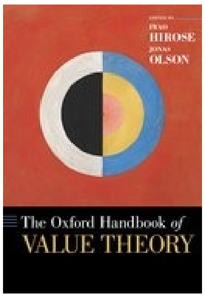


The Oxford Handbook of Value Theory

Iwao Hirose (ed.), Jonas Olson (ed.)

https://doi.org/10.1093/oxfordhb/9780199959303.001.0001

Published: 2015 **Online ISBN:** 9780199984978 **Print ISBN:** 9780199959303

CHAPTER

13 General and Personal Good: Harsanyi's Contribution to the Theory of Value

John Broome

https://doi.org/10.1093/oxfordhb/9780199959303.013.0014 Pages 249–266

Published: 07 April 2015

Abstract

In 1955 John Harsanyi proved a remarkable theorem that connects general good with the personal good of individuals. This chapter interprets Harsanyi's theorem. It explains the meaning of its conclusion, the way it links together intrapersonal and interpersonal aggregation, and in particular the link it makes between the value of avoiding risk and the value of avoiding inequality between people. It explains how the theorem connects prioritarianism with risk avoidance and utilitarianism with risk neutrality. It sets out the theorem's premises and assesses the various objections that have been made to them. Finally, it considers how far Harsanyi's theorem gives support to utilitarianism.

Keywords: utilitarianism, prioritarianism, risk avoidance, risk neutrality, utility, general good, personal good, Harsanyi's theorem

Subject: Moral Philosophy, Philosophy

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

13.1. Introduction

IN 1955, John Harsanyi published a singular contribution to the theory of value (Harsanyi 1955). He proved a theorem that links together the valuation of uncertain prospects for a single person and the valuation of distributions of good across people. The theorem's conclusion is important and remarkable; it is by no means obvious, and it requires some mathematics to uncover it. Perhaps as a consequence, philosophers of value have not always given this theorem the attention it deserves. This chapter describes and interprets the theorem, and explains its importance.

Harsanyi uses the language of economics, and he sets his argument in a framework that is generally taken for granted by economists but not widely accepted in philosophy. He assumes that each person's good consists in the satisfaction of her preferences. But his conclusion is about the relation between general good and the good of individual people. It is independent of particular assumptions about the nature of a person's good, and I shall present it in a way that does not depend on any such assumptions.[1]

13.2. The Theorem

An *outcome* is a state of affairs. Some outcomes are better than others: a relation of betterness holds among outcomes. For example, if the climate warms by two degrees, that is a better outcome than if it warms by three degrees.

p. 250 A *prospect* is a number of possible outcomes, each having some degree of likelihood. Some prospects, too, are better than others, and a relation of betterness holds among them. Betterness among prospects is less fundamental than betterness among outcomes, because what ultimately matters is what actually happens, rather than what might happen. Betterness among prospects is therefore derivative, but it is genuine all the same. A prospect in which it is likely that global warming will stay below two degrees is better than one in which it is likely to go above two degrees.

Not only does betterness simpliciter hold among outcomes and prospects, so does betterness for particular people. It is true of each particular person that some outcomes and prospects are better for her than others. The prospect in which it is likely that global warming will go above two degrees is better for some people—perhaps some of those who live in Siberia—than the prospect in which it is likely to stay below two degrees. Let us say that *personal betterness relations* hold among outcomes and prospects. And let us distinguish the relation of betterness simpliciter by calling it the *general betterness relation*.

For the moment, let us assume that general betterness supervenes on personal betterness for people. That is to say, if neither of two prospects is personally better for anyone than the other, then neither is generally better than the other. Moreover, let us assume that this supervenience is positive. That is to say, if one of two prospects is personally better for someone than the other, and worse for no one, then it is generally better. I call this assumption of positive supervenience the *principle of personal good*. I shall question it in section 13.8.

The personal and general betterness relations have various structural properties. For one thing, they are orderings. Precisely, they are strict partial orderings, which means they are irreflexive and transitive. That is to say: nothing is better than itself, and if one prospect is better than another, which is better than a third, then the first is better than the third.

Next, let us go much further and assume that all these relations satisfy all the axioms of expected utility theory. (I shall question this assumption in sections 13.6 and 13.7.) Different versions of expected utility theory[2] have different axioms. But Harsanyi's Theorem can be proved within many versions, so it does not matter precisely which set of axioms we adopt.

A technical note: the axioms are generally specified for the relation "better than or equally as good as," whereas I take the primitive relation to be betterness. Equality of goodness may be defined in terms of betterness like this: "*a* is equally as good as *b*" means that neither *a* nor *b* is better than the other, and that any third thing *c* is better than *a* if and only if it is better than *b*. This definition has some consequences that may be questioned. One is that, if nothing is better than anything else, then everything is equally good. Nothing in this chapter turns on this choice of primitive, and I shall not discuss it further.

I shall adopt some technical terminology from expected utility theory. Take a set of prospects and outcomes on which there is a betterness relation. Take a function that assigns numbers to prospects and outcomes.
p. 251 The function is said to *represent* the ↳ betterness relation when the number it assigns to one prospect or outcome is greater than the number it assigns to another prospect or outcome if and only if the former is

better than the latter. When a function represents a betterness relation, I call it a *utility function*, and I call the numbers it assigns to prospects and outcomes *utilities*. This defines exactly what I mean by these terms. Many economists use the word "utility" as a synonym for a person's "good," and this practice has recently been spreading among philosophers. It is not my meaning. A person's utilities are defined to represent the person's betterness *order*, not to represent the *quantity* of her good. The question of whether or not they also represent the quantity of her good is a substantive issue that will be considered in section 13.9.

Another piece of terminology: a utility function is said to be *expectational* when the utility it assigns to a prospect is the mathematical expectation of the utilities it assigns to the prospect's possible outcomes.

I have assumed the principle of personal good and I have assumed that both general betterness and personal betterness satisfy the axioms of expected utility theory. These assumptions together have a remarkable consequence. They imply that general betterness can be represented by a utility function that is the sum of utility functions that represent the personal betterness of each person. In brief, general utility can be treated as the total of personal utilities. In symbols:

$$U\left(x\right) = u_1\left(x\right) + u_2\left(x\right) + \ldots + u_n\left(x\right). \tag{*}$$

Here, x is a prospect or outcome, $U(x)$ is its general utility, and $u_1(x) \ldots u_n(x)$ are the personal utilities of all the people. Furthermore, all the utility functions are expectational.

This is Harsanyi's Theorem. Harsanyi's proof was built on Jacob Marschak's (1950) version of expected utility theory, which assumes there are objective chances. But the theorem is robust; it can also be proved in versions that allow for subjective credences.[3]

13.3. Interpretation

What does this theorem tell us? The first thing is displayed on its face by formula (*): this formula has an additive structure. In technical language, it is *additively separable*. An outcome can be evaluated by first evaluating it from the point of view of each person separately, and then adding up the separate evaluations.

This is in itself an impressive conclusion. Addition is a very special operation. Philosophers sometimes take it to be the default mode of combining quantities together. They assume that, by default, a whole is the sum of its parts. When it is not, they think this needs some special explanation such as "organic unity" (see p. 252 Carlson, chapter 15 in this volume). But when addition obtains, this too needs explanation. Why should ↳ the individual utility functions be combined by addition? None of the premises of Harsanyi's Theorem—neither the principle of personal good nor the axioms of expected utility theory—mentions addition. The additive structure arises from the mathematics, in a not very intuitive fashion.[4]

The debate between prioritarians and strict egalitarians about the value of equality is a debate about additive separability (see Holtug, chapter 14 in this volume). Strict egalitarians believe that general betterness depends partly on how people fare relative to each other. A strict egalitarian formula for general utility would contain some terms that embody comparisons between different people's situations. For example, it might contain some measure of the dispersion among individual utilities, such as their variance or the Gini coefficient. Being additively separable, formula (*) contains no such terms.

This does not immediately imply that the formula is opposed to strict egalitarianism. Each person's betterness relation might itself be influenced by the person's standing in comparison to other people. For example, suppose that in some outcome a person is worse off than other people who are no more deserving than her. This may be an unfairness she suffers. Suffering an unfairness is presumably bad for her, so it will influence her personal betterness relation, and will be registered in her own utility. Formula (*) does not rule that out. But it does rule out *communal* egalitarianism, which is the view that equality is a sort of good that belongs to the community as a whole rather than to the individual members of the community (Broome 1991: § 9.2). Harsanyi's Theorem is opposed to strict communal egalitarianism about good.

The second point of interpretation is more difficult. Start by concentrating on the utility function of one person—say the first, $u_1()$. Suppose some outcome a is better for this person than b, which is in turn better for her than c. Because the person's utility function represents her betterness, $u_1(a)$ is greater than $u_1(b)$, which is greater than $u_1(c)$. This is just to say that utilities represent the *order* of the person's good.

But there is more. Utilities are also assigned to prospects. Let us compare two particular prospects. One is the prospect that has a and c as possible outcomes, and gives them equal chances of one-half each. Call this prospect "Gamble." The second is a simple, "degenerate" prospect that has only one possible outcome, b. In this prospect b is certain; call it "Certainty." Suppose Gamble is better on balance for the person than Certainty. What does that tell us?

Gamble is better for the person than Certainty in one respect: it offers a one-half chance of the best outcome a, whereas Certainty offers only the less good b. Gamble is worse for the person than Certainty in another respect: it offers a one-half chance of the worst outcome c, whereas Certainty offers the better b. Whether Gamble is better or worse for the person on balance is determined by putting together the respect in which it is better with the respect in which it is worse. These respects are *aggregated*, we may say, which in this case means they are weighed against each other. The difference in goodness between a and b is weighed against the difference in goodness between b and c.

We are supposing Gamble is better on balance. This means that the difference in goodness between a and b p. 253 counts for more in this aggregation—has a greater weight—than ↳ the difference in goodness between b and c. The fact that one prospect is better than the other tells us how differences in goodness weigh or count in this particular aggregation. The example is a simple one, but the point can be generalized. In general, the relation of betterness among prospects provides a basis for weighing up differences of goodness.

Moreover, utility measures how much these differences in goodness count. The example shows how. Since utility represents betterness and Gamble is better than Certainty, Gamble's utility is higher than Certainty's. Since utility is expectational, the utility of each prospect is the mathematical expectation of the utility of its outcomes. So the utility of Gamble is $½u_1(a)$ + $½u_1(c)$ and the utility of Certainty is $u_1(b)$. Since Gamble is better,

$$½u_1(a) + ½u_1(c) > u_1(b).$$

That is to say:

$$u_1(a) - u_1(b) > u_1(b) - u_1(c).$$

The difference in utility between a and b is greater than the difference in utility between b and c. This represents the fact that the difference in goodness between a and b outweighs, or counts for more than, the difference in goodness between b and c.

Remember this is only how much these differences count in one particular sort of aggregation: aggregation in the context of uncertainty, where a prospect is evaluated from the point of view of a single person. The place of a prospect in the person's betterness ordering depends on aggregating together the different possible outcomes that make up the prospect. Utility measures how much differences in goodness count in this sort of intrapersonal aggregation. We may say it measures a sort of *contributory value* that an outcome has: the contribution the outcome makes to the value of a prospect.

Harsanyi's Theorem tells us about another, interpersonal sort of aggregation, where different people's goods are aggregated and weighed against each other. The formula (*) in Harsanyi's Theorem describes aggregation across people; it specifies how the good of different people goes together to make up general good.

For example, suppose the outcome *d* is better for the first person than another outcome *e*, whereas *e* is better for the second person than *d*. Assume that *d* and *e* are equally good for everyone else. In one respect, *d* is better than *e*—it is better for the first person—whereas in another respect *e* is better than *d*—it is better for the second person. Whether *d* or *e* is better on balance is determined by putting together the respect in which it is better with the respect in which it is worse. This is a matter of aggregating across people: of putting the first person's good together with the second's.

According to Harsanyi's Theorem, the result of this aggregation is given by the total of utilities. Outcome *d* is better than outcome *e* if its total utility is greater. That is to say, if

$$u_1(d) + u_2(d) > u_1(e) + u_2(e).$$

p. 254 In other words, if

$$u_1(d) - u_1(e) > u_2(e) - u_2(d).$$

So *d* is better than *e* if the difference in the first person's utility between *d* and *e* is greater than the difference in the second person's utility between *e* and *d*. In general, differences in people's utility measure how much differences in people's good count in aggregating across people. Utility measures the contribution a person's good makes in this sort of aggregation.

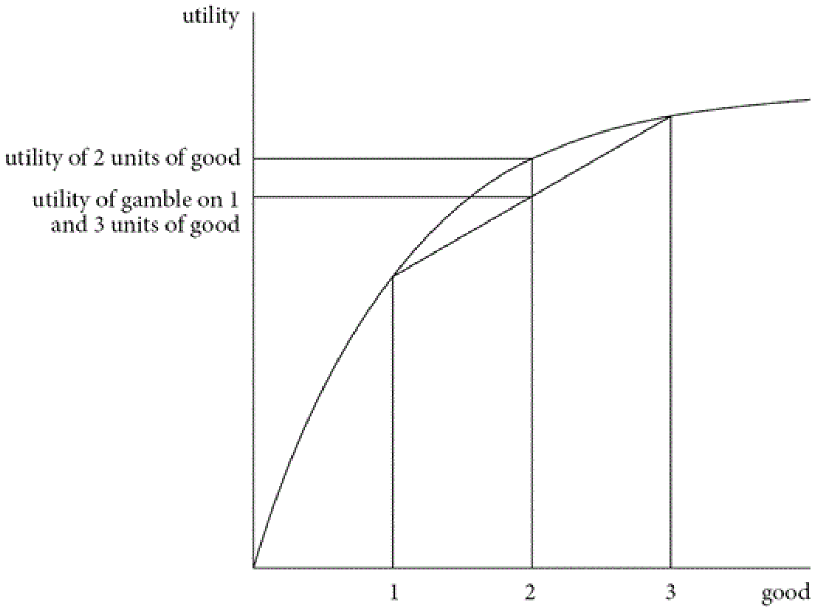
This is the same utility as measures the contributory value of a person's good in aggregation under uncertainty. Harsanyi's Theorem links together the two sorts of aggregation. A person's good has the same contributory value in both. This is remarkable. On the face of it, aggregating under uncertainty and aggregating across people are quite different matters. Harsanyi showed they are linked, so long as the principle of personal good holds and both sorts of betterness satisfy the axioms of expected utility theory.

13.4. Examples: Prioritarianism and Utilitarianism

To illustrate the significance of this conclusion, let us imagine for a moment that we have some quantitative concept of a person's good: we attach meaning to quantities of good. Moreover, let us imagine that this meaning is portable between people, so we can say that a unit of good is equally good for each person. I shall consider in section 13.9 where this quantitative concept might come from; for the moment let us not question it.

Since utility represents betterness, the utility of an outcome is an increasing function of the outcome's goodness for the person. This means the graph of utility against good slopes upwards. Figures 13.1 and 13.2 show two different examples.

Figure 13.1



Risk Aversion and Prioritarianism

Suppose the graph curves downward, as it does in figure 13.1. Take a certainty of two units of good, and compare it with an uncertain prospect that has the same mathematical expectation of good: specifically, a gamble at equal odds between one unit of good and three units. The expected utility of the gamble is the average utility of one unit and three units; it is shown in the diagram. It is less than the utility of the certainty of two units. It follows that the gamble is worse than the certainty, even though its expected goodness is the same. The very uncertainty of the gamble counts against it. In a sense, uncertainty about good is a bad thing. Between any two prospects that have the same expectation of good, the one with less uncertainty is better. This is *risk aversion* about good, to use a common term. It is a consequence of the downward curvature of the utility graph.

Utility measures how much good counts in aggregation under uncertainty. The graph in figure 13.1 shows p. 255 that an increase in good counts for more the less well off the person ↳ is. A one-unit increase from one to two units counts for more than a one-unit increase from two to three units. Good has *diminishing marginal utility*, as economists say. It is this feature of the downward curvature that leads to risk-aversion about good.

Figure 13.1 shows the good of one person only. Now suppose that the relation between good and utility has the same curved shape for everyone. One implication is that risk aversion applies to everyone's good; this is a matter of intrapersonal aggregation under uncertainty. Harsanyi's Theorem tells us much more: it tells us that the same utilities also determine how good is interpersonally aggregated across people. The downward-curving shape of the utility function means that an increase of good counts for more the less p. 256 well off the person is, in aggregation across people as well as ↳ in aggregation under uncertainty. Diminishing marginal utility applies to both sorts of aggregation.

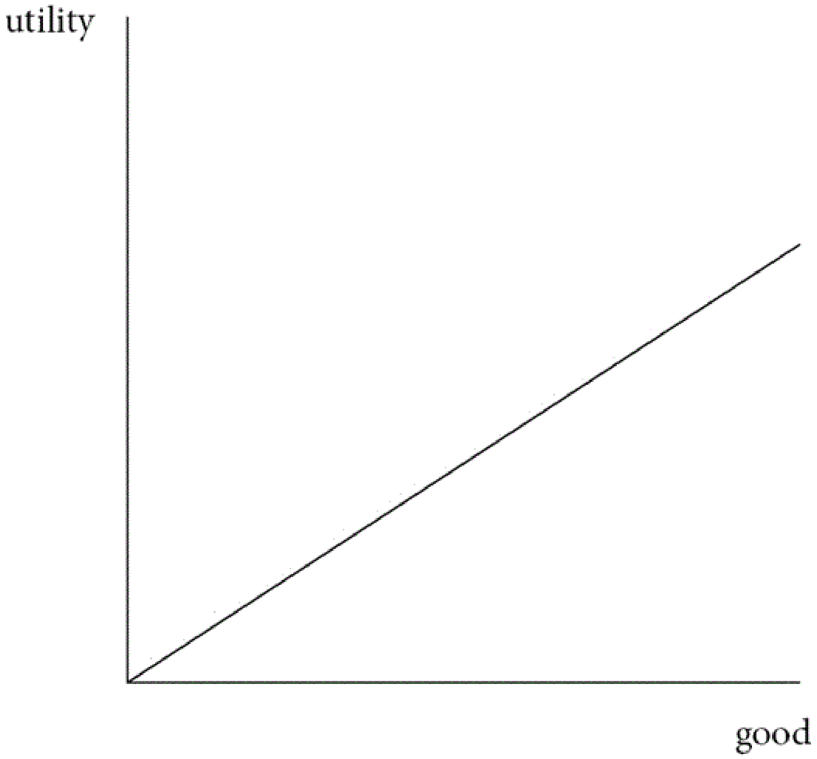
Although I have assumed that the relation between utility and good has the same shape for everyone, I am not yet entitled to assume it is exactly the same relation for everyone. Utility can always be rescaled: if one function represents a person's betterness, so does any other function that is a positive multiple of that function. (The origin of the function can also be changed, but that fact makes no difference to the interpretation of Harsanyi's Theorem, and I shall ignore it.) So betterness can be represented by a family of functions, each a rescaling of the others. Harsanyi's Theorem says that, for each person, *one* utility function can be selected out of the family of functions that represents her betterness such that adding up these function across all the people gives us a utility function that represents general betterness. It does not say that the same function can be selected for each person.

However, we may supplement the theorem by assuming impartiality between people. Take a distribution of quantities of good across people, and imagine permuting the quantities among the people, so that the quantities remain the same, but some of them end up being possessed by different people. Impartiality is the claim that general good remains the same: what matters is quantities of personal good, not which particular people possess them. Impartiality implies that Harsanyi's Theorem selects the same utility function for each person. Given that, and given that the function shows diminishing marginal utility, the formula (*) in Harsanyi's Theorem is a *prioritarian* value function. It tells us that increasing the good of people who have more counts for less than increasing the good of people who have less.

A consequence of prioritarianism is that transferring a quantity of good from a better-off person to a less well-off person makes the world generally better. That is to say, for a given total of people's good, it is better that it is more rather than less equally distributed. Inequality is in this way a bad thing. For this reason, prioritarianism has traditionally been known among economists as *inequality aversion*. Harsanyi's Theorem tells us that prioritarianism or inequality aversion is tightly linked with risk aversion, provided the premises of the theorem hold. This is an important implication of the theorem.

Alternatively, the relation between a person's utility and her good might be linear rather than curved, as it is in figure 13.2. This implies *risk neutrality* about good. Given the supplementary assumption of impartiality, the formula (*) in Harsanyi's Theorem is then a *utilitarian* value function. It values the arithmetic total of people's good. Harsanyi's Theorem tightly links utilitarianism with risk neutrality.

Figure 13.2



Risk Neutrality and Utilitarianism

13.5. Accepting or Rejecting the Conclusion

Many authors resist the tight link Harsanyi’s Theorem makes between interpersonal aggregation across p. 257 people and intrapersonal aggregation under uncertainty. There ↳ seems to be something wrong with “adopt[ing] for society as a whole the principle of rational choice for one man,” as John Rawls (1971: 26–27) put it. Interpersonal aggregation is a moral matter, whereas intrapersonal aggregation is prudential. More specifically, interpersonal aggregation seems obviously to have an aspect of fairness that has no place in intrapersonal aggregation. This point has been made repeatedly in various ways by many authors from Peter Diamond (1967) to Michael Otsuka and Alex Voorhoeve (2009).

Nevertheless, the premises of Harsanyi’s Theorem are plausible on the face of it, so the theorem constitutes a good argument for its conclusion. At least it sets a challenge to anyone who denies the link between interpersonal and intrapersonal aggregation, or more specifically the link between risk aversion and prioritarianism. She needs to identify which of the premises she rejects, and why. For example, when Otsuka and Voorhoeve (2009: 176) explicitly deny the conclusion of Harsanyi’s Theorem by saying

> Some shift is justified in the priority we give to benefiting a person if she is very badly off rather than somewhat badly off when we move from the case of the isolated person to the interpersonal case,

they should also say which of the theorem’s premises they deny.

It is not that the premises are unquestionable; several are open to objections. The significance of the theorem is not that its conclusion is indubitable. Its significance is to present us with a menu of options, of which we must select one. We may accept the theorem’s conclusion, or we may choose one of its premises to reject.

My own view, set out in Broome (1991), is that we should accept the conclusion. This does not mean denying that interpersonal aggregation has an aspect of fairness that intrapersonal aggregation lacks. It means recognizing fairness as a personal good and unfairness as a personal harm. When a distribution of good among people is unfair in some way, this unfairness diminishes the good of individual people. The damage unfairness does to overall goodness appears in the damage it does to the good of individuals, not in the aggregation of good across people.

For example, insofar as it is a bad thing for good to be distributed unequally among people, it is bad because it is unfair to some people. It reduces those people’s good. The badness of inequality is suffered by individuals; it is not some sort of communal badness. That is my account.

Philosophers sometimes think that inequality of good between people cannot itself reduce the good of people, because each person’s goodness must first be determined in order to determine what inequality obtains between different people’s good. But that is a fallacy. Each person’s good and inequality between people’s goods can be simultaneously determined. Here is a crude example taken from Broome (1991: 182). There are just two people, whose respective goods, g_1 and g_2, are given by

$$g_1 = \bar{g}_1 - \max\{0,\ \tfrac{1}{2}(g_2 - g_1)\}$$

p. 258 and

$$g_2 = \bar{g}_2 - \max\{0,\ \tfrac{1}{2}(g_1 - g_2)\},$$

where $\bar{g}_1$ and $\bar{g}_2$ are their respective goods apart from the matter of inequality. If, say, $\bar{g}_1 = 2$ and $\bar{g}_2 = 3$, solving these equations shows that $g_1 = 1$ and $g_2 = 3$.

In that way, fairness can be made consistent with Harsanyi's Theorem.

13.6. Rejecting Completeness

Anyone who nevertheless still wants to reject the conclusion of Harsanyi's Theorem must choose which of the premises to reject. I shall next review the options.

The premise that personal betterness satisfies the axioms of expected utility theory is rarely questioned in this context. Expected utility theory has a strong intuitive attraction in its application to personal good and I shall raise no doubts about it here. However, its application to general good is different. In that application, there are special grounds for questioning the axioms of expected utility theory, and many authors reject one or another of them.

One axiom open to doubt is *completeness*, which says that, of any two prospects, either one is better than the other or they are equally good. This may be doubted on the grounds that the relative contributory values of different people's goods may not be fully determinate. When one of two outcomes is better than the other for one person and worse for another person, which of these outcomes is generally better will depend on the contributory value of one person's good relative to the other's. Many authors doubt that there is always a determinate result.

This is not necessarily a doubt about interpersonal comparability of good. The contributory value of good is not necessarily the same as good itself. Take two people who live very similar lives, but at different dates in history. It might be determinate that their lives are equally good, but not determinate what relative weight they have in determining general good. Some people believe in "discounting" good that comes later in time.[5] They would think the good of the later-living person counts for less. If they also think it is indeterminate what is the right rate of discount, they would think it indeterminate what relative contributory value the two lives possess. However, although this is a possible view, I know of no one who has adopted it. Doubts about completeness of the general betterness ordering arise much more commonly from doubts about the interpersonal comparability of good.

If the general betterness relation is incomplete, that does not necessarily vitiate Harsanyi's Theorem. It depends on how radical the incompleteness is. Some economists believe, or at least profess, that no comparisons at all can be made between the goods of different people: it is never true of two people that one
p. 259 is ↳ better off than the other.[6] If that were so, Harsanyi's Theorem would be completely empty. But if the incompleteness is less radical, the theorem can still have some significance.

It can still be applied in a supervaluationist manner. To explain how, I need first to reformulate the conclusion of Harsanyi's Theorem. I explained in section 13.4 that each person's betterness relation can be represented by a whole family of utility functions, each a rescaling of the others. The theorem says that there are functions, one for each person, that add up to a general utility function. The theorem itself picks the functions, we might say. But now I shall arbitrarily pick in advance one utility function for each person. Let these functions be $u_1()$, $u_2()$, and so on. Then the conclusion of the theorem, given the premises, is that general betterness can be represented by a weighted sum of these functions. There are positive weights a_2, a_3, and so on such that general utility is:

$$U(x) = u_1(x) + a_2 u_2(x) + a_3 u_3(x) + \ldots + a_n u_n(x). \tag{**}$$

(I have scaled the general utility function to make $a_1 = 1$.) This is the same theorem, presented in a different form.

Now, suppose the general betterness ordering is incomplete. Its gaps can be filled in. That is to say, so long as it satisfies the other axioms of expected utility theory, it can be extended to a complete ordering that satisfies all the axioms. Normally, there will be many different extensions of the ordering that have this property. One outcome or prospect is generally better than another if and only if it better according to every one of these extensions.

Harsanyi's Theorem applies to each one. The theorem tells us that, for each extension there will be weights a_2, a_3, and so on such that (**) is true. For each extension, each outcome or prospect x is assigned a general utility by the function $U(x)$, determined as a weighted sum of personal utilities through (**). One outcome or prospect is generally better than another if and only if it is better according to the general utility function determined by every extension.

If the indeterminacy is not great, the weights determined by the different extensions will not be very different from each other. Nor, therefore, will the general utilities determined for each extension through (**). Supervaluation would allow us still to think of overall general utilities, though they would be a bit indefinite. Some of the lessons I have drawn from the case where general betterness is complete will carry over. It will still be fair to say that personal utilities, which measure the contributory value of betterness in intrapersonal aggregation under uncertainty, also contribute to interpersonal aggregation.

So even if the general betterness ordering is incomplete to some extent, Harsanyi's Theorem still makes some link between intrapersonal and interpersonal aggregation. Giving up the assumption of completeness is therefore not a very effective strategy for someone who wishes to deny this link.

p. 260

13.7. Rejecting Strong Independence

A second axiom open to doubt is known as the *strong independence axiom* or *sure-thing principle*. One part of this axiom is as follows. Suppose two different outcomes are equally as good as each other. Then a prospect that is a "mixture" of the two—leading to either one or the other of them—is equally as good as each of them.

Peter Diamond (1967) raised a powerful objection to this premise. He argued that it cannot recognize the sort of fairness that can sometimes be achieved by a random lottery. Suppose two people each have a claim to some valuable thing. Let it be to life-saving. Suppose each person is in mortal danger, and one can be saved but the other will die. Suppose their claims to being saved are equal. Fairness requires their equal claims to be equally satisfied. But that is impossible if one is saved, because the other will not be. It will normally be plainly wrong to save neither, but if one is saved some unfairness cannot be avoided. Nevertheless, a partial fairness can be achieved by holding a lottery: each claimant can be given the same chance of being saved. That is fairer than simply saving one or the other without a lottery.

One possible outcome is that the first claimant is saved. Another is that the second claimant is saved. Let us suppose these two outcomes are equally as good as each other. A lottery is a mixture that leads to either one or the other of these outcomes. So according to the strong independence axiom, it should be equally as good as each of the two possible outcomes. But it is actually better than both of them because it achieves a partial fairness, whereas saving one of the claimant's without holding a lottery does not achieve this sort of fairness. So the strong independence axiom is false. That is Diamond's argument.

It can be answered by treating fairness as a personal good and unfairness as a personal harm, in the way I recommended in section 13.5. The lottery leads to either one person's being saved or the other's being saved. But these outcomes are not exactly the ones that would be achieved by saving one or the other person without a lottery. If there is a lottery, the people are treated fairly to some extent. This is a good that is done them and it adds to their overall good. If there is no lottery, they do not have this good. So the lottery is not a mixture of the two outcomes in which there is no lottery. Therefore the strong independence axiom is not violated.

This response to Diamond is not universally accepted. I know of two different theories of value that are constructed by generalizing Diamond's objection. One comes from Larry Epstein and Uzi Segal (1992); the other from David McCarthy (2006). Both imply that a prospect that is a mixture of two equally good outcomes is always better than each of the outcomes. That is to say, they consistently reject the strong independence axiom. McCarthy's theory is a version of prioritarianism; he calls it *ex-ante prioritarianism*. It has the advantage of being immune to the strictures against prioritarianism that appear in section 13.9 below.

p. 261

13.8. Rejecting the Principle of Personal Good

The most commonly questioned premise of Harsanyi's Theorem is the principle of personal good.

One reason to doubt it is that there are goods other than the good of people, and these must contribute to general good. The good of nonhuman animals is a clear example. It is also easy to take account of in the theorem. We have only to include nonhuman animals along with people, and extend the principle of personal good to make it a principle of personal and animal good. The theorem remains valid with this amendment. Other sorts of good can be accommodated in the same way: the good of ecological systems, for example, if it really exists.

Once extended in this way to include nonpersons, the principle of personal good can seem hard to doubt. Overall good surely depends in a positive way on the good of people, animals, and whatever else has a good that should be counted. However, the premise of Harsanyi's Theorem is that the principle applies to both outcomes and prospects. Several authors doubt it when applied to outcomes, even thought they accept it when applied to prospects.

An argument for applying the principle of personal good to prospects appears in Broome (1991: ch. 8). The arguments I have seen against doing so are indirect. They do not find an independent fault with the principle. Instead, they fault it because it joins with other assumptions to imply Harsanyi's Theorem. In this way, they are not as powerful as Diamond's objection to the sure-thing principle, because Diamond makes the independent objection I have described.

Wlodek Rabinowicz (2002) provides an example. He rejects the conclusion of Harsanyi's Theorem on the familiar grounds I mentioned in section 13.5, that interpersonal and intrapersonal aggregations of good are not as closely parallel as the theorem implies they are. He recognizes that one of the premises of the theorem consequently has to go. He finds the principle of personal good less secure than the others, so this is the one he rejects.

Marc Fleurbaey and Alex Voorhoeve (2013) argue in the same way. They start by rejecting the conclusion of Harsanyi's Theorem on the familiar grounds. Then they adopt a principle they call "the principle of full information," which is nothing other than the sure-thing principle or strong independence axiom.[7] This principle is perfectly consistent with the principle of personal good, so it cannot on its own constitute an objection to that principle. However, the sure thing principle (with the other axioms of expected utility theory), together with the principle of personal good, implies the conclusion of Harsanyi's Theorem; this is just Harsanyi's Theorem itself. So if the sure-thing principle (and the other axioms) are true, and the conclusion of Harsanyi's theorem is false, it follows that the principle of personal good is false. This is Fleurbaey and Voorhoeve's argument against it.

p. 262 All these authors reject the principle of personal good only when it is applied to uncertain prospects. They accept it when it is applied to outcomes; presumably they find it too plausible there to reject it. This leaves them with a puzzle to deal with, as Rabinowicz (2002) recognizes. The outcomes that the principle applies to must contain no uncertainty; otherwise the same objection would recur for them. They will have to be fully specific possible worlds, with all details specified through all of history. We do not in practice encounter outcomes like that. The outcome of any act, or of anything that happens, is uncertain to some extent. Was the fine day's walking you enjoyed yesterday better than staying at home? Who knows? No doubt it was fun, but perhaps it triggered some small change in your body that will eventually take you to an early grave.

So the principle of personal good, which is so plausible, is left with no direct practical applications. This is not fatal to the argument, but it leaves work to be done if the principle is to be applied to outcomes and not prospects. We must be sure that our theory of value can be properly founded on outcomes of this sort that we do not encounter. Jeffrey's decision theory, for one, makes no distinction between prospects and outcomes. Indeed, Bolker's (1966) axiomatization of it is based on an "atomless" set of prospects, which contains no outcomes. So within this theory, a principle of personal good that applies to outcomes but not prospects cannot even be formulated.

13.9. Utility and Goodness

To describe prioritarianism in section 13.4, I presumed we have a quantitative concept of a person's good. This is a big presumption, and we should not make it without some idea of how it might be satisfied. Where could this quantitative concept come from?

First, what do I mean when I call a concept *quantitative*? I mean that the degree to which something possesses the property denoted by the concept can be measured on what is called a "cardinal scale." This means in turn that differences between degrees can always be intelligibly compared.

Take some concept that has an intelligible comparative. For instance, take the concept of heaviness whose comparative is heavier than. Some things are heavier than others, which is to say that things differ in their degree of heaviness. The concept is quantitative if we can always make sense of comparisons between these differences: if we can always make sense of the claim that one difference is greater than another. Heaviness is quantitative if we can always make sense of the claim that the difference in heaviness between one thing *A* and another *B* is greater than the difference in heaviness between a third thing *C* and a fourth *D* (where *A* is heavier than *B* and *C* heavier than *D*). For heaviness we can indeed always make sense of this claim, so heaviness is quantitative. For heaviness, the claim means that, if *A* and *D* were put together on one pan of a pair of scales, ↳ they would outweigh *B* and *C* put together on the other pan. This is how differences in heaviness are comparable. p. 263

By contrast, our ordinary concept of hardness is not quantitative in this sense. It has the intelligible comparative harder than, but we cannot always make sense of comparisons of differences in hardness. Steel is harder than copper and oak is harder than pine, but we could not ordinarily make sense of the claim that the difference in hardness between copper and steel is greater than the difference in hardness between pine and oak.

A concept can sometimes be made quantitative by finding a way to make sense of comparisons of differences. When this is an innovation, it modifies the concept to make it a tighter, more precise one. Heaviness was presumably made quantitative and tightened up by the invention of scales. There are often alternative ways to make a concept quantitative. For purposes of science, various different scales of hardness have been developed, using different means for comparing differences. Each alternative provides a different quantitative concept of hardness.

What basis do we have for a quantitative concept of a person's good? How can we make sense of comparisons of differences in a person's good? I have already described one way in my example of Gamble and Certainty. The difference in goodness between outcomes *a* and *b* can be compared with the difference in goodness between *c* and *d* (where *a* is better than *b* and *c* better than *d*) by comparing the goodness of particular uncertain prospects. The goodness of a gamble at equal odds between *a* and *d* can be compared with the goodness of a gamble at equal odds between *b* and *c*. If the first is better than the second, the difference in goodness between *a* and *b* is greater than the difference in goodness between *c* and *d*.

That is not actually what I said when describing the example of Gamble and Certainty. I presented the example as a way of determining how much differences of goodness *count* in aggregation under uncertainty. I did not present the example as a way of making sense of differences in goodness themselves. I concluded that the difference in goodness between *a* and *b* counts for more than the difference in goodness between *b* and *c*. I did not conclude that the difference in goodness between *a* and *b* is greater than the difference in goodness between *b* and *c*. Why not?

Out of caution. Aggregation of differences under uncertainty is indeed one way to make sense of differences of goodness. So it can provide one quantitative concept of good. If we took this route, utility, which I defined to measure how much goodness counts, would actually be a measure of goodness itself. But there might be alternative, rival ways to make sense of differences, which would provide rival quantitative concepts of good. A case could be made for adopting the measure given by utility, but the case would be questionable if there was a rival.

For example, comparing differences in good for one person with differences in good for another person, in the process of aggregating good across people, might provide a rival scale. However, Harsanyi's Theorem tells us that the same utility functions as specify how good is aggregated under uncertainty also specify how good is aggregated across people. So, provided the premises of Harsanyi's Theorem hold, no rival scale arises from ↳ aggregation across people. This very much strengthens the case for taking utility to measure goodness itself. My own view is that the case is strong enough. We should treat utility as a scale of good. (See also Broome 2004: 86–91; Greaves, forthcoming; McCarthy 2006; Jensen 1995.)

Graphically, this means that the curved graph in figure 13.1 relating utility to goodness makes no sense. Since utility is nothing other than goodness itself, that graph has to be a straight line. One consequence is that risk aversion about good also makes no sense.

Another is that prioritarianism makes no sense if the premises of Harsanyi's Theorem hold. Prioritarianism relies on a distinction between a person's good and how much the person's good counts in aggregation between people. But given Harsanyi's Theorem and the identification of utility with good, no such distinction can be made.

I cannot rule out the existence of some other means to make sense of comparisons of differences that would be a genuine rival to utility as a measure of goodness. But at least a prioritarian who accepts the premises of Harsanyi's Theorem needs to explain what means she has in mind.

13.10. Conclusion

If utility does indeed measure goodness, then the formula (*), together with the supplementary assumption of impartiality mentioned in section 13.4, is the utilitarian theory of value. It says that general good is the total of the good of the people. Harsanyi's Theorem constitutes an argument for utilitarianism. His paper should be considered one of the founding documents of utilitarianism.

Notes

[1] My book *Weighing Goods* (Broome, 1991) is a fuller presentation.

[2] For example, von Neumann and Morgenstern (1944), Marschak (1950), Savage (1954), Jeffrey (1965).

[3] Mongin (1995) contains a proof within Savage's (1954) version. Broome (1990) contains a proof within the Bolker-Jeffrey version (Bolker, 1966, 1967; Jeffrey, 1965).

[4] Interestingly, it arises from very different mathematical sources in different versions of the theorem. In Savage's decision theory, it follows from a theorem about "crosscutting separability," proved in Gorman (1968)—see the sketch proof in Broome (1991). In Jeffrey's decision theory, it follows from a theorem within measure theory, proved by Liapounoff (1940)—see the proof in Broome (1990). There must be a deep parallel in these theorems, but it is beyond my mathematical ability to identify it. Chapter 4 of Broome (1991) gives the most intuitive explanation of additivity that I can find.

[5] This claim is implicit in a lot of economics, but few economists have made it explicitly. One who has is Arrow (1999).

[6] For example, Arrow (1963: 9).

p. 265 [7] One part of the principle of full information as Fleurbaey and Voorhoeve state it is: "When one knows that, in every state of the world with positive probability, one is indifferent between two alternatives, then one should be indifferent between these alternatives." Compare the part of the strong independence axiom that I stated in section 13.7: "Suppose two different outcomes are equally as good as each other; then a prospect that is a 'mixture' of the two—leading to either one or the other of them—is equally as good as each of them." These are the same claim, differently worded.

References

Arrow, K. J. (1963). *Social Choice and Individual Values*, 2nd ed. New Haven: Yale University Press.
Google Scholar Google Preview WorldCat COPAC

Arrow, K. J. (1999). “Discounting, Morality, and Gaming.” In P. R. Portney and J. P. Weyant (eds.), *Discounting and Intergenerational Equity*. New York: Resources for the Future, 13–21.
Google Scholar Google Preview WorldCat COPAC

Bolker, E. D. (1966). “Functions Resembling Quotients of Measures.” *Transactions of the American Mathematical Society* 124: 292–312.
Google Scholar WorldCat

Bolker, E. D. (1967). “A Simultaneous Axiomatization of Utility and Subjective Probability.” *Philosophy of Science*, 34: 333–40.
Google Scholar WorldCat

Broome, J. (1990). “Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism.” *Review of Economic Studies* 57: 477–502.
Google Scholar WorldCat

Broome, J. (1991). *Weighing Goods: Equality, Uncertainty and Time*. Oxford: Blackwell.
Google Scholar Google Preview WorldCat COPAC

Broome, J. (2004). *Weighing Lives.* New York: Oxford University Press.
Google Scholar Google Preview WorldCat COPAC

Diamond, P. A. (1967). “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility: Comment.” *Journal of Political Economy* 75: 765–66.
Google Scholar WorldCat

Epstein, L. G., and U. Segal (1992). “Quadratic Social Welfare Functions.” *Journal of Political Economy* 100: 691–712.
Google Scholar WorldCat

Fleurbaey, M., and A. Voorhoeve (2013). “Decide as You Would with Full Information! An Argument against *Ex Ante* Pareto.” In N. Eyal, S. Hurst, O. Norheim, and D. Wikler (eds.), *Inequalities in Health: Concepts, Measures and Ethics*. New York: Oxford University Press, 113–28.
Google Scholar Google Preview WorldCat COPAC

Gorman, W. M. (1968). “The Structure of Utility Functions.” *Review of Economic Studies* 35: 367–90.
Google Scholar WorldCat

Greaves, H. (Forthcoming) “*Antiprioritarianism.*”
Google Scholar Google Preview WorldCat COPAC

Harsanyi, J. C. (1955). “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility.” *Journal of Political Economy* 63: 309–21.
Google Scholar WorldCat

Jeffrey, R. C. (1965). *The Logic of Decision.* New York: McGraw-Hill.
Google Scholar Google Preview WorldCat COPAC

Jensen, K. K. (1995). “Measuring the Size of a Benefit and Its Moral Weight: On the Significance of John Broome’s Interpersonal Addition Theorem.” *Theoria* 61: 26–60.
Google Scholar WorldCat

Liapounoff, A. (1940). “Sur les fonctions-vecteurs complètement additives.” *Bulletin of the Academy of Sciences of the USSR*, Ser. Math. 4: 465–78.
Google Scholar WorldCat

Marschak, J. (1950). “Rational Behavior, Uncertain Prospects, and Measurable Utility.” *Econometrica* 18: 111–41.
Google Scholar WorldCat

McCarthy, D. (2006). “Utilitarianism and prioritarianism I.” *Economics and Philosophy* 22: 1–29.
Google Scholar WorldCat

Mongin, P. (1995). “Consistent Bayesian aggregation.” *Journal of Economic Theory* 66: 313–51.
Google Scholar WorldCat

p. 266 Otsuka, M., and A. Voorhoeve (2009). “Why It Matters That Some Are Worse Off Than Others: An Argument against the Priority View.” *Philosophy and Public Affairs* 37: 172–99.
Google Scholar WorldCat

Rabinowicz, W. (2002). “Prioritarianism for Prospects.” *Utilitas* 14: 2–21.
Google Scholar WorldCat

Rawls, J. (1971). *A Theory of Justice.* Cambridge: Harvard University Press.
Google Scholar Google Preview WorldCat COPAC

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.
Google Scholar Google Preview WorldCat COPAC

von Neumann, J., and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University press.
Google Scholar Google Preview WorldCat COPAC