Soc Choice Welfare (1993) 10: 57–68

Social Choice and Welfare

A cause of preference is not an object of preference

John Broome

Department of Economics, University of Bristol, Bristol BS8 1TN, UK

Received December 24, 1991 / Accepted June 24, 1992

Abstract. Welfare economists sometimes treat a cause of preference as an object of preference. This paper explains that this is an error. It examines two examples where the error has occurred. One is from the theory of endogenous preferences. The other is from the theory of extended preferences. People have erroneously been led to believe that everyone must have the same extended preferences, and this has led them to think that extended preferences can be the basis of interpersonal comparisons of wellbeing. But actually the basis of interpersonal comparisons must come from elsewhere.

1. Introduction

Welfare economists sometimes treat a cause of preference as an object of preference. Some do so inadvertently, others with enthusiasm. 'If two persons have preferences which appear to differ', says Serge-Christophe Kolm, 'there is a reason for this, there is something which makes them different from each other. Let us place this 'something' within *the object of the preferences* which we are considering.' (Kolm 1972, pp 79–80; translated by Rawls 1982, p 174.) But to treat a cause of preference as an object of preference is an error, and it leads to erroneous conclusions.

In Sect. 2, I shall explain the error in general terms. I shall use simple examples that make it obvious. Then I shall describe actual cases of error from the literature of welfare economics. Section 3 describes one from the literature on endogenous

Note. Presented at the conference on Social Choice and Welfare, Caen, June 1992. I have greatly benefited from correspondence and discussions on the subject of this paper, at the conference and elsewhere, with Kenneth Arrow, John Harsanyi, Susan Hurley, Serge-Christophe Kolm, John Roemer, T.M. Scanlon, Brian Skyrms, Hans-Peter Weikard, Menahem Yaari and an anonymous referee. The research for this paper was funded by the Economic and Social Research Council under grant R000 23 3334.

preferences. Sections 4 and 5 describe an argument about extended preferences that has been influential in the debate about interpersonal comparisons of wellbeing. The argument aims to show that everyone must have the same extended preferences, provided extended preferences are defined widely enough. This has led people to believe that extended preferences can provide the basis for interpersonal comparisons of wellbeing. The argument is based on the causes of preference, and Sect. 5 shows it is erroneous because it confuses causes of preference with objects of preference.

In the literature, this causal argument is sometimes not clearly separated from a quite different argument for the same conclusion. To reduce the risk of misunderstanding, Sect. 6 describes this second argument and raises an objection to it as well. I see no reason why different people should have the same extended preferences. Consequently, I do not think that interpersonal comparisons of wellbeing can be derived from extended preferences. I have no doubt we can compare the wellbeing of different people, but not on the basis of extended preferences.

2. The error

Suppose you have preferences over some range of alternatives, which are represented by a utility function U. $U(x)$ is the utility of an alternative x (which may be a vector). For instance, suppose your preferences between c, the amount of comfort in your life at a particular time, and e, the amount of excitement, are represented by the function

$$U(c,e)=\alpha \log c+(1-\alpha)\log e \ , \tag{2.1}$$

where α is a parameter between 0 and 1. These same preferences might equally well be represented by

$$V(c,e)=\log c+(1/\alpha-1)\log e \ . \tag{2.2}$$

V is a increasing transform of U in which the parameter α happens also to be a parameter of the transformation.

Now suppose there is a causal explanation of why you have the preferences you do. The form of your utility function U depends on some causal variable ϕ (which may be a vector). The function may be written $U_\phi(x)$. An alternative notation is $U(x;\phi)$. For instance, suppose your preferences between comfort and excitement depend on your age at the time; suppose the parameter α happens to be $t/100$, where t is your age. The utility function

$$U(c,e;t)=(t/100)\log c+(1-t/100)\log e \ , \tag{2.3}$$

obtained by substituting $t/100$ for α in (2.1), represents your preferences. Let us call a utility function of this sort, containing one or more causal variables, a *causal function.*

In a causal function, the causal variables play a quite different role from the variables denoting the objects of preference – the 'object variables', let us call them. If the object variables change in such a way as to increase your utility, that means you prefer the new values to the old ones. This is the whole point of representing your preferences by a utility function. But if the causal variables change in such a way as to increase your utility, that does not mean you prefer

the new values to the old ones. Your preferences about the causal variables are not represented in the function at all. In (2.3), if c is greater than e, U increases with t. But this does not mean you like getting older. Your preferences about age are not represented in the function.

To reinforce this obvious point, notice that your preferences could equally well be represented by the function

$$V(c, e; t) = \log c + (100/t - 1) \log e \ ,$$

which is obtained by substituting $t/100$ for α in (2.2). V is an increasing transform of U if the arguments of U are taken to be c and e, and t is treated as a parameter of the function. But t is also a parameter of the transformation, so V is not an increasing transform of U if the arguments of U are taken to be c, e and t. The transformation gives us another causal function representing exactly the same preferences, and exactly the same causal determination of those preferences. But whereas U increases with t provided c is greater than e, that is not so for V. V decreases with t provided e is greater than 1. So the fact that U increases with t cannot possibly indicate that you prefer old age.

Le me write out formally what transformations of causal functions are possible:

> If $U(x; \phi)$ represents preferences over objects x, which are determined by causes ϕ, then another function $V(x; \phi)$ represents the same preferences and the same causal determination if and only if $V(x; \phi) = f(U(x; \phi), \phi)$, where f is increasing in its first argument.

Under a transformation like this, a change in ϕ that increases U need not increase V.

Plainly, it is vital to respect the semicolon in $U(x; \phi)$; it must not be mistaken for a comma. The causal variable ϕ can formally be treated as an argument in the function, but it retains the distinct character of a parameter. A cause of preference must not be mistaken for an object of preference. The point is obvious, and no one would make the mistake in simple cases. But some authors have made it in more complicated cases, and drawn erroneous conclusions as a result. This paper aims to expose their error.

I can mention one complication immediately. Sometimes the causes of a person's preferences are amongst the things the person has preferences about. When I say 'a cause of preference is not an object of preference', I do not mean that no cause of preference is ever an object of preference. When I say 'an alderman is not an almoner', I do not mean that no alderman is ever an almoner. An alderman may also be an almoner, but her role as alderman is different from her role as almoner. It is the same with a cause of preference that is also an object of preference.

Suppose, for instance, to change the example, that your preferences about comfort and excitement are themselves affected by the amount of comfort you are having. Suppose comfort is addictive. Specifically, suppose the parameter α in (2.1) is not $t/100$ as before, but $c^h/100$, where c^h is the amount of excitement you are having. We must be especially careful about notation; c^h is the amount of excitement you are *having*, which determines the form of your preferences, whereas c and e are the amounts of comfort and excitement you are *contemplating* as objects of your preferences. You have a causal function

$$U(c, e; c^h) = (c^h/100)\log c + (1 - c^h/100)\log e \ ,$$

obtained by substituting $c^h/100$ for α in (2.1). For a constant c^h, this function represents your preferences if you are having c^h of comfort. For instance, compare the alternative (60, 40) - 60 units of comfort and 40 of excitement - with (40, 60). $U(60, 40; 60)$ is greater than $U(40, 60; 60)$. This means that, when you are having 60 units of comfort, your prefer (60, 40) to (40, 60). On the other hand $U(60, 40; 40)$ is less than $U(40, 60; 40)$. So if you found yourself experiencing 40 units of comfort, your preferences between the alternatives would change around.

Can we draw any conclusion from the function about some sort of overall preference ordering? Does the function, perhaps, tell us whether you are better off with (60, 40) or with (40, 60)? Certainly not. The causal function represents many different preference orderings, one for each value of the causal variable. It tells us nothing about any overall ordering. It may be tempting to merge c with c^h, and write

$$W(c, e) = (c/100)\log c + (1 - c/100)\log e \ . \tag{2.4}$$

We might hope that W would tell us how well of you are with different values of (c, e). But it does not. If this is not obvious, notice that your preferences and their causes could equally well be represented by the function

$$V(c, e; c^h) = \log c + (100/c^h - 1)\log e \ ,$$

obtained by substituting $c_h/100$ for α in (2.2). Merging c with c^h in V gives us the function

$$W'(c, e) = \log c + (100/c - 1)\log e \ . \tag{2.5}$$

W' has just as good a claim as W to represent how well of you are. But W' and W cannot both represent how well off you are, because W' is not an increasing transform of W. In truth, neither W nor W' has any good claim at all.

3. Endogenous preferences

My first example of error from the literature is Menahem Yaari's (1978) treatment of endogenous preferences. (Yaari is evidently aware of the mistake, and in a later paper on the same subject (Yaari, unpublished), he is careful to avoid it.) Yaari models addiction as follows. Let $(x_1, x_2, \ldots, x_T)$ be a person's sequence of consumptions through time. Each x_t is a vector. For simplicity let us suppose it has only two components: consumption of whisky w_t and consumption of bread b_t. At each time t, the person has a utility function defined on the whole sequence of consumptions:

$$U_t(x_1, x_2, \ldots, x_T) \ .$$

To represent the addictiveness of whisky, Yaari imposes a condition on the form of each utility function. He assumes the person's marginal rate of substitution of whisky for bread is an increasing function of the amount of whisky she has already consumed. That is to say, in the function U_t, the marginal rate of substitution of w_τ for b_τ, for all $\tau \geqq t$, increases with $w_1 + w_2 + \ldots w_{t-1}$. This means

that, in the function U_t, the consumptions $w_1, w_2, \ldots w_{t-1}$ are working as causal variables rather than object variables.

Let us deal with a simple example once again. Suppose there are only two times. At time 2 the person has preferences about the whisky and bread she consumes at that time. Let us suppose her utility function takes the form $\alpha w_2 + b_2$, where α is a parameter. Since whisky is addictive, the parameter α is determined by her consumption w_1 of whisky at time 1. Let us simply make $\alpha = w_1$. Then her causal function is

$$U_2 = w_1 w_2 + b_2 \ . \tag{3.1}$$

This function conforms to Yaari's condition on marginal rates of substitution. In this function, w_1 is a causal variable, whereas w_2 and b_2 are object variables. The difference shows up in the transformations that are possible. For instance, the same preferences and their causation could be represented by the function

$$U_2' = w_2 + b_2/w_1 \ , \tag{3.2}$$

obtained by dividing U_2 by w_1. But they cannot be represented by $w_1 + b_2/w_2$, obtained by dividing U_2 by w_2. U_2', like U_2, also conforms to Yaari's condition on marginal rates of substitution.

There is nothing wrong with having variables of two different sorts, playing two different roles, in the same function. But it creates a risk of error, and Yaari falls into error. He takes for granted an intertemporal Pareto condition: he assumes that one sequence of consumptions $(x_1, x_2, \ldots, x_T)$ is better for the person than another if it gives a higher value for one of her utility functions U_t, and a lower value for none of them. He assumes, then, that increasing U_t for any t is good for the person, other things being equal. But this is not necessarily so. In (3.1), increasing w_1 increases U_2. Is this good for the person? There is no reason to think so, since w_1 is a causal variable. At time 2, the person may have no preferences about how much whisky she has consumed at time 1. The fact that increasing w_1 increases U_2 is merely an artifact of the particular representation we have selected. Increasing w_1 *decreases* U_2' in (3.2); yet U_2' also represents the preferences. Yaari's Pareto condition is therefore mistaken.

4. Extended preferences

My second example is a popular theory about interpersonal comparisons of wellbeing. This theory was originally motivated by the doctrine of ordinalism; were it not for ordinalism, there would be no need for it. One principle of ordinalism says that, of two alternatives a and b facing a person, a would be better for the person than b if and only if the person prefers a to b. I shall call this 'the preference-satisfaction condition on wellbeing'. A second principle is epistemological. It says our knowledge of a person's wellbeing can derive only from preferences. Most ordinalists draw the conclusion that we cannot make interpersonal comparisons of wellbeing. One person's preferences, they say, tell us about her wellbeing, and another person's about hers, but no one's preferences tell us how one person's wellbeing compares with another's. More specifically, these ordinalists conclude that we cannot know whether or not an alternative a would be better for one person than alternative b would be for another.

Some ordinalists, however, have argued that interpersonal comparisons of wellbeing are possible within ordinalism. They appeal to the notion of *extended* preferences. Think of alternatives such as the a and b I have mentioned already as conditions of life: where you live, your consumption of materials goods, the education your receive, and so on. We can imagine living under various alternative conditions of life, and we have preferences amongst them. We can also imagine taking on other people's personal characteristics – their values, physical features and so on – and we can have preferences amongst these things too. Indeed, we can have preferences amongst conditions of life and personal characteristics taken together. Call pairs like this – conditions of life together with personal characteristics – *extended alternatives*. A typical one is (a, χ_i), where a stands for some conditions of life, and χ_i for the characteristics of person i. People have preferences between extended alternatives; call these *extended preferences*. Since they are *preferences*, the argument goes, from an ordinalist point of view they are an acceptable basis for determining whether one alternative is better than another. In particular, we may be able to determine from them whether or not conditions of life a would be better for i than conditions b would be for j.

I think, however, that the attempt to derive interpersonal comparisons of wellbeing from extended preference is unsuccessful. I shall explain why. This does not mean I doubt the possibility of interpersonal comparisons of wellbeing. Far from it. It means I doubt these comparisons can be reconciled with ordinalism. So much the worse for ordinalism, I say.

The idea is that we may be able to determine from extended preferences whether or not conditions of life a would be better for i than conditions b would be for j. How might that be done? Let us take the derivation one step at a time. According to the preference-satisfaction condition on wellbeing, if I prefer (a, χ_i) to (b, χ_j), then (a, χ_i) would be better for me than (b, χ_j). That means, presumably, that it would be better for me to live in conditions a having the characteristics of person i than to live in conditions b having the characteristics of person j. But this is not yet the interpersonal comparison of wellbeing we are looking for. We are looking for the conclusion that a would be better for i than b would be for j. Can we justifiably take this final step?

One minimum condition is necessary before that step could be justified. Suppose you have the opposite extended preference from mine: you prefer (b, χ_j) to (a, χ_i). Then according to the preference satisfaction condition, it would be better for you to live in conditions b having the characteristics of j than in conditions a having the characteristics of i. If that were so, we could scarcely draw the conclusion from the extended preference *I* happen to have that a would be better for i than b would be for j. Obviously, to draw that conclusion we would need some coincidence between the extended preferences of different people.

It would be nice if everyone's extended preferences were identical, because then they would provide a comprehensive basis for every interpersonal comparison we might need to make. Several authors have suggested that, indeed, everyone's extended preferences must be identical, provided we construe personal characteristics widely enough. Kenneth Arrow (1977, p 159), for instance, says:

> We may suppose that everything which determines an individual's satisfaction is included in the list of goods. Thus, not only the wine but the ability to enjoy and discriminate are included among goods. ...If we use this complete

> list, then everyone should have the same utility function for what he gets out of the social state.

The utility function Arrow is referring to is presumably a function representing a person's extended preferences. So Arrow is presumably saying that everyone should have the same extended preferences.

On the face of it, though, this assertion seems definitely incorrect. It seems that different people have different extended preferences. For instance, I myself prefer to live the life of an academic, with my own academic characteristics, even in the conditions allotted to academics in contemporary Britain, to being a financial adviser living in the conditions allotted to financial advisers. I would expect a financial adviser, with her different values, to have the opposite preference. So her extended preferences are different from mine. The reason I have mine is that an academic has some slight change of making a worthwhile contribution to knowledge. I recognize that, if I were a financial adviser, with all the characteristics of a financial adviser, I would not then value knowledge as I do now. Nevertheless, I do value knowledge, and that is why I prefer to be an academic.

Are there any good arguments to counter this prima facie example? Have we any good reason to think, despite the example, that people will all necessarily have the same extended preferences? One argument is based on the causation of preferences. It is particularly associated with John Harsanyi, and is spelt out in most detail in Harsanyi (1977, pp 57–60). The argument appears in rudimentary form in Harsanyi (1955, pp 17–18), and independently in Tinbergen (1957, p 501). It also appears in Kolm (1972, pp 79–80). I shall be making only a single point about Harsanyi's argument. Other important discussions of the argument occur in Hurley (1989, pp 105–20), MacKay (1986) and Scanlon (1991), but these authors do not mention the point I shall be making.

5. The causal argument

A person k has preferences over conditions of life x, which can be represented by a utility function $u_k(x)$. There is a causal explanation of why she has the preferences she does. Let ϕ be a full specification of all the causal variables that influence her preferences. We can represent her preferences and their causes by a causal function $u(x;\phi)$. This function need not be indexed by k because, since ϕ is a complete specification of the causes, anyone who is subject to those same causes will have the same preferences. The function is the same for everyone.

Within ϕ will be a number of things that people have preferences about. For instance, ϕ will include a person's level of education and her ability at tennis – things that many people care about. As I explained in Sect. 4, extended preferences take account of such things. Person k's extended preferences can be represented by an extended utility function $U_k(x,\chi)$. Some components of ϕ will be included in x and some in χ; it does not matter which. But let us make sure that x and χ are specified extremely widely, so that (x,χ) includes everything that anyone has preferences about, and also anything that has a causal influence on preferences.

As yet we have no reason to think everyone will have the same extended preferences; we are trying to develop an argument why they should. So we need the index k on $U_k(x,\chi)$. But there is a causal explanation of why a person has

the extended preferences she has. So we can represent the preferences by a causal function $U(x, \chi; \phi)$, which includes a full specification of the causal influences. (If there are more causal influences on extended preferences than on ordinary ones, ϕ will have to be extended to include them, and we must make sure that (x, χ) has been defined widely enough to include them too.)

The causal function U is a universal function representing extended preferences. It is universal – the same for everyone – because it embodies a complete specification of the causes of preferences. Its form depends on the laws of psychology, which in principle can be discovered from scientific observation. So the function can in principle be found from observable information about the form and causes of preferences. That means it should satisfy the epistemological requirements of ordinalism. A universal function representing extended preferences presumably represents universal extended preferences. So it should constitute a basis for interpersonal comparisons of wellbeing, consistent with ordinalist principles. That is Harsanyi's argument as best I understand it.

However, although U is a universal function, and although it represents preferences, it does not represent universal preferences. U is defined on x, χ and ϕ, but it does not represent a single preference ordering over x, χ and ϕ. It represents many different preference orderings over the objects of preference x and χ, one ordering for each value of the causal variables ϕ. We have definitely not discovered any universal extended preferences.

Perhaps what deceived Harsanyi is that (x, χ) includes all the components of ϕ. We deliberately defined extended alternatives (x, χ) very widely to make sure this was so. It may seem redundant to list all these components twice as arguments of the function U. Can we not think of $U(x, \chi; \phi)$ as a function $W(x, \chi)$ of just x and χ, and might not that function represent universal preferences over x and χ? Certainly not. Although all the variables included in ϕ are also included in (x, χ), it is different *values* of the variables in each case. In ϕ are the values a person actually experiences; these determine the form of the person's preferences. In x and χ are the values the person contemplates as objects of her preference. So the double appearance of the same variables is not redundant.

We *could* create a function of just x and χ by concentrating only on cases where the values of the variables in ϕ happen to be the same as they are in (x, χ). For any values of x and χ, let $\phi(x, \chi)$ be those particular values of the causal variables that are contained in x and χ. Then let

$$W(x, \chi) = U(x, \chi; \phi(x, \chi)) \ .$$

Here is a function of x and χ, but it does not represent preferences of any sort. Take two people i and j with characteristics χ_i and χ_j, and suppose $W(a, \chi_i)$ is greater than $W(b, \chi_j)$. This means that $U(a, \chi_i; \phi(a, \chi_i))$ is greater than $U(b, \chi_j; \phi(b, \chi_j))$. If, by chance, $\phi(a, \chi_i)$ is the same as $\phi(b, \chi_j)$, then $U(a, \chi_i; \phi(a, \chi_i))$ and $U(b, \chi_j; \phi(b, \chi_j))$ are values of the same utility function, representing preferences determined by the causal variables $\phi(a, \chi_i) = \phi(b, \chi_j)$. In this utility function (a, χ_i) has a greater utility than (b, χ_j). Consequently, (a, χ_i) is preferred to (b, χ_j) by anyone who has this particular utility function. (That includes i if she lives in conditions a and j if she lives in conditions b.) But normally $\phi(a, \chi_i)$ will not be the same as $\phi(b, \chi_j)$. If it is not, then $U(a, \chi_i; \phi(a, \chi_i))$ and $U(b, \chi_j; \phi(b, \chi_j))$ are values of *different* utility functions, and the fact that one is greater than the other tells us nothing about anyone's preferences. Normally, then, if $W(a, \chi_i)$ is greater than $W(b, \chi_j)$, that does not

tell us that anyone prefers (a, χ_i) to (b, χ_j). Indeed, it is perfectly possible for everyone to have the opposite preference. So the function W does not represent any sort of preference ordering.

$W(x, \chi)$ corresponds exactly to $W(c, e)$ in (2.4), and is just as worthless. Both result from conflating the object variables and causal variables – not respecting the semicolon. In Sect. 2, I brought out the worthlessness of $W(c, e)$ by comparing it with $W'(c, e)$ in (2.5). I can do the same thing now. The causal function $U(x, \chi; \phi)$ may be transformed into another causal function $V(x, \chi; \phi)$, representing the same preferences and their causes, according to the rule given in Sect. 2. Then a function W' may be defined by

$$W'(x, \chi) = V(x, \chi; \phi(x, \chi)) \ .$$

Since V need not be an increasing transform of U, when ϕ is treated as an argument in the function, W' need not be an increasing transform of W. If W has a claim to represent universal preferences, W' has just as good a claim. But they cannot both represent these universal preferences, since they are not increasing transforms of each other. In truth, neither does. This should have been obvious from the start. $U(x, \chi; \phi)$ represents a lot of different preferences orderings, and contains no information comparing one with another. So no amount of formal manipulation is going to turn it into a function that represents a single universal ordering.

I conclude that the causal argument fails. It does not show that everyone will have the same extended preferences. Serge-Christophe Kolm (1972, pp 79–80, translated by Rawls 1982, p 174) presents the idea of the argument very clearly:

> If two persons have preferences which appear to differ, there is a reason for this, there is something which makes them different from each other. Let us place this 'something' within *the object of the preferences* which we are considering, thereby removing it from the parameters which determine the structure of these preferences. The preferences of these two persons defined in this way are necessarily identical. We may carry out this operation in the case of any society: namely, the operation of placing in the object of preferences everything which would cause differences between the preferences of different members of society. An identical preference of all members of this society obtained in this way is called 'a fundamental preference' of the members of this society. It is a property which describes the tastes and needs of the 'representative individual' of this society.

Kolm seems to think that, by contemplating a cause of preference as an object of preference, I somehow remove myself from its causal influence, so that it ceases to be a parameter determining the structure of my preferences. But that is a fantasy. My position as an academic causes me to have particular values. Since those are my values, I cannot escape them, even when I am forming my preferences about lives in which I would not have those values.

6. An alternative argument

That completes the task I set for this paper: to reveal the error in the causal argument for the coincidence of people's extended preferences. But to reduce the risk of misunderstanding, I need to mention a second argument that is implicit

in Harsanyi's (1977, pp 57–60) presentation of the causal argument, interwoven with the causal argument itself. If I understand him, Arrow (1977, pp 159–60) also uses this argument in support of his claim that everyone will have the same extended preferences. I shall not try to refute this argument definitively, but I shall say where I think it fails.

Suppose it would be better for me to live in conditions a having the characteristics of i than in conditions b having the characteristics of j. If we understand characteristics broadly enough, to include every aspect of a personality, then if I were to have the characteristics of i, I would have nothing left of my own personality at all. Consequently, how good it would be for me to live in conditions a having the characteristics of i is simply how good it would be for i to live in those conditions. Therefore, it is also how good it would be for you or anyone else to live in conditions a having the characteristics of i. If it would be better for me to live in conditions a having the characteristics of i than in conditions b having the characteristics of j, then the same must also be true for anyone. But the preference-satisfaction condition on wellbeing tells us that it would be better for me to live in conditions a having the characteristics of i than to live in conditions b having the characterstics of j if and only if I prefer (a, χ_i) to (b, χ_j). The same is true for anyone. Therefore, if I prefer (a, χ_i) to (b, χ_j), so must anyone.

I agree with the whole of this argument up to the use of the preference-satisfaction condition at the end. But I now wish to question the preference-satisfaction condition when it is applied to extended preferences. The condition says, remember, that one alternative would be better for a person than another if and only if the person prefers it. It can be defended in two ways. The first is to say a person's wellbeing actually *consists* in the satisfaction of her preferences: it is better for you to get what you prefer just because you prefer it. This defence is not available in the context of extended preferences. It implies that, if I prefer (a, χ_i) to (b, χ_j), then (a, χ_i) would be better for me than (b, χ_j) just because I prefer it. But if (a, χ_i) would be better for me than (b, χ_j), that means it would be better for i to live in conditions a than for j to live in conditions b. And that can scarcely be true just because of a preference of mine.

So the preference-satisfaction condition must fall back on the second defence. This is to say that, although a person's wellbeing is conceptually independent of her preferences, nevertheless it happens that people prefer what would be better for them. This defence forces us to weaken the preference-satisfaction condition. If a person's wellbeing is conceptually independent of her preferences, people's preferences are bound to diverge from their wellbeing sometimes. In practice, people make mistakes, possess inadequate information, suffer from failings of rationality and so on. So the condition must be weakened to something like: one alternative would be better for a person than another if the person prefers it and if this preference is rational and well-informed.

This weakening of the preference-satisfaction condition weakens the conclusion that can be drawn from the argument I gave above. The argument cannot show that everyone's extended preferences coincide. At most it can show that rational and well-informed extended preferences coincide: if two people each have a rational and well-informed preference between a pair of extended alternatives, those preferences must be the same. Arrow (1977, p 160) acknowledges this limitation on his conclusion, and so does Harsanyi (1977, pp 59–60). So too

do Ignacio Ortuño-Ortin and John Roemer (1991), and they develop a theory to overcome it.

However, even the limited conclusion is implausible. My preference between the life of an academic and the life of a financial adviser differs from the financial adviser's preference. It is implausible that this is because of some irrationality or lack of information on the part of one of us. It is because we have different values. I value opportunities to add to knowledge; she values an opulent life-style. I very much doubt that either of these values is irrational, ill-informed or even objectively wrong. I doubt that there is objective fact of the matter whether the life of an academic is better or worse than the life of a financial adviser. Life contains goods of different sorts, and there is no objective scale that completely ranks quantities of one good against quantities of another. There is therefore room for differing values – assigning different weights to particular goods – none of which are irrational, ill-informed or objectively wrong.

For this reason, I believe the preference-satisfaction condition is false for extended preferences. I think the authors I have mentioned accept it because implicitly they accept a monistic theory of good. They believe that the goodness of a life consists in one thing – happiness, satisfaction, or something else – and each life delivers a particular quantity of this thing. Consequently, it is always a matter of objective fact which of two lives is better. I am sure that some lives are objectively better than others, but I doubt that every pair of lives can be objectively ranked.

7. Conclusion

The mistake of treating a cause of preference as an object of preference is easily made. In the context of extended preferences it has led people to believe in a universal extended preference ordering. Consequently, they have concluded that interpersonal comparisons of wellbeing can be derived from extended preferences. But actually there is no reason to believe in a universal extended preference ordering. If interpersonal comparisons of wellbeing are possible, as I am sure they are, they must be based on something other than preferences.

References

Arrow KJ (1977) Extended sympathy and the possibility of social choice. Am Econ Rev 67: 219–225. Reprinted in his collected papers vol. 1: Social Choice and Justice. Blackwell 1984: 147–161 (Page references to the reprinted version)

Harsanyi JC (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. J Polit Econ 63: 309–321. Reprinted in his Essays on Ethics, Social Behavior, and Scientific Explanation. Reidel, Dordrecht 1976 (Page references to the reprinted version)

Harsanyi JC (1977) Rational Behavior and Bargaining Equilibrium in Games and Social Situations. Cambridge University Press, Cambridge

Hurley S (1989) Natural Reasons. Oxford University Press, Oxford

Kolm SC (1972) Justice et Équité. Centre national de la récherche scientifique

MacKay AF (1986) Extended sympathy and interpersonal utility comparisons. J Philos 83: 305–322

Ortuño-Ortin I, Roemer JE (1991) Deducing interpersonal comparability from local expertise. In: Elster E, Roemer JE (eds), Interpersonal Comparisons of Well-Being. Cambridge University Press, Cambridge

Rawls J (1982) Social unity and primary goods. In: Sen A, Williams B (eds) Utilitarianism and Beyond. Cambridge University Press, Cambridge
Scanlon TM (1991) The moral basis of interpersonal comparisons. In: Elster J, Roemer JE (eds) Interpersonal Comparisons of Well-Being. Cambridge University Press, Cambridge
Tinbergen J (1957) Welfare economics and income distribution. Am Econ Rev 47: 490–503
Yaari ME (1978) Endogenous changes in tastes: a philosophical discussion. In: Gottinger HW, Leinfellner W (eds) Decision Theory and Social Ethics. Reidel, Dordrecht
Yaari ME Consistent utilization of an exhaustible resource: or how to eat an appetite-arousing cake (unpublished)