting zero for Ann's utility of w and that number for Bob's utility of w, it follows that Bob's utility at any well-being level v must equal Ann's utility at that level plus that number:

$$\text{utility}_{\text{Bob}}(v) = \text{utility}_{\text{Ann}}(v) + \text{constant} \tag{26}$$

This constant cancels out when comparing outcomes. Moral value can thus be represented by summing utilities measured on the same scale. In what follows, we assume the Principle of Impartiality and thus use the same utility function for everyone.

Because well-being is quantitatively measurable as well as interpersonally comparable, utilitarianism makes sense. But whether it is true depends on what the relationship between utility and well-being is. The following figure shows some possibilities.

[16] Harsanyi formulates a version of this principle in Harsanyi 1977b, p. 69. Note that this principle does not presuppose that well-being is quantitatively measurable.

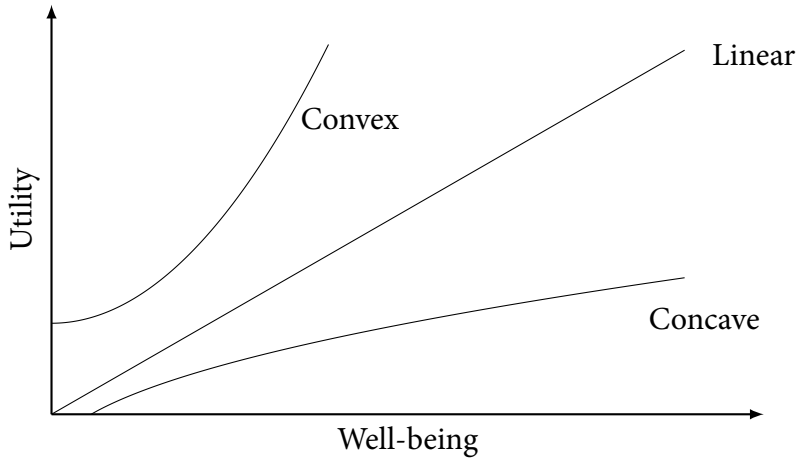


Figure 2: Relationship between Utility and Well-being

The relationship between utility and well-being might take several forms. It might be *linear*: the result of multiplying well-being by a positive constant and adding another constant; for example, utility(w) = $2w + 3$. It could be *concave*: sloping upward but curving downward; for example, utility(w) = $\sqrt{w}$. Or it could be *convex*: sloping upward and curving upward; for example, utility(w) = w^2. (These are just examples chosen for simplicity. Other functions may be more plausible.[17])

The social-aggregation theorem tells us that moral value can be represented by summing utilities. So, if utility is a linear function of well-being, utilitarianism is *true*. When comparing outcomes, constants from multiplication and addition cancel out, so comparing sums of utilities becomes equivalent to comparing sums of well-being.

But, if utility is *not* linear in well-being, utilitarianism is *false*. Consider a concave utility function like utility(w) = $\sqrt{w}$. Given two outcomes — both individuals having five units of well-being versus one having ten units and the other zero — summing utilities would favour the first while summing well-being would rate them equally. In general, a concave utility function yields not utilitarianism but *prioritarianism*, which gives priority to those worse off in absolute terms, as well-being gains at lower levels translate into larger utility gains and thus greater moral value.[18]

This is the basis of Amartya Sen's objection that Harsanyi's theorem does not, by itself, support utilitarianism.[19] Sen is right: we also need the claim that utility is a

[17] For instance, there are concave functions which, unlike utility(w) = $\sqrt{w}$, are defined for negative well-being numbers, for example, utility(w) = $1 - e^{-w}$.

[18] Early discussions of prioritarianism appear in Atkinson and Stiglitz 1980, pp. 333-365 in economics and Weirich 1983 in philosophy. The view was named and popularized by Parfit 1995.

[19] See Sen 1976, pp. 247–251 and Sen 1977, pp. 298–301. See also Sen 1986, pp. 1122-1124 and Roemer

linear function of well-being.

But, even by itself, Harsanyi's theorem remains significant. This is because the theorem links the correct attitude to risk in well-being with the correct attitude to inequality in well-being. If utility is a linear function of well-being, like utility(w) = $2w + 3$, we get *risk-neutrality* in the sense that it is just as good to have a moderate amount of well-being for sure as to take a 50-50 gamble that either doubles or destroys that amount. This is because possible well-being gains always receive the same weight in calculating individual value. But, if utility is a concave function of well-being, like utility(w) = $\sqrt{w}$, we get *risk-aversion* in the sense that it is better to have a moderate amount of well-being for sure than to play double-or-quits with well-being. This is because well-being gains translate into larger utility gains when they happen at lower levels. This kind of risk-aversion is compatible with Expected Utility Theory.

Now, as we saw, a linear utility function yields utilitarianism while a concave utility function yields prioritarianism. Harsanyi's theorem therefore shows that intrapersonal and interpersonal aggregation are tightly linked: risk-neutrality corresponds to utilitarianism, risk-aversion to prioritarianism. Harsanyi's case for utilitarianism could therefore be completed by arguing for risk-neutrality in well-being. This is the content of

> *Bernoulli's Hypothesis* A prospect is at least as good for someone as another prospect if and only if the expected well-being from the first is at least as great as the expected well-being from the second.

This principle implies that well-being, like utility, is an expectational representation of individual value. This in turn implies that the relationship between utility and well-being must be linear.[20]

But, as Broome points out, Bernoulli's Hypothesis is, at first blush, implausible.[21] It would mean that it is best for anyone to play double-or-quits with well-being. But it might be supported indirectly. Given the rest of Harsanyi's theorem and the Principle of Impartiality, everyone's utility scale must be essentially the same. Since the

---

2008, p. 135, as well as surveys in Weymark 1991, pp. 297–315; 2005, and Greaves 2017.

[20] For the Bernoulli Hypothesis, see Broome 1991, p. 142, which Broome attributes to Bernoulli 1738 (translated as Bernoulli 1954). The argument from the Bernoulli Hypothesis to linearity parallels the argument given towards the end of §2. Well-being and utility must both be expectational representations of an individual's evaluation. Therefore, they must be related as two expectational utility scales are related, that is, by a linear transformation. Instead of the Bernoulli Hypothesis, we could also appeal to the following principle: a 50-50 chance of getting x or y is as good for someone as a 50-50 chance of getting z or w if and only if the well-being difference between x and z is as large as that between y and w. See Ellingsen 1994, pp. 135–136 and Fleurbaey and Mongin 2016, p. 294.

[21] Broome 1991, pp. 144–145.

relationship between utility and well-being reflects the correct attitude to risk, this means that everyone must — incredibly — exhibit the same attitude to risk. Given this forced unanimity, risk-neutrality appears privileged due to its symmetry.[22]

OPTION 3: COMPARABLE BUT NON-QUANTITATIVE WELL-BEING

It might be that well-being does not clearly have quantitative structure to begin with. In this case, it is unclear whether utilitarianism even makes sense: we cannot meaningfully sum well-being across individuals. Yet Harsanyi's theorem might still support a form of utilitarianism. Broome suggests that we can use Bernoulli's Hypothesis to stipulatively define quantities of well-being where none existed. Hilary Greaves suggests that well-being's quantitative character is indeterminate among several possible definitions, and Bernoulli's Hypothesis offers a plausible resolution of this indeterminacy. Utilitarianism does not, strictly speaking, make sense; but we can both make it make sense and, at the same time, make it true. This is because it is up to us what differences in well-being amount to, and we can make them coincide with differences in utility.[23]

Consider the following analogy. Suppose we move boxes in a warehouse. We can tell which box is bigger by seeing which fits through the entrance door. But, if your box is "twice the size" of mine, what does this mean? There is no clear answer. It could mean twice the side length, volume, or face area. Similarly with well-being: we might be able to make qualitative comparisons while lacking a clear way to measure differences.[24]

Expected Utility Theory might appear to be a uniquely natural way to make well-being quantitative. When we find that a 50-50 chance of getting a banana or a coconut is equally good for someone as getting an apple for sure, this is like observing that a banana and coconut in one pan balance two identical apples in the other pan of a scale. This suggests that the well-being difference between the apple and the banana equals the difference between the coconut and the apple.[25]

There are, nonetheless, several alternative ways to make well-being quantitative.[26] First, we can use the person trade-off method.[27] This assumes that moral

[22] A similar argument is suggested by Adler 2019, p. 272.

[23] See Broome 1987, pp. 418–421; 1991, pp. 145–148; 2004, pp. 86–91; 2008, p. 230–232; 2015, pp. 262–264, and Greaves 2017, pp. 192–210. See also Jensen 1995, pp. 41–58 and Otsuka 2015, pp. 10–20.

[24] This analogy is from Greaves 2017, p. 194–195.

[25] This analogy is from Broome 2008, p. 231.

[26] See Ellingsen 1994, pp. 118–133, Jensen 1995, pp. 54–58, and Greaves 2017, pp. 200–202.

[27] A version of this approach is used in health economics; see Patrick et al. 1973, pp. 234–242. The label 'person trade-off' is due to Nord 1992, pp. 568–569.

value is the sum of well-being. Pick a good outcome (value 1) and a neutral outcome (value 0). Then assign outcome x the number $1/n$ if having n people at outcome x and one at the zero outcome is as good as having n people at the zero outcome and one at the good outcome. The well-being difference between x and y thus represents how many more people getting y we would need to save from getting the zero outcome in order to equal the value of saving other people getting x. However, this method adds well-being across individuals, so it builds in the summing part of utilitarianism.

Second, we can use the time trade-off method.[28] This assumes that a person's lifetime well-being is the sum of well-being over time. Pick a good state (value 1) and a neutral state (value 0). Then assign state x the number $1/n$ if living n years in state x is as good as living one year in the good state. The well-being difference between x and y thus represents how many more years in state y we would need to match the value of living one year in x. We can extend this method to assign values to entire lives because every life is plausibly just as good as some life spent in a constant state. But this method adds well-being over time, so it rules out effects like diminishing returns to longevity.[29]

Third, we can use what we might call 'the goods trade-off' method. This assumes that an individual's well-being is the sum of different types of goods; for example, pleasure and achievement. Mark a unit on the pleasure scale by picking out two levels (a good level and a neutral one) as well as a zero level for achievement. Then assign outcome x the number n if it takes n units of pleasure to reach an outcome equivalent to x when holding achievement at the zero level. The well-being difference between x and y thus represents how many more units of pleasure it takes to reach x compared to what it takes to reach y, when achievement is held at the zero level. But this method assumes that each good contributes independently to total value, ruling out interaction effects between goods.[30]

Fourth, we can use the just-noticeable-difference method.[31] This assumes hedonism and a limited ability to detect changes in pleasure, with the smallest noticeable increment of pleasure labeled a 'just-noticeable difference'. Assign number n to outcome x if it takes n just-noticeable steps to reach x from the baseline of zero pleasure. This approach presupposes that the just-noticeable steps correspond to equal amounts of well-being across people. This means, however, that a more sensitive person gains more well-being from the same improvement in external circumstances

[28] A version of this approach is used in health economics; see Torrance et al. 1972, pp. 124–125.

[29] See Otsuka 2015, p. 17.

[30] See Broome (1991, pp. 25–26) and Greaves (2017, pp. 201–202).

[31] See Edgeworth 1881, pp. 7, 60, 98–102 and Ng 1975, p. 563, but also Bentham in a 1782 manuscript, as quoted in Stigler 1950, p. 310.

such as income.

A fifth approach also assumes hedonism. Here, we identify pleasure with a natural property such as hormone levels or neuron firing rates, which has its own scale. An outcome's well-being matches that measured level. The well-being difference between x and y thus represents the difference in hormone levels or neuron firing rates between these states. But this rules out that the same physiological changes can have different effects on well-being, depending on circumstances.[32]

Lastly, we can rank not just outcomes (x is better than y) but also changes between outcomes (exchanging x for y is better than exchanging z for w). If these rankings satisfy certain conditions, we can represent them with numbers. For example, after setting a good outcome to 1 and a neutral one to 0, we can assign 1/2 to an outcome x if exchanging x for the good outcome is as good as exchanging the neutral one for x. We can then find outcomes worth 1/4, 3/4, 1/8, and so on. The well-being difference between x and y equals that between z and w if and only if exchanging y for x is as good as exchanging w for z. This method is the most general since it places no substantive constraints on well-being comparisons.[33]

5. Impartial Observer

In addition to his social-aggregation theorem, Harsanyi also proved another theorem, his impartial-observer theorem, based on the idea that a person's moral values coincide with what they would prefer when they are do not know who they are — having an equal chance of being anyone. The theorem begins with the following principle:

> *The Impartial-Observer Principle* A prospect is at least as good as another if and only if the first prospect would be at least as preferred as the second by a rational and sympathetic observer who had an equal chance of taking anyone's place.

This principle is plausible as an expression of ethical impartiality.[34] The observer is rational in the sense of satisfying the axioms of Expected Utility Theory, as well as sympathetic in the sense of satisfying the following principle:

> *The Principle of Acceptance* A prospect is at least as good for someone as another prospect if and only if the first would be at least as preferred

[32] See Otsuka 2015, p. 19.

[33] It was first developed in Alt 1936 (translated as Alt 1971). See also Krantz et al. 1971, pp. 136–198. This method is favoured by, for example, Adler (2019, pp. 55–56) and Stefánsson (2023, pp. 303–306).

[34] See Harsanyi's comments in 1953, pp. 434–435. Harsanyi's preferred name for this principle was the 'equiprobability model of value judgments'; see for example Harsanyi 1978, p. 227.

as the second by an observer who had to take that person's place, sharing their tastes and other traits.

This principle is plausible as an expression of respect for each individual's perspective.[35] The theorem says that, if these two principles are true, then a prospect is better if and only if the sum of utility numbers assigned to it by each individual is higher.

To see how this works, let's again consider dividing up fruit between two individuals, Ann and Bob. Ann likes apples while Bob likes strawberries. How to evaluate the outcome of giving Ann an apple and Bob a strawberry?

To answer this question, we can figure out how the observer evaluates the prospect with a 50% chance of getting an apple while taking Ann's place and a 50% chance of getting a strawberry while taking Bob's place. The observer's evaluation of this prospect is supposed to coincide with the moral evaluation of the outcome of giving Ann an apple and Bob a strawberry.

First, we need to set up individual utility scales for Ann and Bob and a utility scale for the observer. Note that the Principle of Acceptance implies that an individual's preference coincides with the preference which the observer has conditional on taking that individual's place. Since the observer satisfies the axioms of Expected Utility Theory, so must both individuals. For simplicity, let's suppose that both like bananas most and coconuts least, as does the observer. Coconuts and bananas thus can serve as the endpoints of their respective utility scales.

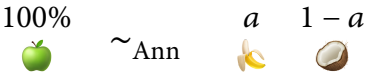
The observer's utility of giving Ann an apple and Bob a strawberry can thus be defined as the number x for which the following holds:

(27) $\quad 50\%\ \text{🍏}_{\text{Ann}}\ \ 50\%\ \text{🍓}_{\text{Bob}} \sim x\ \text{🍌}\ \ 1-x\ \text{🥥}$

In this notation, a subscripted outcome — such as '🍏$_{\text{Ann}}$' — means getting an apple while taking Ann's place together with her tastes and other traits, and '$\sim$' stands for the observer's indifference relation.

Now, the Principle of Acceptance helps us again by connecting the observer's utility of the outcomes on the right-hand side of (27) to Ann's and Bob's individual utilities for apples and strawberries. To see this, note that Ann's utility for getting an apple is defined as the number a for which the following holds:

[35] This principle is first suggested Harsanyi 1955, p. 316, fn. 16 but only stated more explicitly and precisely in Harsanyi 1977b, p. 51–55, where the observer is said to have extended preferences defined over ordered pairs consisting of individual outcomes and subjective circumstances. Mongin 2001, p. 155 argues that Vickrey's (1945) similar but earlier theorem has 'no ethical significance' because it does not include anything like this principle.

(28) 

The Principle of Acceptance implies that the observer must have a corresponding preference on the assumption of taking Ann's place. So (28) implies

(29) 

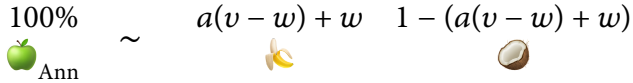
We need to be careful here. The observer's utility for getting an apple while taking Ann's place does not have to be a. This is because the observer might, for instance, like some sorts of fruit less conditionally on taking Ann's place. We thus need to place the endpoints of Ann's scale on the observer's scale. The observer's utility for getting a banana (the top endpoint of Ann's scale) while taking Ann's place is defined as the number v for which the following holds:

(30) 

Similarly, the observer's utility for getting a coconut (the bottom endpoint) while taking Ann's place is defined as the number w for which the following holds:

(31) 

Now, substituting (30) and (31) into (29) and then simplifying in accordance with the laws of probability, we get

(32) 

The observer's utility for getting an apple while taking Ann's place therefore depends on the observer's utilities for getting a banana and a coconut, both conditional on taking her place. These outcomes are assigned utility numbers on the observer's scale, not necessarily the same as on Ann's scale. The observer's utility of getting an apple while taking Ann's place equals Ann's utility of getting an apple multiplied by a weight and incremented by a constant. It is therefore a linear transformation of Ann's utility. Let a' be this transformation:

(33) $$a' = a(v - w) + w$$

Now, let's consider Bob's case. Suppose that Bob's own utility for getting a strawberry is s. Let's also assume that, when taking Bob's place, the observer's utility for getting a banana (the top of Bob's scale) is b, and for getting a coconut (the bottom of Bob's scale) it is c. Applying a similar argument as we did for Ann, we conclude

(34) $\underset{\text{Bob}}{100\%} \sim o(b-c)+c \quad 1-(o(b-c)+c)$

Similarly as before, the observer's utility of getting an apple while taking Bob's place is equal to Bob's utility of getting an apple multiplied by a weight and incremented by a constant. Let o' be this transformation:

(35) $$o' = o(b-c)+c$$

We can then substitute (32) and (34) into (27) and simplify, to get:

(36) $x \quad 1-x \sim 0.5(a'+o') \quad 1-0.5(a'+o')$

This implies

(37) $$x = 0.5a' + 0.5o'$$

If x was greater than $0.5a' + 0.5o'$, the prospect on the left would be preferred by the observer; if x was lesser than that, the prospect on the right would be preferred. But the two prospects are indifferent for the observer, so x must equal $0.5a' + 0.5o'$.

This x represents the observer's utility of giving Ann an apple and Bob a strawberry when facing an equal chance of taking either Ann's place or Bob's place. We found that it equals the weighted average of Ann's utility for getting an apple and Bob's utility for getting a strawberry, plus some constants. That is:

(38) $\underset{\text{Ann}}{50\%} \quad \underset{\text{Bob}}{50\%} \sim 0.5(a'+o') \quad 1-0.5(a'+o')$

The observer's utility scale for prospects where taking anyone's place is equally probable represents moral evaluation. So, to compare two outcomes in terms of moral value, we can simply compare their utility for the observer under this condition of equal probability. We can do this by adding up the utilities which individuals assign on their scales, weighting them based on how the observer's utilities relate to individual utilities when taking each individual's place. The constants involved are uniform across outcomes and thus cancel out when making comparisons.

Put slightly more generally, if we start with any utility scales for Ann and Bob, there are weights and constants such that moral evaluation can be represented by

(39) Ann's weight · Ann's utility + Bob's weight · Bob's utility

Since the unit of a utility scale is arbitrary, we can select individual utility scales with the weights built in. So there exist *some* utility scales for Ann and Bob such that moral evaluation can be represented by

(40) Ann's utility + Bob's utility

This argument can be extended to cases with more numerous societies and to cases where individuals have different best and worst outcomes, or even where no best or worst outcomes exist. This extension proceeds similarly to how we extended the utility scale in the individual case, as covered in §2.

This presentation of Harsanyi's impartial-observer theorem should remove any lingering mystery surrounding it. As in the social-aggregation theorem, the additive structure of utilitarianism comes chiefly from Expected Utility Theory, which itself reflects the laws of probability. The Principle of Acceptance plays a key role in connecting individual utilities with the observer's utility (and thus, moral utility). Its role is analogous to the role of Ex-Ante Pareto in the social-aggregation theorem.[36]

6. Separability

We have covered Harsanyi's two main utilitarian theorems. But he also relied on a third theorem — the separability theorem — which was first developed by Marcus Fleming. Unlike the other theorems the separability theorem does not involve uncertainty, a feature Harsanyi saw as an advantage. The theorem works by setting up utility scales for individual evaluations in such a way that moral evaluation can be represented by a sum of these utilities. The theorem begins with the following principle:

[36] Our presentation of the impartial-observer theorem follows Harsanyi 1977b, pp. 48–60. The result is first presented in Harsanyi 1953, but without anything like the Principle of Acceptance, which is first briefly mentioned in Harsanyi 1955, p. 316, fn. 16. Many presentations — for example, Sen 1986, pp. 1122–1123 — do not make this key principle explicit. The idea of using an impartial observer ignorant of their place in society was later popularised by Rawls (1971, pp. 11–17); his first inchoate appeal to a similar idea can be found in Rawls 1957, p. 656. Rawls's chief differences from Harsanyi are that his observer: should assume they can take the place of representatives of social classes rather than individuals themselves, is not allowed to use probabilities (on the spurious grounds that doing so requires Laplace's Principle of Indifference), and is required to use an extreme 'maximin' decision rule (as it is supposed to be grounded in the special features of the observer's situation). Harsanyi offers a persuasive response to Rawls in Harsanyi 1975a; Rawls 1974 responds; see also Harsanyi 2008.

> *Separability* If two outcomes differ only in their effects on two individuals, then which outcome is better does not depend on the situation of unaffected individuals.

This principle is plausible as an expression of ethical individualism, making moral evaluation depend solely on the evaluations of individuals affected.[37]

The theorem shows how separability, along with certain auxiliary principles, lets us construct utility scales such that moral value can be represented as the sum of individual utilities. This approach connects individual and moral evaluation directly, without appealing to probability.[38]

Let's see how this works by considering three individuals: Ann, Bob, and Cat. Ann likes apples, Bob likes bananas, and Cat likes coconuts. For simplicity, assume these are the only things they like and can get. For each person, more of their pre-

[37] A version of this principle is stated in Fleming 1952, pp. 372–374 who argues it allows us to disregard our ignorance about parts of the world we cannot affect; to use Fleming's examples, present Martians or far-future Earthlings. A similar principle is relevant to the 'Egyptology' objection against average utilitarianism in population ethics; see McMahan 1981, p. 115. Harsanyi sees separability as extending the individualism of Ex-Ante Pareto. If only two individuals are affected and they agree, Ex-Ante Pareto makes moral evaluation depend only on their evaluations; separability does this even when the two disagree. Both principles leave aim to leave 'no room for the separate interests of a superindividual state or of impersonal cultural values'. See Harsanyi 1955, p. 311. Both Harsanyi and Fleming stress that one might be affected in the relevant sense even if one's income is unaffected, if relative income matters for well-being. Fleming (1957) suggests an interesting argument from Ex-Ante Pareto and a dominance principle weaker than Expected Utility Theory to separability, foreshadowing an argument of Thomas 2022, pp. 280–282.

[38] The auxiliary principles, following Wakker's development of Debreu's topological approach to the theorem, are as follows:

> *Non-Triviality* At least three individuals exist who are not universally indifferent.
>
> *Ordering* Both moral and individual evaluations must be complete (for any pair of outcomes, either the first is at least as good as the second, or the second is at least as good as the first) and transitive (if a first outcome is at least as good as a second and the second is at least as good as the third, then the first is at least as good as the third).
>
> *Continuity* If every outcome in a sequence approaching outcome x is at least as good (bad) as some outcome y, then x itself must be at least as good (bad) as y.
>
> *Topological Assumption* The space of possible outcomes for each individual must be connected (cannot be divided into two non-empty open sets) and separable (contains a countable dense subset). The space of social outcomes must have the product topology, meaning convergence for society happens when and only when it happens for each individual.

See Wakker 1989, p. 70 and Debreu 1959 (reprinted in Debreu et al. 1983). It is possible to replace the final two assumptions — which are topological in nature — with algebraic ones instead; see Krantz et al. 1971, p. 301–302.

ferred fruit is always better.

Separability implies that the moral value of changes in someone's holdings can be assessed independently of others' holdings. We can thus represent each person's holdings on a separate vertical axis.

We begin with Ann. Starting from nothing, let a_1 be a sufficiently small increment of apple. This base increment will serve as our measuring rod: changes equivalent in moral value to Ann's getting a_1 instead of nothing will represent one unit of utility.

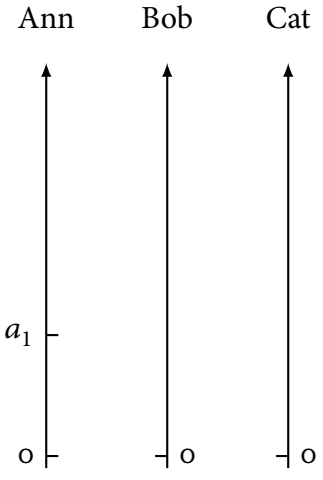


Figure 3: Setting up

Now we turn to Bob. We want to find increments of banana that are equivalent in moral value to Ann's base increment. This will let us divide Bob's holdings into increments each worth one unit of utility.

First, we define b_1 as the increment for Bob such that: Ann's getting a_1 instead of nothing is as good as Bob's getting b_1 instead of nothing. This definition compares two outcomes: one where Ann is at a_1 and Bob has nothing, the other where Bob is at b_1 and Ann has nothing. Next, we define b_2 as the increment for Bob such that Ann's getting a_1 instead of nothing is as good as Bob's getting b_2 instead of b_1. This definition also compares two outcomes: one where Ann is at a_1 and Bob has b_1, the other where Bob is at b_2 and Ann has nothing. We can continue this process as in Figure 4.

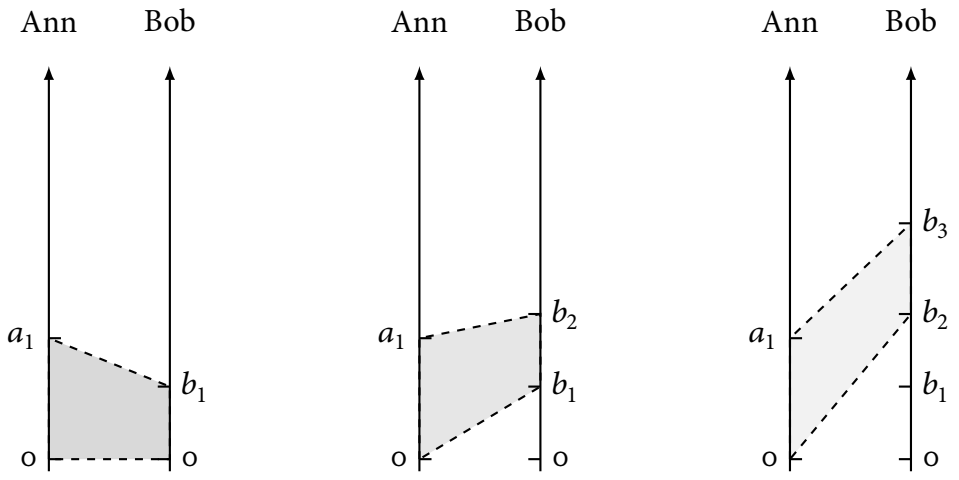


Figure 4: Dividing Bob's holdings

These increments need not be physically equal: what matters is their moral equivalence. In fact, due to diminishing marginal benefit of fruit consumption, we would expect subsequent increments to be physically larger; getting some fruit when you had none typically matters more than getting the same amount when you already have plenty.

Our construction of equivalent increments relies on separability. Why? Without separability, whether getting a_1 instead of nothing creates as much moral value as getting b_1 instead of nothing might depend on Cat's holdings. But separability lets us make these comparisons independently of unaffected individuals.

The construction thus allows us to set up a utility scale for Bob. If Bob's holdings can be divided into m increments, each equivalent in moral value to Ann's base increment, we say Bob's utility is m. These utility numbers are simply counting how many increments of equivalent moral value we can find in someone's holdings, using our chosen increment as a measuring rod.

We follow an analogous procedure for Cat. We want to find increments of coconut that are equivalent in moral value to Ann's base increment. If Cat's holdings can be divided into n increments, each equivalent in moral value to Ann's base increment, we say that Cat's utility is n.

Note that we defined Bob's base increment through the equivalence that Ann's getting a_1 instead of nothing is as good as Bob's getting b_1 instead of nothing. It follows that each of Cat's increments must also create as much moral value as Bob's base increment.

Now we return to Ann to divide her holdings into increments as well. Since we can't use Ann's base increment to measure further changes in her holdings, we use

Bob's base increment, which we defined as equivalent in moral value.

First, we define a_2 as the increment for Ann such that: Ann's getting a_2 instead of a_1 is as good as Bob's getting b_1 instead of nothing. Next, we define a_3 as the increment for Ann such that: Ann's getting a_3 instead of a_2 is as good as Bob's getting b_1 instead of nothing. We can continue this process, as in Figure 5.

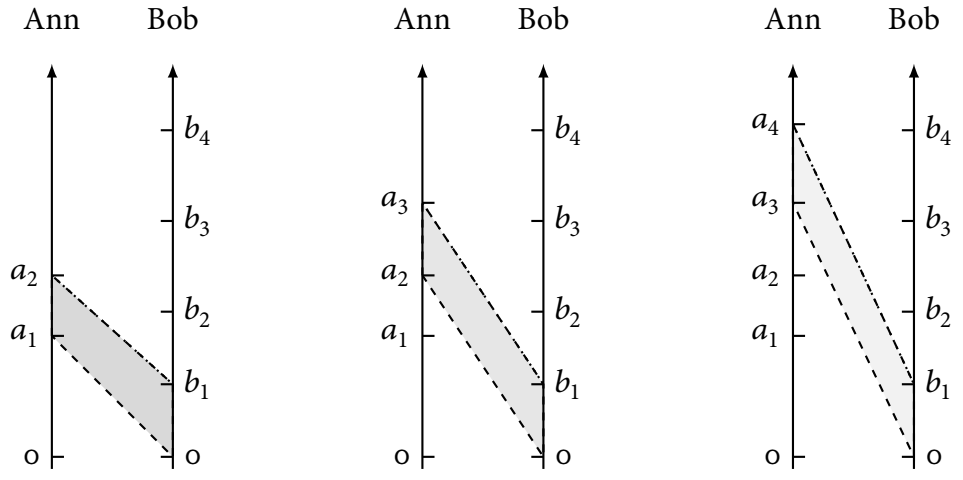


Figure 5: Dividing up Ann's holdings

Through this construction, we divide Ann's holdings into increments each creating as much moral value as Bob's base increment and, therefore, by our earlier equivalence, as much moral value as Ann's base increment. If Ann's holdings can be divided into o such increments, we say that Ann's utility is o.

Now let's see what happens when we transfer holdings between individuals. Start with a situation where Bob has bananas that we have divided into m equivalent increments, and Cat has coconuts we have divided into n equivalent increments. (We imagine, for the time being, that Ann has nothing.) We want to show that this is exactly as good as taking all of Bob's bananas away and giving Cat additional coconuts; specifically, enough extra coconuts to amount to m increments that we marked out on her scale.

Here's what follows from our construction: When Bob loses one increment of banana, this creates exactly as much moral loss as Ann gaining her base increment of apple creates moral gain. Due to separability, this equivalence holds true no matter how many coconuts Cat has. Similarly, when Cat gains one increment of coconut, this creates exactly as much moral gain as Ann losing her base increment of apple creates moral loss. Again by separability, this equivalence holds true regardless of how many bananas Bob has.

These two equivalences together show us that: Bob losing one increment of banana creates exactly as much moral loss as Cat gaining one increment of coconut creates moral gain. And because of separability, this equivalence holds true regardless of what Ann holds, or how many bananas or coconuts Bob and Cat currently have. This is illustrated in Figure 6.

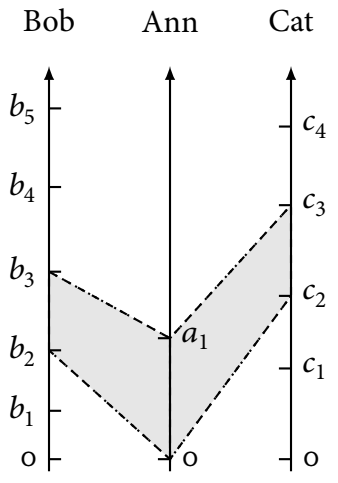


Figure 6: Transferring an increment from Bob to Cat with Ann as bridge

This means that we can take all of Bob's bananas (divided into m increments) and replace them with coconuts for Cat (m increments of coconut), maintaining the same moral value throughout. The end result: starting from any distribution, taking away all of Bob's bananas while giving Cat an equivalent amount of additional coconuts (measured in our specially constructed increments) preserves moral value.

We follow an analogous procedure when transferring holdings from Cat to Ann. Start with Bob having nothing. From our construction, when Cat loses one increment of coconut, this creates exactly as much moral loss as Bob gaining his base increment of banana creates moral gain. Due to separability, this equivalence holds true no matter how many apples Ann has. Similarly, when Bob loses his base increment of banana, this creates exactly as much moral loss as Ann gaining one increment of apple creates moral gain. Again by separability, this equivalence holds true regardless of how many coconuts Cat has.

These two equivalences together show us that: Cat losing one increment of coconut creates exactly as much moral loss as Ann gaining one increment of apple creates moral gain. By separability, this equivalence holds true regardless of what Bob holds, or how many apples or coconuts Ann and Cat currently have. This is illustrated in Figure 7.

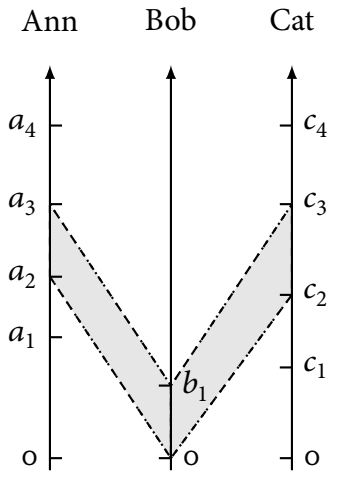


Figure 7: Transferring an increment from Cat to Ann with Bob as bridge

This means we can take all of Cat's coconuts (divided into $n + m$ increments) and replace them with more apples for Ann ($n + m$ increments of apple), maintaining the same moral value throughout. The end result: starting from any distribution, taking away all of Cat's coconuts while giving Ann an equivalent amount of additional apples (measured in our specially constructed increments) preserves moral value.

We have thus shown how to convert any distribution into equivalent holdings for a single person. When we have an outcome where Ann's holdings can be divided into o increments, Bob's holdings into n increments, and Cat's holdings into m increments, this is exactly as good as an outcome where Ann gets $o + n + m$ increments (and others nothing).

This conversion lets us compare any two outcomes. We start with a sufficiently small base increment: one small enough to serve as a measuring rod for dividing everyone's holdings in both outcomes into equivalent increments. We then convert each outcome into equivalent holdings for a single person. Because we assumed that more of your preferred fruit is always better, we can compare these converted outcomes simply by counting increments. By transitivity, the original outcomes must compare in the same way as these equivalent one-person outcomes.

The argument, when generalized, provides a representation of moral value in terms of sums of utilities. In this argument, utility numbers mean simply that, if someone's holdings can be divided into x increments (relative to our measuring rod), their utility is x. To say that one outcome has more total utility than another (and is therefore better) is simply to say that we can divide both outcomes into equivalent increments (defined in terms of some change for someone) such that one outcome

has more of these increments than the other. This should remove much of the mystery from the separability theorem.[39]

Our construction of utilities through the separability theorem raises the question: how do these utilities relate to well-being? The theorem defines utility in terms of morally equivalent changes in holdings, not in terms of any independently specified notion of well-being. The possibilities regarding this relationship are similar to those we discussed in §4. Well-being might fail to be interpersonally comparable or quantitatively measurable. In such cases, utilitarianism does not make sense but might be supported as a subjective requirement (if there is no interpersonal comparability) or through a stipulative way of making well-being quantitative (if there is no clear quantitative measure to begin with). If well-being is interpersonally comparable and quantitatively measurable, utilitarianism straightforwardly follows if everyone's utility is represented by the same linear function of well-being. But because the separability theorem does not appeal to probabilities, we can no longer argue for linearity through risk-neutrality in well-being.[40]

7. Summing Up

Harsanyi put forward three arguments for the summing part of utilitarianism. The social-aggregation theorem shows how moral value inherits additive structure from probability theory, with Ex-Ante Pareto ensuring this structure tracks individual evaluation. The impartial-observer theorem arrives at the same conclusion through a different route, using the idea of sympathetic imagination rather than actual probabilities. The separability theorem shows we can reach utilitarian conclusions even without appealing to probability, by constructing equivalent increments through pairwise comparisons.

These arguments present utilitarianism in a new light. The additive form is not a mysterious assumption but emerges from more basic principles. In the first two arguments, it comes from probability theory's mathematical structure. In the third, it comes from the construction of equivalent increments through separability. The

[39] The separability theorem was first proved by Fleming (1952, pp. 375–379) and discussed by Harsanyi (1955, pp. 309–312). Our approach follows Wakker 1989, pp. 41–77, which is similar to Broome 1991, pp. 82–89. A proof is also provided in Harsanyi 1977b, pp. 69–82. For a more mathematically complex but general approach, see Debreu 1959, §3. Wakker (1989, pp. 46–47) provides references to other proofs of similar results.

[40] In this context, Maskin (1978, p. 94) proposes the following kind of principle: if outcome x is at least as good as y, then multiplying everyone's well-being by positive a and adding b preserves this relationship. See also Blackorby et al. 1980, p. 27. Arguments for linearity using this last principle appear in Brown 2007, pp. 334–336 and Nebel 2021, pp. 584–585.

question then shifts from whether moral value has additive structure to how the utilities in this structure relate to well-being.

This relationship remains a crucial open question. If well-being lacks a clear quantitative structure or interpersonal comparability, utilitarianism might still be supported as a subjective requirement or through stipulation. If well-being is both quantitative and comparable, utilitarianism follows if utility is linear in well-being. The social-aggregation and impartial-observer theorems suggest arguing for this through risk-neutrality. And alternative routes can be found for the separability theorem.

Thus, while Harsanyi's contribution does not completely establish utilitarianism, it transforms how we think about aggregation in ethics. The key question becomes not whether to add up, but what to add up.

8. Further Reading

John Broome. *Weighing Goods: Equality, Uncertainty and Time*. Basil Blackwell, Oxford, 1991.

John Broome. General and Personal Good: Harsanyi's Contribution to the Theory of Value. In Iwao Hirose and Jonas Olson, editors, *The Oxford Handbook of Value Theory*, pages 249–266. Oxford University Press, Oxford, 2015.

John C. Harsanyi. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, 1977. (pp. 48–83)

Michael D. Resnik. *Choices: An Introduction to Decision Theory*. University of Minnesota Press, Minneapolis, 1987. (pp. 196–205)

9. Acknowledgements

We wish to thank Richard Yetter Chappell and Dean Spears for valuable comments.

References

Adler, Matthew (2012) *Well-Being and Fair Distribution*, Oxford: Oxford University Press.

Adler, Matthew D. (2019) *Measuring Social Welfare: An Introduction*, Oxford: Oxford University Press.

Adler, Matthew D. and Chris William Sanchirico (2006) 'Inequality and Uncertainty: Theory and Legal Applications', *The Journal of Political Economy* 155 (2): 279–377.

Alt, Franz (1936) 'Über die Meßbarkeit des Nutzens', *Zeitschrift für Nationalökonomie* 7 (2): 161–169.

— (1971) 'On the Measurability of Utility', in John S. Chipman, Leonid Hurwicz, Marcel K. Richter, and Hugo F. Sonnenschein, eds., *Preferences, Utility, and Demand: A Minnesota Symposium*, pp. 424–431, New York: Harcourt Brace Jovanovich.

Atkinson, Anthony B. and Joseph E. Stiglitz (1980) *Lectures on Public Economics*, London: McGraw-Hill.

Bentham, Jeremy (1789) *An Introduction to the Principles of Morals and Legislation*, London: T. Payne and Son.

Bernoulli, Daniel (1738) 'Specimen theoriae novae de mensura sortis', *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5:175–192.

— (1954) 'Exposition of a New Theory on the Measurement of Risk', *Econometrica* 22 (1): 23–36.

Blackorby, Charles, David Donaldson, and John A. Weymark (1980) 'On John Harsanyi's Defences of Utilitarianism', URL https://john-weymark.github.io/files/BDW80CORE.pdf.

Border, K. C. (1985) 'More on Harsanyi's Utilitarian Cardinal Welfare Theorem', *Social Choice and Welfare* 1 (4): 279–281.

Broome, John (1987) 'Utilitarianism and Expected Utility', *The Journal of Philosophy* 84 (8): 405–422.

— (1991) *Weighing Goods: Equality, Uncertainty and Time*, Oxford: Basil Blackwell.

— (2004) *Weighing lives*, Oxford: Oxford University Press.

— (2008) 'Can There Be a Preference-Based Utilitarianism?', in Marc Fleurbaey, Maurice Salles, and John A. Weymark, eds., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pp. 221–238, Cambridge: Cambridge University Press.

— (2015) 'General and Personal Good: Harsanyi's Contribution to the Theory of Value', in Iwao Hirose and Jonas Olson, eds., *The Oxford Handbook of Value Theory*, pp. 249–266, Oxford: Oxford University Press.

Brown, Campbell (2007) 'Prioritarianism for Variable Populations', *Philosophical*

Studies 134 (3): 325–361.
Coulhon, T. and Philippe Mongin (1989) 'Social Choice Theory in the Case of von Neumann-Morgenstern Utilities', *Social Choice and Welfare* 6 (3): 175–187.
Debreu, Gerard (1959) 'Topological Methods in Cardinal Utility Theory', Tech. Rep. 76, Cowles Foundation Discussion Papers.
Debreu, Gerard, Gerard Debreu, and Werner Hildenbrand (1983) 'Topological Methods in Cardinal Utility Theory', in *Mathematical Economics: Twenty Papers of Gerard Debreu*, Econometric Society Monographs, pp. 120–132, Cambridge University Press.
Diamond, Peter A. (1967) 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment', *Journal of Political Economy* 75 (5): 765–766.
Edgeworth, Francis Ysidro (1881) *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*, London: C. Kegan Paul & Co.
Ellingsen, Tore (1994) 'Cardinal Utility: A History of Hedonimetry', in M. Allais and O. Hagen, eds., *Cardinalism*, pp. 105–165, Dordrecht, Netherlands: Kluwer Academic Publishers.
Epstein, Larry G. and Uzi Segal (1992) 'Quadratic Social Welfare Functions', *Journal of Political Economy* 100 (4): 691–712.
Fishburn, Peter C. (1984) 'On Harsanyi's Utilitarian Cardinal Welfare Theorem', *Theory and Decision* 17 (1): 21–28.
Fleming, Marcus (1952) 'A Cardinal Concept of Welfare', *The Quarterly Journal of Economics* 66 (3): 366–384.
— (1957) 'Cardinal Welfare and Individualistic Ethics: A Comment', *The Journal of Political Economy* 65 (4): 355–357.
Fleurbaey, Marc (2018) 'Welfare Economics, Risk and Uncertainty', *Canadian Journal of Economics* 51 (1): 5–40.
Fleurbaey, Marc and Philippe Mongin (2016) 'The Utilitarian Relevance of the Aggregation Theorem', *American Economic Journal: Microeconomics* 8 (3): 289–306.
Fleurbaey, Marc and Alex Voorhoeve (2013) 'Decide as You Would with Full Information! An Argument Against Ex Ante Pareto', in Ole Norheim, Samia Hurst, Nir Eyal, and Dan Wikler, eds., *Inequalities in Health: Concepts, Measures, and Ethics*, pp. 113–128, Oxford: Oxford University Press.
Fontaine, Philippe (2010) 'The Homeless Observer: John Harsanyi on Interpersonal Utility comparisons and bargaining, 1950–1964', *Journal of the History of Economic Thought* 32 (2): 145–173.
Greaves, Hilary (2017) 'A Reconsideration of the Harsanyi–Sen–Weymark Debate on Utilitarianism', *Utilitas* 29 (2): 175–213.
Hare, R. M. (1982) 'Ethical Theory and Utilitarianism', in Amartya Sen and Bernard

Williams, eds., *Utilitarianism and Beyond*, pp. 23–38, Cambridge University Press.
Harsanyi, John C. (1953) 'Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking', *Journal of Political Economy* 61 (5): 434–435.
— (1955) 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economy* 63 (4): 309–321.
— (1975a) 'Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory', *American Political Science Review* 69 (2): 594–606.
— (1975b) 'Nonlinear Social Welfare Functions', *Theory and Decision* 6 (3): 311–332.
— (1977a) 'Morality and the Theory of Rational Behavior', *Social Research* 44 (4): 623–656.
— (1977b) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press.
— (1978) 'Bayesian Decision Theory and Utilitarian Ethics', *American Economic Review* 68 (2): 223–228.
— (1979) 'Bayesian Decision Theory, Rule utilitarianism, and Arrow's Impossibility Theorem', *Theory and Decision* 11 (3): 289–317.
— (1982) 'Morality and the Theory of Rational Behaviour', in Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond*, pp. 39–62, Cambridge University Press.
— (1992) 'Game and Decision Theoretic Models in Ethics', in *Handbook of Game Theory with Economic Applications*, vol. 1, pp. 669–707, Amsterdam: North Holland.
— (2008) 'John Rawls's Theory of Justice: Some Critical Comments', in Marc Fleurbaey, Maurice Salles, and John A. Weymark, eds., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pp. 71–79, Cambridge: Cambridge University Press.
Herstein, I. N. and John Milnor (1953) 'An Axiomatic Approach to Measurable Utility', *Econometrica* 21 (2): 291–297.
Jeffrey, Richard C. (1971) 'On Interpersonal Utility Theory', *Journal of Philosophy* 68 (20): 647–656.
Jensen, Karsten Klint (1995) 'Measuring the Size of a Benefit and its Moral Weight on the Significance of John Broome's: "Interpersonal Addition Theorem"', *Theoria* 61 (1): 25–60.
Jensen, Niels Erik (1967) 'An Introduction to Bernoullian Utility Theory: I. Utility Functions', *Swedish Journal of Economics* 69 (3): 163–183.
Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky (1971) *Foundations of Measurement, Volume I: Additive and Polynomial Representations*, San Diego: Academic Press.

Kreps, David M. (1988) *Notes on the Theory of Choice*, Boulder: Westview.
Leibniz, G. W. (1700) 'Observationes de principio juris', *Monatlicher Auszug*, pp. 371–382.
Marschak, Jacob (1950) 'Rational Behavior, Uncertain Prospects, and Measurable Utility', *Econometrica* 18 (2): 111–141.
Maskin, Eric (1978) 'A Theorem on Utilitarianism', *Review of Economic Studies* 45 (1): 93–96.
McMahan, Jefferson (1981) 'Problems of Population Theory', *Ethics* 92 (1): 96–127.
Mill, John Stuart (1969) *Essays on Ethics, Religion and Society*, vol. X of *Collected Works*, Toronto: University of Toronto Press.
Mongin, Philippe (2001) 'The Impartial Observer Theorem of Social Ethics', *Economics and Philosophy* 17 (2): 147–179.
Mongin, Philippe and Claude d'Aspremont (1998) 'Utility theory and ethics', in Salvador Barbera, Paul Hammond, and Christian Seidl, eds., *Handbook of Utility Theory Volume 1: Principles*, pp. 371–481, Dordrecht: Kluwer Academic Publishers.
Nebel, Jacob M. (2021) 'Utils and Shmutils', *Ethics* 131 (3): 571–599.
— (2022) 'Aggregation without Interpersonal Comparisons of Well-Being', *Philosophy and Phenomenological Research* 105 (1): 18–41.
Ng, Yew-Kwang (1975) 'Bentham or Bergson? Finite Sensibility, Utility Functions and Social Welfare Functions', *The Review of Economic Studies* 42 (4): 545–569.
Nord, Erik (1992) 'Methods for Quality Adjustment of Life Years.', *Social science & medicine* 34 (5): 559–569.
Otsuka, Michael (2015) 'Prioritarianism and the Measure of Utility', *Journal of Political Philosophy* 23 (1): 1–22.
Parfit, Derek (1995) *Equality or Priority*, University of Kansas, Department of Philosophy, The Lindley Lecture, 1991.
Patrick, Donald L., J. W. Bush, and Milton M. Chen (1973) 'Methods for Measuring Levels of Well-Being for a Health Status Index', *Health Services Research* 8 (3): 228–245.
Rabinowicz, Wlodek (2002) 'Prioritarianism for Prospects', *Utilitas* 14 (1): 2–21.
Rawls, John (1957) 'Justice as Fairness', *Journal of Philosophy* 54 (22): 653–662.
— (1971) *A Theory of Justice*, Cambridge: Harvard University Press.
— (1974) 'Some Reasons for the Maximin Criterion', *American Economic Review* 64 (2): 141–146.
Resnik, Michael D. (1983) 'A Restriction on a Theorem of Harsanyi', *Theory and Decision* 15 (4): 309–320.
— (1987) *Choices: An Introduction to Decision Theory*, Minneapolis: University of Minnesota Press.

Roemer, John E. (2008) 'Harsanyi's Impartial Observer Is Not a Utilitarian', in Marc Fleurbaey, Maurice Salles, and John A. Weymark, eds., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pp. 129–135, Cambridge: Cambridge University Press.

Selinger, Stephen (1986) 'Harsanyi's Aggregation Theorem without Selfish Preferences', *Theory and Decision* 20 (1): 53–62.

Sen, Amartya (1976) 'Welfare Inequalities and Rawlsian Axiomatics', *Theory and Decision* 7 (4): 243–262.

— (1977) 'Non-Linear Social Welfare Functions: A Reply to Professor Harsanyi', in Robert E. Butts and Jaakko Hintikka, eds., *Foundational Problems in the Special Sciences: Part Two of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada-1975*, pp. 297–302, Dordrecht: D. Reidel Publishing Company.

— (1986) 'Social Choice Theory', in *Handbook of Mathematical Economics*, vol. 3, pp. 1073–1181, Amsterdam: North Holland.

Sidgwick, Henry (1907) *The Methods of Ethics*, London: Macmillan, 7 edn.

Stefánsson, H. Orri (2023) 'In Defence of Pigou–Dalton for Chances', *Utilitas* 35 (4): 292–311.

Stigler, George J. (1950) 'The Development of Utility Theory. I', *Journal of Political Economy* 58 (4): 307–327.

Thomas, Teruji (2022) 'Separability and Population Ethics', in *The Oxford Handbook of Population Ethics*, pp. 271–295, Oxford University Press.

Torrance, George W., Warren H. Thomas, and David L. Sackett (1972) 'A Utility Maximization Model for Evaluation of Health Care Programs', *Health Services Research* 7 (2): 118–133.

Vickrey, William (1945) 'Measuring Marginal Utility by Reactions to Risk', *Econometrica* 13 (4): 319–333.

von Neumann, John and Oskar Morgenstern (1944) *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Wakker, Peter P. (1989) *Additive Representations of Preferences: A New Foundation of Decision Analysis*, Dordrecht: Kluwer.

Weirich, Paul (1983) 'Utility Tempered with Equality', *Noûs* 17 (3): 423–439.

Weymark, John A. (1991) 'A Reconsideration of the Harsanyi–Sen Debate on Utilitarianism', in Jon Elster and John E. Roemer, eds., *Interpersonal Comparisons of Well-Being*, pp. 255–320, Cambridge: Cambridge University Press.

— (1994) 'Harsanyi's Social Aggregation Theorem with Alternative Pareto Principles', in Wolfgang Eichhorn, ed., *Models and Measurement of Welfare and Inequality*, pp. 869–887, Berlin: Springer.

— (2005) 'Measurement Theory and the Foundations of Utilitarianism', *Social*

Choice and Welfare 25 (2/3): 527–555.