UTILITARIAN METAPHYSICS?

John Broome

Published in *Interpersonal Comparison of Well-Being*, edited by Jon Elster and John Roemer, Cambridge University Press, 1991, pp. 70–97.

This paper outlines the core argument of my book *Weighing Goods*, Blackwell, 2001.

The argument of this paper was first developed in an unpublished paper I wrote while I was a Visiting Fellow at All Souls College, Oxford, financed by a grant from the Economic and Social Research Council. I thank the College for its hospitality, and the ESRC for its support.

Section 1

I am going to make some comparisons between interpersonal comparisons of good and intrapersonal intertemporal comparisons of good. I shall set them in the context of formal theorems that connect together these different dimensions of comparison. I hope to shed some light on the foundations of utilitarianism. Not all of utilitarianism, though; only this one part:

> The Utilitarian Principle of Distribution. One alternative is at least as good as another if and only if it has at least as great a total of people's good.

Briefly: good is the total of people's good. Utilitarians value only the total of good, regardless of whom it comes to. They do not value equality in the distribution of good.

It has often been suggested[1] that utilitarianism is associated with a particular metaphysical theory of personhood: a "disuniting" theory that denies the full unity of a person over time. But I do not think the connection between utilitarianism and a disuniting metaphysics has yet been properly made out.[2] I hope, with the help of the formal theorems, to show how one might derive the Utilitarian Principle of Distribution from a disuniting metaphysics. I shall sketch out a line of argument, though it is far from complete.

I am not myself committed to either the Utilitarian Principle or a disuniting metaphysics. Deriving one from the other can serve either to support the one or to weaken the other.

Throughout this paper, I shall speak of good rather than wellbeing or preference-satisfaction or something else. This is chiefly for the sake of generality. I want to be neutral between competing theories of what a person's good consists in. I want to talk about how goods coming at different times or to different people are compared, irrespective of what these goods might actually be. Amongst the competing theories of good is the theory that a person's good consists in the satisfaction of her preferences. In speaking of good rather than preference-satisfaction, I am not denying this theory, but only allowing for other possibilities as well. In speaking of good rather than wellbeing, I am allowing for the

possibility of goods that cannot plausibly be counted as part of wellbeing.[3]

I shall be concerned with how goods coming at different times or to different people go together to make up overall good. This is a matter of assessing, not just the relative magnitudes of the goods, but also their relative values or how much they count in the overall judgment. These may or may not be different things, and I do not want to preempt that issue at the beginning. Comparing goods suggests to me only the first task of assessing relative magnitudes. I want to include the second task too. So from now on I shall use the more general metaphor of weighing goods.

Sections 2 and 3 of this paper describe the theorems and consider what conclusion can be drawn from them. Section 2 is largely a summary of a previous article of mine, "Utilitarianism and expected utility",[4] which deals with the interpersonal weighing of goods. Section 3 extends the argument to include intertemporal weighing too. The crucial issue turns out to be the truth of something I call "the Principle of Temporal Good." If true, this principle will give strong support, through the theorems, to the Utilitarian Principle of Distribution. Section 4 suggests that the Principle of Temporal Good is prima facie doubtful. But Section 5 argues that, nevertheless, it may be possible to defend it on the basis of a disuniting metaphysical theory. This is how I think a disuniting theory can support utilitarianism. Section 6 compares this argument with Derek Parfit's argument for the same conclusion. It argues that his is unsuccessful.

Section 2

Suppose there are h people. Suppose we are interested in prospects drawn from some set of uncertain prospects - the same set for everybody. Take the j'th person and consider her "betterness relation":

_ is at least as good as _ for j,

where each blank is to be filled in with a prospect from the given set.

This relation will have many of the same properties as a rational preference relation.

In fact, I argued in "Utilitarianism and expected utility" that j's betterness relation will satisfy the axioms of expected utility theory (when it is inserted into the theory in the place of a preference relation). I expressed doubts about one of the axioms: completeness. But I argued that the major consistency axioms of the theory, including the controversial Sure-Thing Principle, apply to this relation. I call a relation "coherent" if it satisfies the axioms of expected utility theory. So my conclusion, setting aside the reservation about completeness, is that each person's betterness relation is coherent.

Since it is coherent, j's betterness relation can be represented by a utility function in the manner of expected utility theory. Write this function U_j. $U_j(A)$ is at least as great as $U_j(B)$ if and only if A is at least as good for j as B. The function is unique up to positive linear transformations. And U_j is "expectational," by which I mean that the utility of an uncertain prospect is the expectation of the utilities of its possible outcomes.

Consider next the general betterness relation

_ is at least as good as _,

defined on the same set of prospects. "Utilitarianism and expected utility" argued that this relation too will satisfy the axioms of expected utility theory, if we set aside doubts about completeness that arise in this context too.

These doubts are actually particularly serious for the general betterness relation. It used commonly to be claimed that interpersonal comparisons of good cannot be made. I take this to mean that one person's good cannot be weighed against another's: if, that is, A is better than B for one person, and B better than A for another, it is undetermined whether, on balance, A is better or worse than B. If this is so, then the general betterness relation is incomplete. Nevertheless, I am simply going to assume it is complete, and set aside this worry about interpersonal comparability. This is a gap in the argument. It needs to be pointed out, because some authors have mistakenly believed that the theorem discussed in this section (Theorem 1 below) actually overcomes the worry: it demonstrates that interpersonal comparisons can be made.[5] Comparability, however, is an assumption of

the theorem, not a conclusion.

Given that, then, the general betterness relation is coherent. So it can be represented by an expectational utility function U, unique up to positive linear transformations. U(A) is at least as great as U(B) if and only if A is at least as good as B.

In "Utilitarianism and expected utility" I also argued for this

> Principle of Personal Good. If two prospects are equally good for everybody, they are equally good. And if one prospect is at least as good as another for everybody and definitely better for somebody, it is better.

The following theorem can then be proved:[6]

> Theorem 1. Suppose the general betterness relation is coherent, and so is each person's. And suppose the Principle of Personal Good is true. Then the general betterness relation can be represented by a utility function U that is the sum of utility functions U_1 ... U_h representing the people's betterness relations:
>
> $$U(A) = U_1(A) + ... + U_h(A) \quad \text{for all prospects A.}$$

Remember that there are many utility functions representing a person's betterness relation, each a positive linear transform of the others. Theorem 1 says you can pick one function for each person, out of all her functions, in such a way that the total of all the functions you pick represents the general betterness relation.

Since the assumptions of this theorem can be supported by argument, its conclusion deserves some credence. And at first sight it looks like the Utilitarian Principle of distribution. It seems to say that good is the total of people's good.[7] But actually we are not - or not yet - entitled to draw this inference from the theorem. The theorem says that general utility is the total of individual utilities. And an individual's utility is defined to represent her betterness relation. But it is defined only to represent the order of betterness for her: of two prospects the one with the greater utility is better for her. Utility does not necessarily represent the quantity of her good. But unless it does, Theorem 1 says nothing about the total of people's good.

However, it can be argued that utilities do indeed represent quantities of good. If this can be established, then Theorem 1 will genuinely support the Utilitarian Principle. So let us pursue this argument. Theorem 1 itself can contribute to it. But first let us see what can be said independently of Theorem 1.

If a utility function represents a person's betterness relation, then so does any positive 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 positive linear transform of her good. The characteristic of a positive linear transformation is that it preserves the order of differences. Take any four prospects A, B, C and D, which give person j these quantities of good: $G_j(A)$, $G_j(B)$, $G_j(C)$ and $G_j(D)$. Then the most that can be expected in general from j's utility function U_j is that the utility difference $\{U_j(A)-U_j(B)\}$ should be at least as great as $\{U_j(C)-U_j(D)\}$ if and only if the difference in good $\{G_j(A)-G_j(B)\}$ is at least as great as $\{G_j(C)-G_j(D)\}$.

Do utilities represent quantities of good to this extent? Take an example. Suppose the person is faced with a choice between two uncertain prospects. One gives her equal chances of getting ten units of wealth or fourteen units; the other equal chances of five or twenty units. In a table, the alternatives are:

States of nature (equally likely)		States of nature (equally likely)	
H	T	H	T
10	14	5	20
Prospect A		Prospect B	

Table 1

Suppose, say, that the utility difference $\{U_j(10)-U_j(5)\}$ is at least as great as $\{U_j(20)-U_j(14)\}$. Does it follow that $\{G_j(10)-G_j(5)\}$ is at least as great as $\{G_j(20)-G_j(14)\}$?

The utilities are defined to represent the order of goodness. So they tell us which of A or B is the better prospect for the person. The expected utility of A, $\frac{1}{2}\{U_j(10)+U_j(14)\}$, is at least as great as the expected utility of B, $\frac{1}{2}\{U_j(5)+U_j(20)\}$. Therefore A is at least as

good for the person as B. The judgment between the two can be looked at this way. It is a matter of weighing against each other the possible loss from ten units to five units in state H against the possible gain from fourteen units to twenty units in state T. Since $\{U_j(10)-U_j(5)\}$ is at least as great as $\{U_j(20)-U_j(14)\}$, the loss counts at least as much as the gain.

So A is at least as good for the person as B. The question is, again, does it follow that $\{G_j(10)-G_j(5)\}$ is at least as great as $\{G_j(20)-G_j(14)\}$?

It would follow if it were necessarily good for a person to maximize the expectation of her good. Then A, being at least as good for the person as B, would necessarily have at least as great an expectation of good. That is, $\frac{1}{2}\{G_j(10)+G_j(14)\}$ would necessarily be at least as great as $\frac{1}{2}\{G_j(5)+G_j(20)\}$. Hence $\{G_j(10)-G_j(5)\}$ would necessarily be at least as great as $\{G_j(20)-G_j(14)\}$. But it is quite plausible that it is not necessarily good for a person to maximize the expectation of her good. For one thing, good may not even be an arithmetic quantity, so there may not even be such a thing as the expectation (in probability theory's sense) of the person's good. And secondly, even if good is an arithmetic quantity, it may be good for a person to be risk-averse about her good. A is less risky than B, and this may make it at least as good as B even if its expectation of good is less. It is perfectly possible, that is, even though A is at least as good as B, for its expectation of good $\frac{1}{2}\{G_j(10)+G_j(14)\}$ to be less than B's expectation of good $\frac{1}{2}\{G_j(5)+G_j(20)\}$. If so, then $\{G_j(10)-G_j(5)\}$ will be less than $\{G_j(20)-G_j(14)\}$. Utility will not then be a positive linear transform of good.

But there is a plausible retort to this argument. It might be said that it is precisely in comparisons of the sort we are making that the notion of quantities of good gets its meaning. In assessing alternatives in the face of uncertainty, possible gains and losses are weighed against each other. In the example, the possible loss in wealth from ten units to five is weighed against the possible gain from fourteen to twenty. We know the former counts at least as much as the latter in determining which of the alternatives is better. This

would be naturally expressed by saying that the difference between ten units of wealth and five amounts to at least as great a quantity of good as the difference between twenty and fourteen. In the last paragraph I said that the difference in good between ten units of wealth and five might actually be less than the difference in good between twenty and fourteen. But what can this mean, if it is not that it counts less in determining which alternative is better? I would have to say that the difference in good is less, but it counts for at least as much in determining the relative goodness of the alternatives. But it looks like an empty gesture to maintain a distinction between quantities of good and how much these quantities count.

I think this is a good retort. It is hard to see what use we can have for the notion of quantities of good except when weighing up differences in good in assessing alternatives.[8] So it is in weighing up differences that we can expect the notion to get its meaning. 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 is the distribution of good between people. Let us consider that.

Suppose we have to compare these two distributions of wealth for two people:

People		People	
1	2	1	2
10	14	5	20
Distribution A		Distribution B	

Table 2

In favor of A is that it gives the first person ten units of wealth instead of five. In favor of B is that it gives the second person twenty units instead of fourteen. How should these conflicting considerations be weighed against each other? That is what Theorem 1 is about. 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. Each person has many functions representing the order of her good, each an increasing

linear transformation of the others, and the theorem gives no guidance about which is the right one. But it does say there is a right one. (Remember it is an assumption of the theorem that the people's good can be compared in the first place.) And once we have it, differences in utility determine the weights that should be given to opposing considerations. In the example, once we have functions U_1 and U_2 for the people, we compare $\{U_1(10)-U_1(5)\}$ with $\{U_2(20)-U_2(14)\}$. If the former is greater, distribution A is better; if the latter, B. This is a context where differences of good are being weighed. So according to what I said above, it gives us grounds for saying that these differences of utility represent differences in quantities of good. Suppose $\{U_1(10)-U_1(5)\}$ is greater than $\{U_2(20)-U_2(14)\}$, so A is better. Then we have grounds for saying that person 1 gains more in good from having ten units of wealth instead of five than person 2 gains from having twenty units instead of fourteen. Since the same utility functions supply the right weights in any distributional comparison, we have grounds for saying that these functions represent quantities of good.

But these grounds are unlikely to convince a nonutilitarian. They beg the question. They insist that, when comparing the goodness of alternatives by weighing differences in good, the greater difference must always be the one that wins. The greater difference always counts more in the comparison. But this simply assumes that the better alternative is always the one with the greater total of good. And that is what has to be proved.

The strength of the utilitarian case, however, is this. The functions U_1 and U_2, which supply the weights when comparing the goodness of alternative distributions of wealth, are utility functions for the people. By their definition, therefore, they represent the order of goodness for the people of uncertain prospects. So they <u>also</u> supply the weights when comparing the goodness of alternative prospects in the face of uncertainty. These functions, then, serve the same purpose in two contexts. This very much strengthens the claim that they represent quantities of good. This is the effect of Theorem 1. Theorem 1 provides a strong case for saying that utilities represent quantities of good. 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 quantities of good and how these quantities count in determining the goodness of alternatives: utility tells us how good counts, but it may be distinct from good itself. We have, though, been shown no way of assigning meaning to quantities of good apart from how they count. And, without that, the distinction now seems emptier than ever.

But that is not the end of the argument. All the nonutilitarian needs to do is supply a way of assigning meaning to quantities of good. What she needs is a another context in which differences of good are weighed against each other. The one to turn to, I think, is the weighing up of goods that come at different times in a person's life. The good in a life may be distributed across time in different ways. It may be spread evenly or unevenly, or it may come early or late. In comparing the goodness of different possible lives, good at one time has to be weighed against good at another. Perhaps this is where quantities of good get their meaning. So now I turn to these intertemporal weighings.

Section 3

Suppose there are a number of "times," say T altogether.[9] For some person j and time t consider the relation

_ is at least as good as _ for j at t,

where each blank is to be filled in with a prospect from the same set as before. It is natural to be dubious about the existence of dated betterness relations like this. But we do often speak of how good things are for a person at one time or another. For instance, we might say that a vaccination is bad for you at the time, because it makes you sick for a while, but it is good for you in the long run. To be sure, we might be a little reluctant to say it is bad for you at the time. If we think it is actually good for you, that may make us reluctant to say it is bad for you at any time. But I think this reluctance is easily overcome. There is really no difficulty in recognizing that the vaccination may be bad for you at one time, even though the badness may be outweighed by a greater benefit later. So I think there is no real

difficulty in recognizing the existence of dated betterness relations.

Furthermore, the arguments that show the coherence of the individual and general betterness relations apply to dated betterness relations too.[10] There is, once again, a reservation to be made about completeness. But apart from that, each dated betterness relation is coherent.

The real worry over dated betterness relations is not, I think, whether they exist, but whether they capture all that is good or bad for a person. The claim that they do is expressed in this

> Principle of Temporal Good. If two prospects are equally good for a person at every time, they are equally good for her. And if one prospect is at least as good as another for the person at every time and definitely better for her at some time, it is better for her.

This principle is the main subject of this paper. I shall consider arguments for and against it. But first I want to explain its importance.

A simple reinterpretation of Theorem 1 gives us

> Theorem 2. Suppose a person's betterness relation is coherent, and so is each of her dated betterness relations. And suppose the Principle of Temporal Good is true. Then the person's betterness relation can be represented by a utility function U_j that is the sum of utility functions U_{j1} ... U_{jT} representing her dated betterness relations:
>
> $$U_j(A) = U_{j1}(A) + ... + U_{jT}(A) \quad \text{for all prospects A.}$$

Putting Theorems 1 and 2 together gives us

> Theorem 3. Suppose the general betterness relation is coherent, and so is each person's betterness relation, and so are all of each person's dated betterness relations. And suppose that both the Principle of Personal Good and the Principle of Temporal Good are true. Then the general betterness relation can be represented by a utility function U that is the sum of utility functions U_j

representing each person's betterness relation. And each of these in turn is the sum of utility functions U_{jt} representing the person's dated betterness relations:

$$U(A) = U_1(A) + ... + U_h(A)$$
$$= U_{11}(A) + ... + U_{1T}(A) + ... + U_{h1}(A) + ... + U_{hT}(A)$$

for all prospects A.

The various betterness relations are, I claimed, coherent, setting aside doubt about their completeness. And I claimed that the Principle of Personal Good is true. So the truth of the conclusions of these theorems turns on the Principle of Temporal Good.

Suppose the conclusions <u>are</u> true; what would that tell us? It would tell us that the same utility functions determine how differences in good weigh against each other in <u>three</u> different contexts: across time within a life, across people, and across states of nature when there is uncertainty. That would follow from a simple extension of the arguments in Section 2. And it would make it very hard to resist the conclusion that these functions represent quantities of good. I can think of no other context where the notion of quantities of good could get its meaning. Whenever there is weighing up to be done, according to the theorems, the total of utility always determines what is best. The same utilities always determine how much good counts in every context. In these circumstance, it would be impossible to maintain the distinction between good and how much it counts. Utility would represent quantities of good. And, given that, the best alternative would always be the one with the greatest total of good. This Utilitarian Principle of Distribution would be irresistible.

So if the Principle of Temporal Good is true, that will give very strong support to the Utilitarian Principle. That is the significance of the Principle of Temporal Good.

This conclusion, notice, is derived from considerations of <u>meaning</u> only. The Utilitarian Principle looks like a substantive and important principle that is opposed to egalitarianism. But the effect of this argument in its defence is to make it much less important. The argument concludes, in effect, that the notion of a quantity of good acquires

its meaning in such a way that it turns out to be best to maximize the total of people's good. I think there is a lesson to be learnt from this. The substantive and important questions about equality will be to do with equality in income or perhaps other things that come with a natural metric. Questions about equality in the distribution of good seem at first to be more fundamental. But in the end they boil down to questions of meaning. They only seem substantive if one assumes in advance a natural metric for good. I very much doubt if there is a such a metric apart from the ones, derived from comparisons in various contexts, that I have considered.

Section 4

The prima facie evidence is against the Principle of Temporal Good.[11] Consider this example. Imagine for simplicity that there are only two times, and compare this pair of prospects for a person:

		States of nature (equally likely)				States of nature (equally likely)	
		H	T			H	T
Times	1	x	y	Times	1	x	y
	2	x	y		2	y	x
		Prospect A				Prospect B	

Table 3

In this table "x" and "y" refer to the quantities of good that come to the person at the times.[12] Assume x and y are different. Prospects A and B are equally good for the person at both times, since at both times they each give her an equal chance of x and y units of good. The Principle of Temporal Good, then, implies that prospects A and B are equally good for her. But it is plausible that actually B is better for her. This is because B gives the person, for sure, x+y units of good altogether, whereas A gives her either 2x or 2y. A is risky, then. So if there is any value in avoiding risk to the person's good, B is better. Its

superior value appears from the standpoint of the person as a whole, and does not show up in either of the times taken separately.

Compare this formally similar example that has two people instead of two times:

		States of nature (equally likely)				States of nature (equally likely)	
		H	T			H	T
People	1	x	y	People	1	x	y
	2	x	y		2	y	x
		Prospect A				Prospect B	

Table 4

Prospects A and B are equally good for each person. So the Principle of Personal Good implies they are equally good. And in this case there is no analogous reason for doubting this conclusion. There is no standpoint analogous to the standpoint of the person as a whole. There is no plausible reason why it should be good to avoid risk to the total good of the two people taken together. The Principle of Personal Good is not threatened by the example in the same way as the Principle of Temporal Good.

On the face of it, however, there seems to be an opposite threat. Although there is nothing in the interpersonal example to recommend B over A, there may seem to be something to recommend A over B: A leads inevitably to an equal distribution of good, and B to an unequal one. If this really makes A better than B, then that contradicts the Principle of Personal Good.

But what, if anything, is good about equality and bad about inequality? It can only be that equality is fair and inequality unfair. Inequality is unfair to the people at the bottom. This is a harm done them as individuals. Equality is good because it avoids this individual harm.[13] In Table 4 the figures stand for the people's total good. Any harm of unfairness that may be caused by inequality has already been taken into account. Once this has been done the value of equality can then provide no further reason for favoring prospect A over B. So

this value turns out to be no threat to the Principle of Personal Good.

This pair of examples, then, shows a disanalogy between the Principle of Personal Good and the Principle of Temporal Good. The latter is on more shaky ground because prospects can be judged from a standpoint that links different times together: the standpoint of a person as a whole. There is no analogous standpoint that links different people together.

Section 5

The example of Table 3 constitutes a prima facie case against the Principle of Temporal Good. Now I shall consider how, nevertheless, this principle might be defended. I shall consider, in particular, whether any metaphysical argument might be made for it.[14]

Some metaphysical theories about personhood deny the full unity of a person over time, and instead assimilate the relationships between different times in a person's life to the relationships between different people. "We regard the rough subdivisions within lives as, in certain ways, like the divisions between lives," says Derek Parfit.[15] I shall call this vaguely defined class of theories "disuniting." It seems plausible that a disuniting theory might be able to give support to the Principle of Temporal Good. It might assimilate the Principle of Temporal Good to the Principle of Personal Good. Since the latter is plausible, this might support the former. Specifically, a disuniting metaphysics might deny the standpoint of the person as a whole.

On the face of it, the two principles are metaphysically not closely analogous. The Principle of Personal Good is about putting together the good of different things - people. The Principle of Temporal Good is about putting together the good of a single thing - a person - that comes to her at different times. At all times in her life, a person is the same person, one thing. This is what makes available the standpoint of the person as a whole. It is the source of reasonable doubts about the Principle of Temporal Good. Across people, there is nothing analogous to this fact that a person is the same person at all times. There

is nothing that is the same at all people.

But a disuniting metaphysical theory makes an analogy where, on the face of it, there is not one. It supposes a person is in some way made up of temporal segments. Each segment is a thing on its own. That a person is the same person at different times is more accurately expressed by saying that person-segments existing at different times make up a single person. This fact, then, is analogous to the fact that different people make up a single society. And this theory supposes that the good of a person at a time is the good of the person-segment that exists at that time. The Principle of Temporal Good, therefore, is about putting together the good of different things, just like the Principle of Personal Good. With this analogy in place, the reasons that support the latter principle may be able to support the former.

This, then, is how a metaphysical argument might go. The rest of this section lays out an argument along these lines. But it leaves some gaps.

The beginning of the task will be to establish that people are made up of temporal segments in the first place. In this paper, though, I am not concerned with the merits of the disuniting theory itself; only with its implications. So I shall take this much for granted.

The notion "made up of" is vague. Societies are made up of people and water is made up of molecules, but the relation between a society and its people is not very similar to the relation between water and its constituent molecules. Different disuniting theories are possible, each with its own account of the relation between a person's segments and the person. One view is that this relation is membership: a relation formalized in set theory. Another is that it is the relation of part to whole: a relation formalized in mereology.[16] All the theories, though, must share this implication: the properties of a person must supervene on the properties of her segments.[17] If a person might have been different in some way without any of her segments being different, then she could not be said to be made up of her segments.

The Principle of Temporal Good says the good of a person supervenes on her dated

goods: her good could not have been different without one of her dated goods being different. (It also says the direction of supervenience is positive: more good for the person implies more good for her at some time. But in this paper I shall leave aside this question of direction. This is one gap in the argument.) The disuniting metaphysics treats the person's good at a time as the good of one of her segments. So, granted the metaphysics, the Principle of Temporal Good says that the good of a person supervenes on the good of her segments.

The disuniting metaphysics implies, as I say, that the good of a person (being one of her properties) supervenes on the properties of her segments. But that it supervenes on the good of her segments is a big further step. Compare a different example. I am made up of my spatial parts. My properties supervene on the properties of my parts. But my good does not supervene on the good of my parts. It is even true that my parts have their own goods: exercise is good for my muscles and fish is said to be good for my brain. But things can be good for me without being good for any of my parts.

The theory that a person is made up of temporal segments, then, is not enough. Something needs to be added to it.

Another example shows the sort of thing it needs to be. A society is made up of people. And it is natural to think that the good of the society does supervene on the good of its members; nothing can be good for the society without being good for one of its members. This, at any rate, is an implication of the Principle of Personal Good. So what is the difference between the example of me and my parts and the example of a society and its members? It must be something to do with the way in which the parts and the members go to make up me and the society. And this must be something to do with the relations amongst them. It must be that my parts are related together in such a way that they make up an aggregate, me, whose good does not supervene on the good of the parts. On the other hand, this must not be true of the members of a society.

So we need to examine the relations between a person's temporal segments. Are they

such that the segments form an aggregate, the person, whose good does not supervene on the good of the segments? (I shall call such a good "autonomous".) Could one argue that they do not?

Here is a way. Suppose we add to the disuniting metaphysics this further premise: that the relations between a person's segments, in virtue of which the segments make up a person (I shall call these "the unifying relations"), are not significant in respect of good ("ethically significant," I shall say). I shall discuss later whether this premise is defensible. But first I shall show it is enough to give us the argument we are looking for. An outline of the argument is this. If these relations between segments are not significant in respect of good, then it makes no difference to good whether or not a collection of segments makes up a person. And this implies that, when a collection does happen to make up a person, this person cannot have an autonomous good of her own. What follows fills out this outline.

I need first to define more precisely the notion of ethical significance. Our premise is that the unifying relations are not ethically significant. What does this mean? Clearly it should be understood counterfactually. A first approach is: if the unifying relations between a person's segments did not hold, then the good and bad in the world would be just as it actually is. This formula, however, picks up causal factors that I do not mean to be picked up. If these relations did not hold, that is likely, for causal reasons, to make a difference to the good of individual segments. Suppose, for instance (as Parfit believes) that amongst the unifying relations are relations of memory. If some wizened person-segment remembers the achievements of some youthful person-segment, that is part of what makes these two segments components of the same person; it is one of the unifying relations. Now, if this relation did not hold, so the wizened segment did not have this memory, this segment would probably be less content than it actually is. There would therefore be less good in the world than there actually is. One could say, then, that this relation of memory is ethically significant in a causal way. But when I say that the unifying relations are not ethically

significant, I do not mean to deny their causal significance. I mean they have no ethical significance apart from a causal one.

What I mean by the premise is this: if the unifying relations between a person's segments did not hold, but the good and bad of each segment was just as it actually is, then the good and bad in the world would be just as it actually is. Granted this, we can argue as follows that a person can have no autonomous good.

Take a person, and imagine what it would be like if the unifying relations between her segments did not hold, but the good and bad of each segment remained the same. Imagine, say, that half way through her life, the person was magically swept out of existence, and in her place appeared a new person similar in all respects, who then lived out the rest of her life. The details of the magic required for this trick will depend on what the unifying relations are, and there are different theories about that. There might have to be an infinitesimal gap in spatio-temporal continuity, say. Or the new person might have to be made out of new matter. Or something else. But, whatever it is, imagine the magic done in such a way that it leaves unaltered the good and bad of every segment. For one thing, all the segments' experiences, including memory-experiences, will have to be the same as the experiences of the actual person.

Now, suppose that the person had an autonomous good independent of the good of her segments. According to our premise, if the magic I described was done, the good or bad in the world would be just as it actually is. So the person's good would still exist. But there would be no one it could belong to. There would be no one taking the place of this person. Instead there would be a conflation of two people. A conflation is not a person, and it is not the sort of thing that could possibly have an autonomous good. So the person could not have had an autonomous good after all. QED.

This argument leaves something to worry about. The magic it requires may seem to be impossible. It may seem that magic could not possibly sever the unifying relations without altering the good or bad of segments. To put it another way: the premise from

which I am arguing is a counterfactual conditional that may seem to have an impossible antecedent. We already know the antecedent is causally impossible because severing the unifying relations will cause the good or bad of segments to alter. But it may also seem metaphysically impossible. Whether or not one takes this view will depend on one's theory of what the good or bad of segments consists in. But here is a plausible example. It might plausibly be claimed that it is bad for a segment to be deceived about its past by its memory-experiences. And the magic will surely cause bads of this sort. After the magic has happened, later segments will have memory-experiences of actions that will lead them to believe these actions were done by earlier segments of the same person. But actually they were not. Because the magic has severed the unifying relations, the person who performed these actions was someone different. So the later segments are deceived by their memory-experiences.

I think, however, that this worry can be overcome. The antecedent of the conditional is not impossible, though it may be a more remote possibility than I have described so far. If necessary, we can envision the possibility as follows.[18] Take a possible world in which there are two perfect duplicates of our person, living duplicate lives on duplicate planets. Call them "P" and "Q." And in this possible world, let the counterparts of segments from the first half of the actual person's life be the corresponding segments of P's life, and let the counterparts of segments from the second half of the actual person's life be the corresponding segments of Q's life. In these circumstances, I cannot see how the good and bad of each segment could fail to be just the same as the good and bad of its counterpart. Certainly, no segment in the possible world is deceived by its memory-experiences; each segment will have genuine memories of the actions of earlier segments of the same person. But this collection of counterpart segments in the possible world is not linked by the unifying relations. So the counterparts of the actual segments have just the same good and bad as the actual segments, but the unifying relation does not hold between them. That is what was wanted.[19]

I conclude, then, that the argument is successful. The premise that the unifying relations are not ethically significant, taken together with a disuniting metaphysics, implies that a person does not have an autonomous good. Her good supervenes on the good of her segments. This is (apart from the matter of direction I mentioned) the Principle of Temporal Good. So it only remains to ask whether the premise can be defended.

Derek Parfit's work[20] provides one model for a defence. It contains extensive arguments about what "matters" (as Parfit puts it) ethically in the relations between person-segments. Parfit argues for a disuniting metaphysics, and he argues that the unifying relations between a person's segments, in virtue of which they make up a person, are psychological connections. He includes connections of memory, intention and so on: one segment of a person remembers what another did, one carries out intentions formed by another, and so on. Normally (excluding some fictional cases of fission and suchlike), a chain of segments will make up a person if and only if each segment is connected psychologically in the appropriate way to another.

On Parfit's account, then, our premise amounts to the claim that these psychological connections are not ethically significant. Parfit calls this view "extreme," but says it is defensible on the basis of his arguments.[21] So here is one way the premise might be defended.

Parfit's work also offers another way. Parfit believes that, speaking roughly, a person's life may be divided into periods that are not very closely connected psychologically. An old person is not very closely connected with the young person she once was. These are the "rough subdivisions within lives" mentioned in the quotation at the beginning of this section. Suppose we treat each of these periods as a segment.[22] They are connected psychologically, and they therefore make up a person. But the connections may be weak enough to be ethically insignificant, or at least unimportant. Parfit does not consider this an extreme view; indeed it appears to be his own. It is a defence of our premise, if an imprecise one.

So Parfit's arguments may offer a basis for the premise that the unifying relations are not ethically significant. Is it a metaphysical basis? There are two steps to the argument: first that the unifying relations are psychological connections, and second that psychological connections are not ethically significant. Parfit's defence of the first is purely metaphysical; it is about the nature of a person. But he does not really defend the second. It may turn out to require a less purely metaphysical argument.[23]

Section 6

Section 5 does not contain a complete argument. But suppose a complete argument was made out deducing the Principle of Temporal Good from a disuniting metaphysics. Section 3 shows that the Principle of Temporal Good, if true, lends very strong support to the Utilitarian Principle of Distribution. So what would have been achieved is an argument from the disuniting metaphysics to the Utilitarian Principle.

This link between a disuniting metaphysics and the Utilitarian Principle has been argued for before. But I believe, firstly, that my argument, if completed, could provide a new and much tighter link. The Principle of Temporal Good is a precise objective for the metaphysical argument to aim for. If it can reach it, then the theorems and their associated formal arguments will bring the Utilitarian Principle in train.

And I believe, secondly, that the previous arguments have not actually succeeded in making the link properly. I am going to review the most thorough attempt - Derek Parfit's[24] - in order to explain where I think it fails.

Parfit's argument starts from his disuniting metaphysics, which implies that "the rough subdivisions within lives [are], in certain ways, like the divisions between lives."[25] He takes this to imply that the distribution of good across time should be regarded, to an extent, in the same way as distribution across people. If, for instance, it is unfair to impose a burden on one person for the sake of benefitting someone else to a greater extent, then it may be unfair to impose a burden on a child for the sake of benefitting her to a greater

extent in later life. Or if some value should be attached to equality in the distribution of good between people, then some value should be attached to evenness in the distribution of good through a life. To put it another way: principles of distribution,[26] whatever they are, should be applied, not to people, but to smaller units: person-segments.

Parfit takes this to be a consequence of his disuniting metaphysics. Because the subdivisions within lives are, in certain ways, like the divisions between lives, then to some extent the same principles of distribution should apply. This is not obvious. It is not obvious that the ways in which the divisions are alike are ways that make it right to regard distribution similarly. But Parfit supposes that the ways in which they are alike include all the ways that matter ethically. So the conclusion follows. Let us take that for granted.

The disuniting metaphysics, then, implies that distribution between lives and within lives should be regarded similarly to some extent. But it does not obviously have any implications about whether it is the Utilitarian Principle or some other that should be applied both between and within lives. It is this step that is negotiated by the theorems and the argument of Section 3 above. That is what I think this formal argument contributes.

Parfit, however, believes he can get from the metaphysics to the Utilitarian Principle without the help of the formal argument. At least, he believes that, without its help, he can use the metaphysics to make the Utilitarian Principle plausible.[27] Here is part of his argument:[28]

> Consider the relief of suffering. Suppose that we can help only one of two people. We shall achieve more if we help the first; but it is the second who, in the past, suffered more. Those who believe in equality may decide to help the second person. This will be less effective; so the amount of suffering in the two people's lives will, in sum, be greater; but the amounts in each life will be made more equal. If we accept the Reductionist View [Parfit's disuniting metaphysics], we may decide otherwise. We may decide to do the most we can to relieve suffering. To suggest why, we can vary the example. Suppose that we

> can help only one of two nations. The one that we can help the most is the one whose history was, in earlier centuries, more fortunate. Most of us would not believe that it could be right to allow mankind to suffer more, so that the suffering is more equally divided between the histories of different nations. In trying to relieve suffering, we do not regard nations as the morally significant unit. On the Reductionist View, we compare the lives of people to the histories of nations. We may therefore think the same about them. We may believe that, when we are trying to relieve suffering, neither persons nor lives are the morally significant unit. We may again decide to aim for the least possible suffering, whatever its distribution.

I do not think, however, that this argument has moved us forward. We have already granted, given the disuniting metaphysics, that we should take the same attitude to distribution within a life as we take to distribution across lives. This is because the boundaries within lives are like the boundaries between lives. So we do not regard people as the morally significant units. This only means that, if we are concerned with distribution at all, we shall be concerned with distribution between what <u>are</u> the morally significant units, namely person-segments or whatever these divisions of a person are. So certainly the fact that a person has suffered more in the <u>past</u> will not make us give extra weight to relieving her suffering now. But if she is suffering more <u>now</u>, we may give extra weight to it. We may be concerned to equalize the distribution of good between person-segments. So all this argument does is remind us that we have changed the units of distribution. It does not suggest that we should be less interested in distribution between them.[29]

Parfit also says:[30] on the Reductionist View "it becomes more plausible to focus less upon the person, the subject of experiences, and instead to focus more upon the experiences themselves. It becomes more plausible to claim that ... we are right to ignore whether experiences come within the same or different lives." Once again I can make the same answer. Focusing on experiences is not the same thing as aiming to maximize the total

goodness of experiences. Suppose an experience is "the morally significant unit," so the unit has shrunk to something even smaller than the person-segment. We still might be interested in equality between units. We might be interested in equalizing the goodness of experiences.

But Parfit argues that if the morally relevant unit is shrunk so small as to be just an experience, then distributive principles become less plausible.[31] His argument seems to be this. If we apply distributive principles with experiences as the units, we shall give the greatest importance to reducing the badness of the worst experiences, to reducing the greatest suffering. Is this plausible? What is so bad about great suffering that it should require such particular attention? Well, suffering strikes us intuitively as <u>really</u> bad when it is the same person who is constantly suffering. But once we have shrunk our morally relevant unit, the fact that it is the same person constantly suffering cannot actually count. So we have removed the reason for giving special importance to reducing the badness of the worst experiences.

I think this argument is inadequate. The question is: should we apply distributive principles across experiences? If we can reduce the badness of an experience by some particular amount, is that a better thing to do if the experience is originally a very bad one than if it is originally less bad? This is a subtle quantitative question, and you cannot answer it without having some metric of badness for experiences. Parfit's argument does not have the necessary quantitative precision. My argument, on the other hand, examined the metric of goodness and badness in detail, and that was crucial to the argument.

I conclude, therefore, that Parfit's argument does not get where it wants to go. Even if we grant, on the basis of the disuniting metaphysics, that the units of distribution are person-segments or something smaller, that does not establish the Utilitarian Principle of Distribution.

Section 7

My own argument is not complete. But I hope it could be more successful in the end. Its linchpin is the Principle of Temporal Good. The disuniting metaphysics may be able to support this principle, and it in turn supports the Utilitarian Principle of Distribution. This second step is achieved using the formal theorems of Sections 2 and 3. These theorems are needed, and Parfit's argument fails because he does not use them.

NOTES

1. The most thorough argument to this effect is Parfit's (1984, 329-357). An article by Mirrlees (1982) makes the same suggestion, implicitly and perhaps not deliberately. Mirrlees's argument provided the stimulus for mine. On the other hand, Rawls (1974) has argued in general that moral theory is independent of metaphysical arguments about personhood. He agrees that utilitarianism may be implied by a disuniting theory of personhood. But he claims that one's theory about personhood will be determined by one's moral theory and not by independent nonmoral considerations. I think, though, that a disuniting metaphysics might be defended on nonmoral grounds, as indeed it has been by Lewis (1983c, 76-77) and Parfit (1984, Part III). So if the disuniting metaphysics can really be shown to support the Utilitarian Principle of Distribution, that will provide a metaphysical defence of the principle.

2. Section 6 explains why I find Parfit's argument inadequate. Schultz (1986) offers a critique of a different sort.

3. For instance, it is plausibly good for a person to achieve literary immortality (especially if that is what she wants), but not plausibly a part of her wellbeing. She will be dead when it happens.

4. Broome (1987).

5. Harsanyi seems to think this (1955), and so does Jeffrey (1971).

6. The first proof was Harsanyi's (1955). There is a rigorous version of Harsanyi's proof in Fishburn (1984). Proofs within different versions of decision theory include Broome (forthcoming) and Hammond (1983).

7. Harsanyi (1977) claims that the theorem directly supports utilitarianism.

8. I am not appealing to a verificationist theory of meaning. Verificationism has been influential in economics. "Utility" has often been said to be meaningless except in so far as it represents choices, because only choices make it empirically verifiable (eg Arrow 1973, 104). That is not what I am saying. I am not saying: weighing differences in good gives the notion of quantities of good its meaning because it makes the notion empirically verifiable. It does not do that anyway. The result of weighing differences in good is a judgment that one or the other alternative is better, and that is not empirically verifiable. I am simply appealing to a general and obvious connection between meaning and use.

9. The number of times has to be finite if the existing proofs of Theorem 2 are to work. I do not assume that everyone is alive at all times. It may be that all prospects are equally good for a person at all times when she is not alive, but I do not assume that either.

10. This needs better justification than I can give it here. My argument in favor of coherence is contained in another article (Broome 1990b). It is conducted in terms of reasons, and the reasons could be confined to reasons directed towards the good of a person at a time. The argument would still work, and would then show the coherence of the person's dated betterness relation.

11. It is easy to think of things that seem good for a person without being good for her at any particular time. One example is success. Suppose someone works to achieve some aim, and succeeds. Her success seems, prima facie, to be good for her. But it is often hard to know at what date it can be good for her. Derek Parfit's example (1984, 151) is a person who works for much of her life to save Venice from the sea. If Venice really is saved, this will make her work worthwhile, whereas if Venice is eventually swamped her work will have been pointless. That her work is worthwhile seem good for her, but it is hard to know when she receives this good. Whether or not Venice is saved may not even be determined until after her death.

Examples like this are inconclusive. Someone who believes in the Principle of Temporal Good can deal with them in two ways. She can deny that there really are goods of this sort; for instance, she might deny that success (as opposed to feeling successful or believing you are successful) is

really a good. Or she can assign some date to the good. For instance, she might date the good of success to all the dates when a person is working for it.

The example in the text is a different sort because it links time and uncertainty.

12. For simplicity, think of them as quantities (numbers). But actually they might be elements in some partially ordered scale. It is not necessary for this example that good should be measurable on a numerical scale.

13. This individualist account of the value of equality is spelt out in Broome (1989 and 1990a).

14. A different argument might rely on some particular theory of what a person's good consists in. Take, for instance, the theory that it consists in good feelings such as pleasure. Good feelings must all occur at some time or other. So it may look as though the Principle of Temporal Good would follow from this theory. But actually that is not so. This feeling theory of good says only what the goodness of an outcome consists in; it says nothing about the goodness or badness of risk. Look again at the example of Table 3. The symbols "x" and "y" stand for quantities of good, which according to the feeling theory are quantities of good feelings. A subscriber to the feeling theory might reasonably think that prospect B is better than A, because from the standpoint of the person as a whole it is less risky. The feeling theory of good is consistent with recognizing the standpoint of the person as a whole. So this theory, at least, cannot support the Principle of Temporal Good. It may be that some other theory of good might do so, though.

15. Parfit (1984, 333-334).

16. This is a popular view, but it weakens the analogy between a person and a society. It is not plausible that a person is a part of a society, at least as mereology conceives a part. In mereology, a part of a part is a part. But it is not very plausible that a part of a person, such as her finger, is a part of society. For an account of mereology see Simons (1987).

17. In a trivial sense, this must be so. Suppose the person has the property F. If, instead, she had not had this property, then all of her segments would have been different in at least this respect: they would have had the property of being segments of a person who does not have property F. But I am speaking of supervenience in a nontrivial sense. To define it adequately I would need

to rule out in some way such trivial properties of segments. Lewis's way (1983a, 359) would be to define supervenience in terms of the intrinsic or nonrelational properties of the segments. But this will not do, at least for my purposes. Whether or not a person has the property of leading a worthwhile life may depend on whether or not she is in love at some time, and we would want this to be consistent with supervenience. But being in love is not an intrinsic property, at least as Lewis understands "intrinsic" (see also Lewis 1983b).

18. In this paragraph I am adopting the terminology and modal semantics of Lewis (1986).

19. Taking the counterfactual in this more elaborate way also makes the final step of my argument particularly clear. If the person had an autonomous good, then so would the collection of counterpart segments in the possible world. But it is particularly clear that this collection is not the sort of thing that can have a good of its own. It is a collection of segments living on different planets.

20. Parfit (1984, Part III).

21. Parfit (1984, 343).

22. Theorems 2 and 3 do not require there to be many "times." Two is enough.

23. Susan Wolf (1986) has argued, in response to Parfit, that what matters ethically is not a metaphysical question. She says there is a reason why we should care about people rather than, say, person-segments: if we cared about person-segments and not people, the world would be less good than it actually is. This reason is independent of metaphysics; it will still exist even if a disuniting metaphysics is true. I am sure Wolf is right. But it does not affect what I am saying. I am concerned with whether the unifying relations are ethically significant as I defined ethical significance. This is not a question about what we should care about. (I believe, too, that what Parfit means by "mattering" is much closer to what I mean by "ethically significant" that it is to Wolf's meaning.)

24. Parfit (1984, 329-347).

25. Parfit (1984, 333-334) quoted earlier.

26. Parfit himself confines the term "distributive principles" to principles other than the utilitarian

one. My usage differs from his.

27. Parfit (1984, 342).

28. Parfit (1984, 341).

29. Nagel (1979, 124-125 note) makes this point too, in response to an earlier argument of Parfit's. Parfit (1984, 343-344) replies to Nagel, but I do not think his reply adds much to the earlier argument.

30. Parfit (1984, 341).

31. Parfit (1984, 345).

REFERENCES

Arrow, Kenneth J (1973) "Some ordinalist-utilitarian notes on Rawls's theory of justice" Journal of Philosophy 70 245-263. Reprinted in his Collected Papers Volume 1: Social Choice and Justice Blackwell 1984. (Page reference to the reprinted version.)

Broome, John (1987) "Utilitarianism and expected utility" Journal of Philosophy 84 405-422

Broome, John (1989) "What's the good of equality?" in Hey (1989)

Broome, John (1990a) Weighing Goods Blackwell

Broome, John (1990b) "Rationality and the sure-thing principle" in Meeks (1990)

Broome, John (forthcoming) "Bolker-Jeffrey decision theory and axiomatic utilitarianism" Review of Economic Studies

Elster, Jon (ed) (1986) The Multiple Self Cambridge University Press

Fishburn, Peter C (1984) "On Harsanyi's utilitarian cardinal welfare theorem" Theory and Decision 17 21-28

Hammond, Peter J (1983) "Ex-post optimality as a dynamically consistent objective for collective choice under uncertainty" in Pattanaik and Salles (1983) 175-205

Harsanyi, John C (1955) "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility" Journal of Political Economy 63 309-321

Harsanyi, John C (1977) "Morality and the theory of rational behavior" Social Research 44, reprinted in Sen and Williams (1982) 39-62

Hey, John (ed) (1989) Current Issues in Microeconomics Macmillan

Jeffrey, Richard (1971) "On interpersonal utility theory" Journal of Philosophy 68 647-656

Lewis, David (1983a) "New work for a theory of universals" Australasian Journal of Philosophy 61 343-377

Lewis, David (1983b) "Extrinsic properties" Philosophical Studies 44 197-200

Lewis, David (1983c) Philosophical Papers, Volume 1 Oxford University Press

Lewis, David (1986) On the Plurality of Worlds Blackwell

Meeks, Gay (ed) (1990) Rationality, Self-Interest and Benevolence Cambridge University Press

Mirrlees, J A (1982) "The economic uses of utilitarianism" in Sen and Williams (1982) 219-238

Nagel, Thomas (1979) Mortal Questions Cambridge University Press

Parfit, Derek (1984) Reasons and Persons Oxford University Press

Pattanaik, P K and Salles, M (eds) (1983) Social Choice and Welfare North-Holland 175-205

Rawls, John (1974) "The independence of moral theory" Proceedings of the American Philosophical Association 48 5-22

Schultz, Bart (1986) "Persons, selves and utilitarianism" Ethics 96 721-745

Sen, Amartya and Williams, Bernard (eds) (1982) Utilitarianism and Beyond Cambridge University Press

Simons, Peter (1987) Parts: A Study in Ontology Oxford University Press

Wolf, Susan (1986) "Self-interest and interest in selves" Ethics 96 704-720