THE JOURNAL OF PHILOSOPHY

VOLUME LXXXIV, NO. 8, AUGUST 1987

UTILITARIANISM AND EXPECTED UTILITY*

THE literature of economics contains a formal theorem that looks on the face of it like an argument for utilitarianism. It would be remarkable if a formal argument could establish a moral theory. This one does not; but I think it can contribute to our understanding of utilitarianism. This paper explores its contribution.

Granted some superficially plausible assumptions, the theorem shows that a utility function representing "social preferences" can be written as a sum of utility functions representing the preferences of individuals. As it stands, this theorem is uninformative, because the notion of social preferences is obscure. And I shall be showing that, when the notion is elucidated, the assumptions lose their plausibility. I shall therefore reinterpret the theorem in terms of *good* rather than *preferences,* in a way that makes the assumptions reasonably acceptable. The reinterpreted theorem shows that a utility function representing general good can be written as a sum of utility functions representing the good of individuals. This seems to suggest that good is the total of individual good, so there is no value in equality in the distribution of good. I shall be considering how far the theorem really licenses this utilitarian conclusion.

I. THE FORMAL THEOREM

Suppose there are *h* people. Each has preferences among a set of alternative prospects, the same set for everyone. Each person's preferences satisfy the axioms of expected-utility theory—I shall call such preferences *coherent.* Expected-utility theory tells us that coherent preferences can be *represented* by a utility function. This function assigns a utility to each prospect in such a way that, of any

* This is a revision of part of a paper I wrote while I was a Visiting Fellow at All Souls College, Oxford, supported by a grant from the Social Science Research Council. Some of the revision was done while I was a Visiting Fellow at the Australian National University. I thank these institutions, and also Martin Browning, Angus Deaton, Jan Graaf, James Mirrlees, Derek Parfit, and Amartya Sen.

0022-362X/87/8408/0405$01.80

two prospects, the preferred one has the higher utility. The function will also be *expectational,* by which I mean that the utility it assigns to a prospect whose results are uncertain is the mathematical expectation of the utility it assigns to the results. If a person's preferences are coherent, there are actually many expectational utility functions that will represent them, all positive linear transforms of each other.[1]

Suppose there are also social preferences among the same set of prospects. If these too are coherent, they can be represented by an expectational utility function. Once again there are actually many expectational utility functions that will represent them, all positive linear transforms of each other.

Suppose next that the social preferences are *Paretian,* that is:

> If everyone is indifferent between some pair of prospects, then the social preferences are indifferent too, and, if at least one person prefers the first of two prospects to the second and no one prefers the second to the first, then the first is socially preferred to the second.

The subject of this paper is:

> *Theorem 1.*[2] Assume that each person's preferences are coherent, and that social preferences are coherent and Paretian. Then there are expectational utility functions $U_1, \ldots U_h$ representing the individual preferences, and an expectational utility function U_g representing social preferences, such that for any prospect P
>
> $$U_g(P) = U_1(P) + \ldots + U_h(P)$$

A difficulty in understanding Theorem 1 is that it says nothing about *which* of the many utility functions that represent a person's

[1] In Richard Jeffrey's version of expected-utility theory, *The Logic of Decision,* 2nd ed. (Chicago: University Press, 1983), a wider range of transformations is allowed. Theorems 1 and 2 below are true within Jeffrey's theory, subject to some extra assumptions (see John Broome, "Bolker-Jeffrey decision theory reveals some cracks in an axiomatic buttress of utilitarianism," Discussion Paper 175, Economics Dept., Univ. of Bristol); but in this paper I confine my attention to more restrictive theories such as Leonard Savage's, see *The Foundations of Statistics,* 2nd ed. (New York: Dover, 1972).

[2] Among the published proofs of this theorem are: Robert Deschamps and Louis Gevers, "Separability, Risk-bearing and Social Welfare Judgments," in *Aggregation and Revelation of Preferences,* Jean-Jacques Laffont, ed. (Amsterdam: North-Holland, 1979), pp. 145–160; Peter C. Fishburn, "On Harsanyi's Utilitarian Cardinal Welfare Theorem," *Theory and Decision,* XVII, 1 (July 1984): 21–28; Peter J. Hammond, "Ex-ante and Ex-post Welfare Optimality under Uncertainty," *Economica,* XLVIII (August 1981): 235–250; Peter J. Hammond, "Ex-post Optimality as a Dynamically Consistent Objective for Collective Choice under Uncertainty," in *Social Choice and Welfare,* P. K. Pattanaik and M. Salles, eds. (Amsterdam: North-Holland, 1983), pp. 175–205; and John C. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy,* LXIII, 4 (August 1955): 309–321.

preferences is the one that is added to others to make up social utility. If U_1 is a utility function for person 1, then so is $2U_1$. But, if $2U_1$ were to replace U_1 in the formula above, the weight of person 1's preferences in social preferences would be increased. Social preferences would be more inclined to favor person 1 in the distribution of goods. So the implications of Theorem 1 for distribution are not clear cut.

One conclusion about distribution can be drawn nevertheless. The theorem links people's attitudes to uncertainty with the attitude of social preferences to inequality. The shape of a person's utility function indicates her attitude to uncertainty. But, when people's functions are added up to make social utility, as Theorem 1 says they can be, their shape also indicates the attitude of social preferences to inequality.

Suppose, for instance, that each person cares only about her own wealth and is risk averse about it. Then the utility function representing each person's preferences will be a strictly concave (downwards bending) function of her wealth. The social-utility function, by Theorem 1, will be a sum of such strictly concave functions. A function of this sort tends to favor equality in the distribution of wealth. The tendency may be weak, because the function may give more weight to some people's wealth than other's. But strict concavity amounts to "diminishing marginal utility of wealth," whose egalitarian tendency, though weak, is well known. Contrast it with the case where each person is risk neutral about her wealth. Then each person's utility function will be linear in her wealth. Social utility, by Theorem 1, will be linear in each person's wealth. And a linear function definitely attaches no value to equality in the distribution of wealth.

The linking of attitudes to uncertainty and attitudes to inequality is one remarkable consequence of Theorem 1. And the theorem has also been taken as an argument for utilitarianism.[3] But, as I say, I think it needs to be reinterpreted if its assumptions are to be acceptable. The reinterpreted theorem also makes a link between uncertainty and inequality, which I shall assess. And I also assess how far it supports utilitarianism.

Section II of this paper explains why a reinterpretation is needed. Section III describes the reinterpretation. Sections IV and V assess the theorem's assumptions as reinterpreted. Sections VI and VII consider what conclusions can be drawn.

[3] See John Harsanyi in "Morality and the Theory of Rational Behaviour," *Social Research,* XLIV, 4 (Winter 1977), reprinted in Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (New York: Cambridge, 1982), pp. 39–62.

One preliminary point needs to be made. Many of the proofs of Theorem 1 allow for subjective probabilities. They use versions of expected-utility theory in which the probabilities a person attaches to different events are revealed by her preferences. Different people's preferences may reveal different probabilities attached to the same event. The proofs of Theorem 1, however, also prove[4]

> *Theorem 2.* Assume that each person's preferences are coherent, and that social preferences are coherent and Paretian. Then the social preferences and all the individual preferences must agree about the probabilities of every event.

Since it is unlikely that everyone will agree about probabilities, this theorem tells us that, as a general rule, coherent Paretian social preferences cannot exist. But, if they do, then Theorem 1 tells us that they can be represented by a utility function that is the sum of individual-utility functions.

II. THE LIBERAL ARGUMENT

Are the assumptions of Theorem 1 plausible?

To test the Paretian assumption, consider this example. There are only two people. Both are concerned only about their own wealth, and they are risk neutral about it. One of two possible events will occur. Compare these prospects

	Event 1	Event 2
Prospect *A*	(2,2)	(2,2)
Prospect *B*	(3,0)	(0,3)

The parentheses show what wealth will result for the two people from each prospect if a particular event occurs. The first person attaches probability .7 to event 1 and .3 to event 2; the second person .3 and .7. Consequently, both people prefer prospect *B* to *A*. The Paretian assumption says that for this reason *B* is socially preferred to *A*. Can this be justified?

This depends on what, exactly, is meant by 'socially preferred'. Social preference is an ambiguous notion. "*B* is socially preferred to *A*" may mean simply that *B* is better than *A*. Under this interpretation, the Paretian assumption might be defended on the grounds that *B* is better than *A* because it is better for both people. This type of argument is explored in Section III. But it is not really available here because, although both people prefer *B* to *A*, *B* cannot actually be better for both of them. It can only be better for person 1 if event 1 is much more probable than event 2. And then it is not better for person 2.

[4] See the proofs by Deschamps and Gevers and Hammond mentioned in note 2.

Alternatively, "*B* is socially preferred to *A*" may mean that, if society had a choice between *A* and *B*, it should choose *B*. I take this to mean that the social system—perhaps the economic or political system, perhaps the government—should bring about *B* rather than *A*. There is no necessary implication that *B* is better than *A*.

It can plausibly be argued that, since both people prefer *B* to *A*, *B* is socially preferred to *A* in this second sense: it should come about rather than *A*. The argument is that the role of the social system is to act as a mechanism whereby people's wants get put into effect. So far as possible, it should simply bring about what people want. Normally, people's wants conflict, and then it should act as a fair mechanism for resolving the conflict. But, when there is no conflict, it should simply do what everyone wants. This argument does not insist that *B* is better than *A*. What it values is the nature of the social system, not its outcome. It says that the social system should be such that *B* comes about rather than *A*. But the occurrence of *B* is not necessarily any benefit added to the existence of the system that brings it about. Someone who accepts this argument might also believe that *B* is actually worse than *A*.

I call this *the liberal argument*. It supports the Paretian assumption under this interpretation of social preference, and I think it is the only argument that can do so. But someone who used this argument could not then defend Theorem 1's other main assumption that social preferences are coherent. The axioms of expected-utility theory are appropriate for an agent that makes choices, and derives reasons for making them from the value it attaches to their results. But the liberal argument does not conceive social preferences as representing the choices of any agent at all.[5] The social system does not really choose; it is a mechanism through which a result emerges from the choices of many agents. Social preferences represent the results that should emerge from this mechanism. And the reasons why the results should be one thing rather than another are not derived from the value of the results themselves, but from the merits of the mechanism.

Furthermore, anyone who supports the Paretian assumption has to tolerate incoherent social preferences. Theorem 2 says that social preferences, if they are to be Paretian, will be incoherent whenever people disagree about probabilities. Indeed, preferring *B* to *A* is

[5] A complication is that the working of the social mechanism often depends on the decisions of an agent. For instance, a minister in the government may actually decide between *A* and *B* in the example. And rationality requires an agent to have consistent preferences. This complication is discussed in my "Should social preferences be consistent?" Discussion Paper 180, Economics Dept., Univ. of Bristol.

already very near to incoherence. Any decent mechanism for resolving conflicting preferences would bring about (2,2) for sure in preference to either (3,0) for sure or (0,3) for sure, because (2,2) has more wealth in total and has it equally distributed. But it then violates the axioms of expected-utility theory[6] for it to bring about *B*, which is a prospect of either (3,0) or (0,3), in preference to *A*, which is (2,2) for sure. Theorem 2 is a snare in the path of any argument designed to justify the assumptions of Theorem 1. It says that the coherence and Paretian assumptions for social preferences are inconsistent with each other, unless people's probabilities coincide. So any argument that supports the Paretian assumption, even when people's probabilities differ, has to give up the coherence assumption as a general rule. It will then be hard for it to explain why coherence should be required in the one special case when people's probabilities happen to agree.

The liberal argument, then, cannot provide a foundation for both the Paretian and coherence assumptions of Theorem 1.

III. THE REINTERPRETATION

Now I return to the other type of argument that might be used to support the Paretian assumption. It depends on this *principle of personal good,* as I call it:

> If two prospects are equally good for everyone, they are equally good, and, if the first of two prospects is better for someone than the second, and at least as good for everyone, then it is better.

In brief: the goodness of a prospect depends only—and positively —on how good it is for people.

This principle does not actually support precisely the Paretian assumption. We have already seen that it does not support it in the example of prospects *A* and *B*. To bring it to bear on Theorem 1 we shall have to reinterpret the theorem. We shall have to replace preference relations with relations of betterness. Instead of "Person *i* prefers *P* to *Q* or is indifferent between them" we must have "*P* is at least as good for person *i* as *Q*." Instead of "*P* is socially preferred or indifferent to *Q*" we must have "*P* is at least as good as *Q*."

These betterness relations ("good relations") may satisfy the axioms of expected-utility theory. If they do—if they are coherent —they can be represented by expectational utility functions. For instance, there will be a utility function U_i such that $U_i(P)$ is greater than $U_i(Q)$ if and only if *P* is better than *Q* for person *i*.

[6] Specifically Savage's sure-thing principle (21–23).

The reinterpreted theorem[7] is:

Theorem 1′. Assume that each person's good relation is coherent, and that the general good relation is coherent and satisfies the principle of personal good. Then there are expectational utility functions $U_1, \ldots U_h$ representing individual good, and an expectational utility function U_g representing general good, such that for any prospect P

$$U_g(P) = U_1(P) + \ldots + U_h(P)$$

This theorem has more acceptable assumptions than the original does. In the rest of this section I explain the reinterpretation in more detail. Then in Sections IV and V I assess the reinterpreted assumptions.

The reinterpretation replaces social preference with general good. This is more a disambiguation than a reinterpretation. As I said earlier, 'is socially preferred to' can in one sense be taken to mean "is better than."

The reinterpretation replaces individual preference with individual good. This *is* a significant change. I am making it chiefly for the sake of generality. The argument we are discussing aims to say something about how the good of individuals is aggregated together to constitute general good. For that, there is no need to adopt any particular theory about what the good of individuals consists in. One theory is that it consists in the satisfaction of their preferences. This theory is, in a way, built into Theorem 1. But Theorem 1′ can accommodate any adequate theory of individual good.

We have seen already, however, that the preference-satisfaction theory of good is not adequate. It cannot be quite right. In the example above, *B* is preferred to *A* by both people, but it cannot be better for both. So Theorem 1′ cannot accommodate precisely this theory. But it may be able to accommodate a modified theory in which a person's good consists in the satisfaction of what her preferences would be if they were "purified" by full information and sound deliberation.

In the example, each person's preference for B may be rational: she may have judged the probabilities rationally on the information she has. So 'better for' does not mean the same as 'rationally preferred by'. 'Better for' requires that the probabilities, as well as being rationally arrived at, are correct. The principle of personal good,

[7] Peter Hammond also reinterprets the theorem in a way that is, if I understand it, similar to mine; see "On Reconciling Arrow's Theory of Social Choice with Hansanyi's Fundamental Utilitarianism," in George Feiwel, ed., *Arrow and the Foundations of the Theory of Economic Policy* (London: MacMillan, 1987), pp. 179–222.

then, works with the correct probabilities, which are the same for everyone, rather than with the probabilities people happen to believe in. In this way it avoids the trap set by Theorem 2; the principle is consistent with assuming that the general-good relation is coherent.

That is an advantage of my reinterpretation. A disadvantage is that it has to assume the existence of correct probabilities, which many subjectivists deny. I have been saying that the term 'better for' presupposes a notion of correctness for probabilities. So these subjectivists ought not to use this term. But the important thing is that they cannot avoid the trap of Theorem 2. Whatever argument they produce for the Paretian assumption or any similar principle, it will have to allow for disagreement about probabilities. And then social preferences will normally be incoherent. The subjectivists' argument will therefore have to tolerate incoherence. So it will not be able to offer a basis for Theorem 1 or any similar theorem.

These subjectivists, then, will not be impressed by a theorem of this sort. I therefore have to part company with them. I continue on the assumption that correct probabilities exist.

IV. THE PRINCIPLE OF PERSONAL GOOD

Now I come to assess the assumptions of Theorem 1, starting with the principle of personal good.

Consider first this principle as it applies to outcomes rather than prospects. It might be thought that some outcomes are simply good, or perhaps good for society, without being good for anybody. An example particularly germane to this paper is equality in the distribution of good between people. Two different ideas need to be looked at here.

One is the idea that, for the same total of personal good, it is better for it to be more equally distributed than less. This may suggest that equality is in a sense a good over and above the good of people. Utilitarianism denies this idea, and Theorem 1′ seems designed to support utilitarianism in this; section VII considers how far it really does so. Therefore, it had better not be simply an assumption of the theorem that the idea is wrong. And, fortunately, it is not, because the idea is consistent with the principle of personal good. Suppose, for instance, that general good is the product of individual goods—$G_g = G_1 G_2 \dots G_h$—and that all individual goods are positive. Then, for a given total of individual good, general good is increased by distributing it more equally. But, nevertheless, general good depends only—and positively—on individual good. And this is all the principle of personal good requires.

The second idea is that it is sometimes better to increase equality by taking good from people who have a lot, even if this does not benefit people who have less. This *does* conflict with the principle of personal good. And I think it is wrong. If such a change is really an improvement, that can only be because it improves fairness. The less well off are more fairly treated. But then this fairer treatment is a sort of good that is done them. Either there is such a good, in which case the change benefits the less well off after all, or there is not, in which case the change is not an improvement.

To speak generally, if any ostensible improvement cannot be pinned down as good for someone, I am not inclined to believe it is really an improvement. And from now on I am simply going to take the principle of personal good for granted when applied to outcomes: for outcomes, good depends only on people's good.

But there is a special difficulty about applying it to prospects. The sense in which the notion of good applies to prospects is a derivative and attenuated one. What is really good is not the prospect but the event it is the prospect of. Suppose some good event is going to happen to me. The prospect of this event is a good prospect. But when the good in my life is catalogued the prospect will not be recorded as an extra item besides the event. Certainly, the good prospect might *cause* some good that must be recorded: enjoyable anticipation, for instance. But it is not because of such effects that the prospect is good; it would still be good even if it did not cause any. Suppose now that some good event is likely but not certain to happen to me. That is a good prospect for me. But, if unluckily the event does not happen, then my life has not been made better by the fact that I once had this good prospect. The fact that something good might have happened is not itself something good that did happen.

Applied to this attenuated sort of good, the principle of personal good encounters a difficulty. Take this example (again, the parentheses show two people's wealth):

	Event 1	Event 2
Prospect *C*	(1,1)	(1,1)
Prospect *D*	(0,2)	(2,0)

Suppose that the events are equally probable. Suppose *C* is better than *D* for both people; this will be true if it is good for both of them to avoid risk. (This may be obscure; I shall come back to it.) Then the principle of personal good says that *C* is better than *D* because it is better for both people. Consider an act of choosing *C* rather than *D*. The principle offers as a reason in favor of this act that its consequence will be better for both people. But actually the consequence will be better for one and worse for the other: one will have one unit of wealth when she would otherwise have had nought, but the other will have one unit when she would otherwise have had two. The principle of personal good is trying to insert between the act of choice and its consequence a sort of quasi consequence, the prospect. And it applies to such quasi-consequences ways of thinking that may be very plausible when applied to real consequences but cannot be taken for granted here. How can it be argued that choosing *C* rather than *D* has results that are better for both people when, plainly, it does not?[8]

To deal with this point, we have to consider in detail what reasons there really are for choosing *C* rather than *D*, or *D* rather than *C*. Generally, we have to consider what reasons should guide choices between uncertain prospects. Talk about the goodness of alternatives ultimately aims to supply reasons for choice. So when the talk about goodness is in doubt we must look directly at the reasons.

To do this properly we need to have in mind a particular agent. It may be that a reason for one agent is not a reason for another, or that different agents should give different weights to the same reasons. So let us fix on an impartial agent with a general duty to act rightly but no special duty to any particular person. Perhaps the government might suit this role.

What reasons might such an agent have for choosing one alternative rather than another? There are, first of all, reasons of the sort that underlie the liberal argument described in section II. Suppose everyone prefers one of two alternatives to the other. Democratic principles suggest that this is the one that should come about, whether or not it is better for everybody, or for anybody. As I said, this is an argument about how the social system should operate as a mechanism. It is not necessarily a reason for any particular agent to bring about this alternative. But our agent may be, like the government, part of the social system. Its decisions may indeed determine what the social system brings about. In that case, these democratic considerations will constitute a reason for the agent to act in a particular way.

But we are not now concerned with reasons of this sort. We are concerned with good, and they are specifically not directed at good. I explained in section II that reasons like these are likely to lead to incoherent preferences, and that is why I turned to consider good instead. So from now on, I shall leave them aside and consider only good-directed reasons.

[8] I asked this, too forcefully, as a rhetorical question in my "Trying to Value a Life," *Journal of Public Economics*, IX (1978): 91–100. There is actually a good answer to it, which I am about to give. But most of that paper is, I still think, correct.

By a "good-directed" reason I mean a reason that is directed toward good in the outcome of a choice. The notion of a good-directed reason, then, does not depend on the notion of good applied to prospects. So I can use it to define good applied to prospects. When I say that one prospect is better than another, I mean that an impartial agent has stronger good-directed reasons for bringing about the first (if it can) than the second. And when I say that one prospect is better for a person than another, I mean that there are stronger reasons directed toward this person's good for bringing about the first than the second.

This way of defining good for prospects meets one essential requirement. Suppose we compare two prospects that each lead for sure to a particular outcome. The better prospect, as I have defined it, is the one that leads to the better outcome. Theorem 1′, which is formally about the goodness of prospects, will therefore tell us something about the goodness of outcomes too.

With these definitions, it is easy to produce an argument for the principle of personal good applied to prospects. Suppose two prospects are equally good for everybody. This means that for each person the reasons directed toward her good for bringing about one are exactly as strong as the reasons directed toward her good for bringing about the other. The impartial agent should act rightly, and beneficence is at least a part of acting rightly. So these will all be reasons for the agent too. And, since the reasons directed toward each person's good are evenly balanced, the agent's reasons directed toward everybody's good taken together will be evenly balanced too. As I say, I am taking it for granted that, for outcomes, good depends only on people's good. So these are all the good-directed reasons there are. Therefore, all the agent's good-directed reasons are evenly balanced. That is to say, the two alternatives are equally good. This proves the first part of the principle of personal good. The second part can be proved in a similar way.

I think this argument comes to the right conclusion. But it skips over what is really the main point. We need to consider the weighting of reasons in more detail.

Take the example of prospects *C* and *D*. Person 1's reasons in favor of *C* boil down to this: if event 1 comes about, she will get one unit of wealth instead of nought. Her reasons in favor of *D* boil down to: if event 2 comes about, she will get two units of wealth instead of one. We assumed that *C* is actually better for her. The former reason, that is to say, outweighs the latter. This is because we assumed that it is good for her to avoid risk. The value of risk avoidance appears in expected-utility theory in the form of weights attached to gains and losses of wealth. Formally it appears in a strictly concave utility function. Here it says that the difference between one unit and nought units has more weight than the difference between two units and one unit. Consequently, the reason in favor of *C* outweighs the reason in favor of *D*.

Our impartial agent has four good-directed reasons to weigh up. Two of them, directed toward person 1's good, are the ones I have just mentioned. And there are two symmetrical ones directed toward person 2's good. How should these reasons be weighed up in determining what the agent ought to do? We do not have to worry here about weighing against each other reasons directed toward different people's good. But I do want to argue that the two reasons directed toward person 1's good must be given the same relative weight as they receive when determining what is best for person 1.

The relative weight given to these reasons in the latter application is the means by which expected-utility theory takes account of whether or not it is good for person 1 to avoid risk, and the degree to which it is good. In our example we are assuming that risk avoidance is indeed good for person 1. So, in determining which prospect is better for her, the reason in favor of *C* outweighs the reason in favor of *D*. If, in determining which alternative our agent ought to choose, we did not give these reasons the same relative weight, the fact that avoiding risk is good for person 1 would not be properly taken account of in our determination. In the example, one reason for the agent to choose prospect *C* rather than *D* is that *C* is less risky for person 1, and avoiding risk is good for person 1. In expected-utility theory this reason appears, not strictly as a separate reason on its own, but in the weighting of other reasons. So this weighting must be preserved.

This is the point of treating prospects as "quasi consequences" and applying the notion of good to them. There is a type of reason for making a choice, namely avoiding risk,[9] which does not appear in the value of true consequences. But it can be represented in the value of prospects.

V. COHERENCE

Theorem 1′ requires individual and general good to satisfy the axioms of expected-utility theory. I mention here only the ones that are likely to give trouble: the completeness axiom and the various consistency axioms. The completeness axiom says that, of two prospects, either one is better than the other or else they are equally

[9] If, alternatively, it is good for a person to take on risk or to be neutral about risk, that too will be a reason for choice.

good. I do not detail the consistency axioms, except to mention that the most controversial is the axiom of strong independence.

I see no reason to think that either the individual- or the general-good relations are complete. For an individual, it is plausible that there are goods that cannot be weighed against each other. For general good, there is the added difficulty that it may not be possible to weigh one person's good against another's. For the sake of Theorem 1′ we shall simply have to assume there are no such intrapersonal or interpersonal incommensurabilities. This is a weakness of the theorem.[10]

On the other hand, I believe there are good arguments to show that the individual- and general-good relations satisfy the consistency axioms. I have argued in another paper[11] that the preferences of a rational agent will satisfy these axioms. It follows that each individual-good relation will satisfy them, because a person's good relation is what her preference relation would be were she rational and self-interested, and were her probabilities correct.[12] Similarly, the general-good relation would be the preference relation of a rational impartial agent whose preferences were determined by the weighing up of good-directed reasons in the manner described in section IV. So this relation, too, must be consistent.[13]

VI. UNCERTAINTY AND INEQUALITY

I have done what I can to justify the assumptions of the reinterpreted Theorem 1′. One source of doubt about them is that they presuppose a notion of correctness for probabilities. Another is that they ignore any problems there might be about the commensurability of goods. But apart from these doubts, I think the assumptions are acceptable.

[10] Richard Jeffrey, in "On interpersonal utility theory," this JOURNAL, LXVII, 20 (October 21, 1971): 647–656, uses Theorem 1 to support the possibility of interpersonal comparisons. I think this is wrong.

[11] John Broome, "Rationality and the Sure-thing Principle," in *Rationality, Self-Interest and Benevolence*, Gay Meeks, ed. (New York: Cambridge, forthcoming).

[12] This does not mean that a person's good consists in the satisfaction of her preferences, a theory about the nature of good mentioned in section III. The determination is in the opposite direction.

[13] The argument of this paragraph is slightly too quick. There might be no such thing as rational self-interested preferences, because it might not be rational to be self-interested. Similarly, it might not be rational for an impartial agent to determine its preferences by the weighing up of good-directed reasons, because, as I explained in section IV, the liberal argument of section II may also supply reasons to such an agent, and these are not good directed. Nevertheless, if a person's preferences *were* determined by her own interests only, and if an impartial agent's preferences *were* determined by good-directed reasons only, then these preferences should satisfy the consistency axioms. It is easy to check that my arguments in "Rationality and the Sure-thing Principle" would apply to such preferences.

So, subject to these qualifications, the conclusion of Theorem 1′ follows. Individual and general good can be represented by utility functions in such a way that general utility is the sum of the individual utilities. What does this tell us? In section I, I distinguished two conclusions one might draw from Theorem 1. First, the theorem made a link between people's attitudes to risk and the attitude of social preferences to inequality. Second, the theorem seemed to be an argument for utilitarianism. We must make a similar distinction for the reinterpreted Theorem 1′.

The first conclusion to be drawn from Theorem 1′ is that the value of equality is linked with the attitude people should take to uncertainty. If people should be neutral about risks to their wealth, then it is wrong to favor equality in the distribution of wealth. And, if people should be risk averse about wealth, there is some presumption in favor of equality in wealth. This link is explained in section I. The only difference is that now we are dealing with what is good for people faced with uncertainty, rather than with the attitude they actually take. Notice, though, that what is good for people here need not be determined by anything other than the people's own tastes. A person's tastes help to determine what is good for her.[14] If she likes apples more than pears then, other things being equal, an apple is better for her than a pear. Shifting attention from preferences to good does not deny the importance of tastes.

That inequality should be linked to uncertainty in this way strikes me as a remarkable consequence of Theorem 1′.

VII. UTILITARIANISM

Does Theorem 1′ give any more general support to utilitarianism?

At most it can support only a part of it. It has nothing to say about the utilitarian thesis that one should act so as to bring about the best result; it is only about what result is best. It has nothing to say about the utilitarian thesis that a person's good consists in pleasure or the satisfaction of her desires; it is only about how different people's good is aggregated. And it simply assumes the utilitarian thesis that goods are always commensurable.

The one part of utilitarianism that the theorem does seem to support is the thesis that general good is the sum of people's good, so there is no value in an equal distribution of good. But so far we are not entitled to suppose it says even this. It says that general utility is

[14] I distinguish tastes from preferences. I think that, other things being equal, satisfying a person's tastes is necessarily good for her. But I do not think the same about her preferences. A taste can supply a reason for a preference, but a preference can also be based on other reasons, or on no reason.

the sum of individual utilities. But utilities have been defined only to represent the *order* of good: of two prospects the one with the higher utility is the better. They do not necessarily represent *degrees* of good.

A case can be made out nevertheless for saying that utilities do actually represent degrees of good. Theorem 1′ itself can contribute to this case. But first let us see what argument can be made independently of Theorem 1′.

If a utility function represents the order of a person's good, then so does any linear transform of it. So the most that can be expected of a person's utility function in general is that it should be a linear transform of her good. The characteristic of a linear transformation is that it preserves the order of differences. Take four prospects M, N, P, and Q, each good for a person to some degree. Write these degrees $\mathrm{G}(M)$, $\mathrm{G}(N)$, $\mathrm{G}(P)$, and $\mathrm{G}(Q)$. Then the most that can be expected in general from a utility function for the person is that the utility difference $[\mathrm{U}(M) - \mathrm{U}(N)]$ should be greater than $[\mathrm{U}(P) - \mathrm{U}(Q)]$ if and only if $[\mathrm{G}(M) - \mathrm{G}(N)]$ is greater than $[\mathrm{G}(P) - \mathrm{G}(Q)]$.

Do utilities represent degrees of good to this extent? Suppose a person is faced with a choice between getting one unit of wealth for sure or alternatively taking a gamble at equal odds of either no units or two units. Suppose that

$$\mathrm{U}(2) - \mathrm{U}(1) < \mathrm{U}(1) - \mathrm{U}(0)$$

The latter difference in utility outweighs the former, and the best choice is the one unit for sure. Is it then necessarily the case that

$$\mathrm{G}(2) - \mathrm{G}(1) < \mathrm{G}(1) - \mathrm{G}(0) \qquad ?$$

This will be so if it is best for the person to maximize the expectation of her good.[15] Since utility is defined so that it is best for her to maximize the expectation of her utility, utility will then represent degrees of good to the required extent. But it is quite plausible that it is not best for a person to maximize the expectation of her good. For instance, the notion of degrees of good may not even make sense; good may not be an arithmetic quantity. Or it might be good for a person to be risk averse about good.[16] Then her utility will be a

[15] Many decision theorists seem to have thought people should be expected-good maximizers. Daniel Bernoulli, "Exposition of a New Theory on the Measurement of Risk," Louise Sommer, trans. *Econometrica*, XXII, 1 (January 1954): 23–36, thought people should maximize expected *emolumentum*, which the dictionary translates as "benefit or advantage." Jeffrey, *The Logic of Decision*, thinks they should maximize expected "desirability."

[16] Such a view is implicit in Kenneth J. Arrow, *Social Choice and Individual Values*, 2nd ed. (New Haven, Conn.: Yale, 1963), p. 10.

strictly concave function of her good, so we might consistently suppose in the example that actually, say,

$$G(2) - G(1) = G(1) - G(0)$$

But there is a possible retort to this. It might be said that it is precisely in comparisons of the sort we are making that the notion of degrees of good gets its meaning.[17] In making decisions in the face of uncertainty, gains and losses are weighed against each other. In the example, the possible gain in wealth from one unit to two is a reason in favor of taking the gamble. The possible loss from one unit to none is a reason against. The latter is the stronger reason. This is naturally expressed by saying that the difference between one unit and no units of wealth amounts to a greater difference in good than the difference between one unit and two. According to the supposition in the previous paragraph, these differences in good are actually the same. But what can this mean, if it is not that they are evenly balanced when weighed against each other as reasons for choosing? According to the supposition, the differences are equal but they do not *count* equally as reasons. But maintaining a distinction between amounts of good and how these amounts count looks like an empty gesture.

I think this is a good retort. It is hard to see what use we can have for the notion of degrees of good except when weighing up differences in good as reasons for making a choice. So it is in weighing up differences that we can expect the notion to get its meaning. Decision making under uncertainty, however, is not the only context in which differences in good are weighed against each other. Perhaps the notion gets its meaning elsewhere. Another context where differences of good are weighed is in the distribution of good between people. Let us consider that.

Suppose there is a choice between the distributions of wealth (1,1) and (0,2). A reason in favor of the first is that it gives person 1 one unit of wealth instead of none. A reason in favor of the second is that it gives person 2 two units instead of one. How should these reasons be weighed against each other? This is what Theorem 1′ is about. (Remember that it is simply an assumption of the theorem that reasons like these *can* be weighed against each other; interpersonal comparisons are possible.) It says that we can find the right weights from the people's utility functions. To do so, we have to make sure that for each person we have picked the appropriate utility function.

[17] Compare J. A. Mirrlees, "The Economic Uses of Utilitarianism," in Sen and Williams, pp. 63–84.

Each person has many functions representing the order of her good, and the theorem gives no guidance about which is the right one. But it does say that there is a right one. And once we have it, differences in utility determine the weights that should be given to opposing reasons. In the example, once we have functions U_1 and U_2 for the people, we compare $[U_1(1) - U_1(0)]$ with $[U_2(2) - U_2(1)]$. If the former is greater, the distribution (1,1) is better; if the latter, (0,2). This is a context where we are weighing up reasons. So according to what I said above, it gives us grounds for saying that these differences of utility represent differences in degrees of good. Suppose $[U_1(1) - U_1(0)]$ is greater than $[U_2(2) - U_2(1)]$. Then we have grounds for saying that person 1 gains more in good from having one unit of wealth instead of none than person 2 gains from having two units instead of one. Since the same utility functions supply the right weights in any distributional comparison, we have grounds for saying that these functions represent degrees of good.

But these grounds are unlikely to convince a nonutilitarian. They beg the question. They insist that, when weighing reasons, the stronger reason must always be the one that represents the greater difference in good. This simply assumes that the better alternative is always the one with the greater total of good. And that was what had to be proved.

The strength of the utilitarian case, however, is this. The functions U_1 and U_2, which supply the weights when weighing reasons in distributing wealth, are utility functions for the people. Therefore, they also supply the weights when weighing reasons in making decisions under uncertainty. So these functions serve the same purpose in two contexts. This very much strengthens the claim that they represent degrees of good. This is the effect of Theorem 1′. Theorem 1′ provides a strong case for saying that utilities represent degrees of good.[18] And, having done so, it also says that the better of two alternatives is always the one with the greater total of good.

The answer to the nonutilitarian's objection is this. The objection relies on a distinction between degrees of good and how these degrees count in weighing reasons: utility tells us how good counts, but utility may be distinct from good itself. But we have been shown no way of assigning meaning to degrees of good apart from how they count. And, without that, the distinction now seems emptier than ever.

[18] Compare John C. Harsanyi, "Nonlinear Social Welfare Functions: A Rejoinder to Professor Sen," in *Foundational Problems in the Social Sciences,* R. Butts and J. Hintikka, eds. (Boston: Reidel, 1977), pp. 293–296.

But this is not the end of the argument. All the nonutilitarian has to do now is supply a suitable way of assigning meaning to degrees of good. What she needs is *another* context in which differences of good are weighed against each other. The one to turn to, I think, is the weighing up of good at different periods of a person's life. I intend to pursue this in another paper.

Imagine for a moment, though, that the argument had come to a decisive end in the defeat of the nonutilitarian. Imagine we had managed to derive from Theorem 1′ the utilitarian conclusion that good is the sum of people's good. What would this conclusion now amount to? The way we would have come to it shows it is less significant than it seems. It seems anti-egalitarian. But our argument was simply about meaning. In order to make sense of the question whether or not there is value in an equal distribution of good, we have to assign a meaning to the notion of degrees of good. And it happens that the most natural way of doing that prevents us from attaching value to equality in the distribution of good. That is all. It suggests that the question whether there is value in equality in the distribution of good is unimportant.

Furthermore, the argument also shows that the utilitarian conclusion adds nothing at all to the conclusion we reached in section VIII. There we made a connection between uncertainty and inequality. The same utility functions that represent what is best in the face of uncertainty also represent what is best in distribution between people. And it is simply because the same functions appear in both contexts that we have now decided they represent degrees of good. Theorem 1′ gives no more general support to utilitarianism than that.

The argument, to summarize, is not ended. But, if it were, it would have achieved less than might have been expected of it.

JOHN BROOME

University of Bristol