Oxford Economic Papers 48 (1996), 177–193

THE WELFARE ECONOMICS OF POPULATION

By JOHN BROOME

Department of Moral Philosophy, University of St Andrews, Fife KY16 9AL

Intuition suggests there is no value in adding people to the population if it brings no benefits to people already living: creating people is morally neutral in itself. This paper examines the difficulties of incorporating this intuition into a coherent theory of the value of population. It takes three existing theories within welfare economics—average utilitarianism, relativist utilitarianism, and critical-level utilitarianism—and considers whether they can satisfactorily accommodate the intuition that creating people is neutral.

1. The problem

MANY of the actions we take now will affect the future of the world's population. Some governments have deliberate population policies, and all governments have policies that affect population incidentally. For instance, the structure of a country's tax system will affect how easy it is for a couple to find a home, and so it will influence couples' decisions about when to have children. If welfare economics is to deal properly with policies like these, it must be able to set a value on changes in population.

It needs to be able to evaluate a range of alternatives involving different populations. Let X be the set of alternatives. Each $x \in X$ is a possible history for the world, in which particular people live, and in which each of these people has a particular quality of life. Let $J(x)$ be the set of people who live at some time in x. Let I be $\bigcup_X J(x)$, the set of people who live in some history or other. If a person i lives in x, so $i \in J(x)$, let $W_i(x)$ be i's lifetime well-being in x. If $i \notin J(x)$, let $w_i(x)$ be Ω. The symbol Ω is just a notational device; it indicates that a particular person does not live in a particular history. The variable $w_i(x)$ takes values in the real numbers augmented by Ω. The vector

$$w(x) = (w_1(x), w_2(x), w_3(x), \dots)$$

shows which people live in x and what their levels of well-being are if they live. I shall call this vector a distribution of well-being.

I assume a person's well-being is measured on a cardinal scale with intervals that are comparable between people. I define the zero on the scale to be the well-being of a person who leads a life without experiences of any sort: a life lived in a coma throughout. I assume that to live such a life would be equally as good, or bad, for one person as for any other. So setting zero at this level makes levels of well-being comparable between people.

Our problem is to find, between alternative histories, which is better than which. That is to say, we want to find the form of the betterness relation between histories, which I shall represent by $\succeq$. The notation $x \succeq y$ means x is at least as good as y. $x \succ y$ means x is better than y, and $x \approx y$ measns x and y are equally good. The relation $\succeq$ is transitive and reflexive, as a matter of logic: as a matter of logic, x is at least as good as itself, and if x is at least as good as y and y at least as good as z, then x is at least as good as z (see Broome, 1994, 173–5).

I shall assume that the relative goodness of two histories x and y depends only on their respective distributions of well-being. Furthermore, increasing a person's well-being makes a history better. That is:

Principle of personal good Suppose $J(x) = J(y)$. If $w(x) = w(y)$, then $x \approx y$. If $w_i(x) \geqslant w_i(y)$ for all $i \in J(x)$, and if $w_i(x) > w_i(y)$ for some $i \in J(x)$, then $x \succ y$.

This principle may seem to imply welfarism. But it does not, because I include within a person's well-being much more than is normally included in her welfare. For instance, if a person is treated unjustly, that is bad for her and counts negatively in her well-being, even if she feels no resentment.

Until Section 9, I shall assume the betterness relation can be represented by a value function v. Since the principle of personal good implies that a history's goodness depends only on its distribution of well-being, we can treat v as a function of the distribution

$$v(w_1(x), w_2(x), w_3(x), \ldots)$$

I assume the existence of a value function because much of the literature on the value of population takes one for granted, and I want to discuss the literature. I think there is unlikely to be a value function if we allow for infinite time, and hence for an infinite number of people (see Broome, 1992, 104–5). So, having assumed there is a value function, I may as well also assume that $J(x)$ is a finite set for all x.

2. The basic intuition

What is the form of the betterness relation? I do not know, but I think that in some way or other it must incorporate one basic intuition: that adding people to the population is not in itself valuable. If parents want a child, that is a reason for their having one. If extra people will add to the well-being of existing people in some way, that is a reason for having them. But there is nothing to be said for adding extra people on account of those extra people themselves. If the extra people would lead good lives and their existence would not harm existing people, there is nothing against adding them to the population, but there is nothing in favour of it either. Adding people with good lives is ethically neutral in itself.

I want first to show this is indeed an intuition many people share. To begin with, think about our common private morality. Suppose a couple are wondering whether to have a child. Their own interests are finely balanced; on balance they would neither gain nor lose by having a child. But, for some unimportant reason, they decide not to have one. Are they doing anything morally wrong? Few of us would say they are. Suppose the couple live in circumstances that make it almost certain that their child would lead a good life, and also that its existence would not harm other people. We would still not think they are doing wrong if they choose not to have the child. But if adding a person to the world is valuable in itself, in choosing not to have a child they make the world less good than it would have been had they made the opposite choice. Other things being equal, it is surely wrong to make the world less good than it would otherwise have been. Since our intuition tells us they are doing nothing wrong, it tells us that adding a person is not valuable for its own sake.

In passing, notice that this case also reveals a major limitation on the intuition that adding a person is morally neutral. Suppose the couple happened to have a genetic defect that meant any child of theirs would lead a short life, full of nothing but pain and misery. Then we would think it wrong for them to have a child. This suggests that adding a person is morally neutral only if the person would have a good life. It is morally bad if the person would have a bad life.

Next, as another indication of how common the basic intuition is, notice it is almost universally taken for granted in the welfare economics of life-saving. What are the benefits of saving a person's life? If her life is saved, the person lives longer than she would have done, and her extra term of life is one benefit of saving her. But if she is young, saving her life will probably also have the further effect of adding new people to the population. Most young people, if they survive, have children who would never have existed had they died. Indeed, they will probably have grandchildren and a whole line of descendants. But economists almost never count the well-being of the new people created as a benefit of saving a young person's life. (An exception is Arthur 1981.) Why not? It must be because of this same intuition that adding people is not valuable in itself. This is not the only place where this intuition is at work in welfare economics; there is another example in Section 4.

The intuition can also be backed by arguments. Two were first expressed by Jan Narveson (1967). Here is one of them. One history is surely only better than another if it is better for someone. But a history in which someone exists is not better for that person than a history in which she does not exist. If it was better for her, then the history where she does not exist would be worse for her. But that cannot be so, because if she does not exist she had no degree of well-being at all.

Narveson's second argument is this. Many of us think our moral duty is to promote people's well-being. But to whom do we owe this duty? A plausible suggestion is that we owe it to the people whose well-being we ought to promote. But we cannot owe anyone a duty to bring her into existence, because failing in such a duty would not be failing anyone. So if our duty is to promote people's well-being, and if that duty is owed to the people themselves, it is not fulfilled by bringing new people into existence.

3. Difficulties

So the intuition is common and defensible. Nevertheless, it is very hard to embody it in a consistent theory about the value of population. Derek Parfit

(1984, 351–79) has demonstrated that it leads to difficulties and contradictions. This section mentions some of them.

Suppose we wanted to express the intuition formally as a condition on the betterness relation. How should we do so? The intuition is that adding people is morally neutral. So it is natural to understand it as saying that, when two histories share the same population, except that one has some extra people who do not exist in the other, then the relative goodness of the histories depends only on their relative goodness for the people who exist in both. It does not depend on the well-being of the extra people. We must remember the proviso that this only applies if the lives of the extra people are good. If they are bad, that counts against the history in which they exist.

If J is a set of people, let $\succeq_J$ stand for the relation of betterness for the set J, so $x \succeq_J y$ means x is at least as good for the set as y. I shall leave the meaning if this expression vague, but I shall put this minimal definitional constraint on it:

Definitional constraint on betterness for people If $w_i(x) \geqslant w_i(y)$ for all $i \in J$, and $w_i(x) > w_i(y)$ for some $i \in J$, then $x \succ_J y$. If $w_i(x) = w_i(y)$ for all $i \in J$, then $x \approx_J y$.

Let w_0 be the lower bound on well-being below which a life will cease to count as good; w_0 need not be zero as I defined it, but in my examples I shall assume it is. Then we might express the basic intuition by:

Constituency condition Let x and y be histories such that $J(x) \subset J(y)$. Then $x \succeq y$ if and only if $x \succeq_{J(x)} y$, provided $w_i(y) \geqslant w_0$ for all $i \in J(y) - J(x)$.

That is to say: when two histories share the same population except that one has some extra people who do not exist in the other (and those people lead good lives), then one is at least as good as the other if and only if it is at least as good for the people who exist in both. The people who exist in both alternatives form a constituency that determines which is the better alternative.

The consituency condition seems a natural way to express the basic intuition. However, it is undoubtedly false. It contradicts the transitivity of betterness, which I have already said is a truth of logic.

Transitivity of betterness If x, y, and z are such that $x \succeq y$ and $x \succeq z$, then $x \succeq z$.

The contradiction is shown by this example

Example 1

$$w(a) = (1, 1, 2, \Omega, \Omega, \dots)$$

$$w(b) = (1, 1, \Omega, \Omega, \Omega, \dots)$$

$$w(c) = (1, 1, 1, \Omega, \Omega, \dots)$$

The constituency condition implies that $a \approx b$ and $b \approx c$, but that $a \succ c$. This contradicts the transitivity of $\succeq$. So the constituency principle is false.

I think most people's intuition would be able to cope with this example. To flesh it out, suppose it represents the dilemma of a couple wondering whether to have a child. The first two people are the couple; the third the child. In *b*, the couple have no child. In *a*, they have one and her well-being is 2. In *c*, they have the child but her well-being is only 1. Our basic intuition is that adding a person to the population is morally neutral, so it does not matter morally whether the couple have the child or not. A different intuition tells us that, if the couple do have a child, they should make sure she is as well off as possible. Therefore, faced with a choice between *a*, *b*, and *c*, the couple should definitely not choose *c*, but it does not matter which of *a* or *b* they choose. However, if *a* was not available, it would not matter which of *b* or *c* the couple chose.

So we learn from this example, not only that the constituency principle is false, but also that it does not adequately express our intuitions in the matter of population. The constituency principle cannot cope with the example, but our intuitions can.

But now look at another example, which is a version of Parfit's 'mere addition paradox' (Parfit 1984, 419–41).

Example 2

$$w(a) = (4, 4, 6, 1, \Omega, \dots)$$

$$w(b) = (4, 4, 5, \Omega, \Omega, \dots)$$

$$w(c) = (4, 4, 4, 4, \Omega, \dots)$$

Between *a* and *b* the constituency consists of the first three people. Both options are equally good for the first two. But *a* is better for the third person, so *a* is better than *b* according to the constituency condition. For the same reason *b* is better than *c*. Between *c* and *a* the constituency consists of all four people. Which is better for these four? That is not determined by my minimal definitional constraint on the idea of betterness for a set of people, since *c* is better for one of the people and *a* better for another. Nevertheless, we can fairly assume that *c* is better than *a* for the four people together, because *c* has a greater total of well-being, and furthermore has it more equally distributed. So the constituency principle says *c* is better than *a*. In this example too, therefore, it contradicts the transitivity of betterness.

I think this example may trouble even our natural intuitions. Think of parents with one child already, who are wondering whether to have another. In *b* they have no second child. In *a* and *c* they do have a second child, but in *a* well-being is unequally distributed between their children and in *c* it is equally distributed. Faced with a choice among all three options, this couple might find themselves in a genuine moral quandary. They might think *a* is better than *b* because it is better for their existing child, and worse for no one. They might think *b* better than *c* for the same reason: it is better for the existing child and worse for no one. And they might think *c* is better than *a* because it is better in total for the children and also distributes well-being equally between them. So they might not know the right thing to do. I do not think intuition provides a clear guide in this case.

Example 1, then, shows the constituency principle is wrong and does not adequately capture our intuitions, and Example 2 shows our intuitions themselves may be in some disarray. It is plainly not going to be easy to fit the basic intuition into a coherent account of the value of population. At the very least, we must expect to have to modify and weaken it. But some theories of population within welfare economics offer the hope of accommodating it to some extent. I shall review three of them.

4. Average utilitarianism

According to average utilitarianism, the betterness relation can be represented by this value function:

$$v(x) = \frac{1}{n(x)} \sum_{i \in J(x)} w_i(x)$$

where $n(x)$ is the number of people who exist in x.

Although average utilitarianism is very commonly adopted by economists, few have defended it explicitly. So in describing the motivation that lies behind it, I have to speculate a bit. However, I am confident that a part of the motivation is the basic intuition that adding people is not valuable for its own sake. Welfare economists are typically interested in making people as well off as possible. We want to improve the well-being of the people who are alive, but we have no interest in having a lot of people around. We know our aim is not achieved by maximizing the total of well-being in the world. This total can be increased in two ways: by creating new people as well as by improving the well-being of people who exist. Since we are only interested in the latter goal, we do not want to maximize this total. On the other hand, maximizing people's average well-being seems to achieve our aim. It ensures that the people who exist are, on average, as well off as they can be. For this reason, average utilitarianism seems to capture the basic intuition. The popularity of average utilitarianism in economics is, I believe, further evidence of the strong grip this intuition has on economists.

However, if I am right about this motivation for average utilitarianism, average utilitarianism lets us down. I said the typical aim of welfare economists is to improve the well-being of people who exist. That is to say, we want to take the people who exist and make them as well off as possible. Average utilitarianism aims to make the people who exist as well off as possible, but that is not quite the same thing. Just as there are two ways of increasing the total of people's well-being, there are two ways of increasing the average of people's well-being. One is to make existing people better off; the other is to create new people whose well-being will be above the average well-being of the people who already exist. We are only interested in the first of these ways; we only want to make existing people better off. Average utilitarianism, on the other hand, is in favour of adding people to the population if their well-being will pull up the average, even if adding these people brings no benefit to the people who exist. But this contradicts the basic intuition.

I see average utilitarianism as a failed attempt to capture this intuition. There is a lot else wrong with it too. Its faults have been described in many places (there is a survey in Broome, 1992, 117–21), and I am not going to rehearse the rest of them here.

5. Population-relativity

Partha Dasgupta does not accept the basic intuition (see Dasgupta 1994, 118), but he has developed an attractive theory of population (most recently presented in Dasgupta 1994) that can be used to accommodate it. This section explains the theory and how the basic intuition can be fitted into it. The two succeeding sections describe some difficulties in the theory itself.

Dasgupta's theory is relativist in a particular sense. I have been assuming up to now that there is a single betterness relation, but Dasgupta does not believe that. Amartya Sen (1982) suggested at one time that the notion of betterness might be relative to the point of view of particular people. Some option x might be better than y from my point of view, but worse from your point of view. Of course, x might be better for me than y, and worse for you than y, but Sen did not mean that. He was not simply pointing out that different people's interests diverge. He was suggesting that *moral* betterness might differ according to people's points of view. This is the idea that Dasgupta takes up. Specifically, Dasgupta suggests that the betterness relation is relative to a particular population. Betterness from the point of view of one population is different from betterness from the point of view of another. I shall discuss the basis of this idea in Section 6.

So for each population of people J, there is a betterness relation $\succeq^r_J$ relative to J. Dasgupta assumes that $\succeq^r_J$ gives special weight to the well-being of the members of J, and for simplicity he assumes it can be represented by a relative value function v^r_J that is a weighted sum of the well-being of the people who exist

$$v^r_J(x) = \sum_{i\in J(x)\cap J} w_i(x) + t \sum_{i\in J(x)-J} w_i(x)$$

The members of J get a weight of one, and other people a weight of t. $0 \leqslant t \leqslant 1$.

Suppose a population J^* has to make a choice amongst a set of options $S \subset X$. Which should it pick? In this section I shall describe what Dasgupta's theory recommends. I shall examine the grounds for it in Section 7.

The theory recommends a two-stage decision procedure. Stage one goes like this. First, divide the options into classes according to their population, so that each class consists of all the options that have a particular population. For a population J, its class is $S(J) = \{x \in S \mid J(x) = J\}$. Then, from each of these classes, pick the best option in the class according to the relative value function of that class's population. That is to say, pick $\bar{x}(J) = \text{argmax}_{x\in S(J)}\, v^r_J(x)$. Dasgupta does not say what to do if more than one option maximizes the value function in a class, so let us assume this never happens. According to the formula above, $\bar{x}(J)$ will be the option that maximizes the unweighted total well-being of the population J, since all the options in the class contain only this population. Do the same for each class of options, selecting the unweighted maximum for each. Let $\bar{X}$ be the set of these maximizing options, consisting of $\bar{x}(J)$ for each J. That is the end of stage one.

The second stage of the decision process is for the decision-making population J^* to compare the options $\bar{x}(J)$ that were selected at the first stage, and pick the best according to its own relative value function. That is to say, it chooses $\text{argmax}_{x \in \bar{X}}\, v^r_{J^*}(x)$. In this second stage, therefore, population J^* gives more weight to its own well-being than to others'.

How does the two-stage procedure work in our examples? In example 1, suppose the decision-making population, choosing between options a, b, and c, is the first two people: the couple as I fleshed out the example in Section 3. Suppose a population's value function gives zero weight to people outside that population, so $t = 0$. Options a and c form a class with the same population including the child. Therefore, at stage one we choose between these two options, using the value function of the population that includes the child. This is an unweighted function, and it puts a above c. Option b is in a class on its own, so it also is selected at stage one, automatically. Stage two requires the couple to choose between a and b using their own value function. Since this gives no weight to the child, the result is a tie between a and b, so it does not matter which the couple chooses. In sum, they should not choose c, but it does not matter which they choose out of a and b. If, on the other hand, only b and c were available as options, the two-stage procedure would say it does not matter which of the two is chosen. These are exactly the conclusions I claimed in Section 3 to be supported by intuition.

We reached them only by assuming $t = 0$. If a population's betterness relation gave any positive weight to the well-being of people outside the population, the conclusion would be that the parents ought to have a child. This would not conform to the basic intuition.

The results of the two-stage procedure in this example violate property β, one of the standard conditions of rational choice (see Sen, 1969). Although b and c are both in the choice set when the options are b and c, b but not c is in the choice set when the options are expanded to a, b, and c. There is no harm in that; it is precisely what we need. Intuition, which we are trying to capture formally, itself violates the standard conditions. The merit of Dasgupta's two-stage procedure is that it can deliver results in tune with intuition, without implying the logical impossibility of an intransitive betterness relation. All Dasgupta's relative betterness relations are transitive, or course.

In example 2, suppose the decision-making population is the first three people: the parents and first child as the example is fleshed out. Options a and c form a class with the same population. Since c has the greater unweighted total of well-being, it is selected from this class at stage one. Options b and c are then compared at stage two, using a value function weighted towards the first three people. Provided the weight t given to the fourth person is less than a quarter, b emerges as the right choice. In this example, the two-stage procedure delivers a definite answer, whereas in Section 3 I suggested that our intuitions might not be able to cope with this example so well. So it seems that Dasgupta's theory may not only be able to represent our intuitions properly when they are clear, but it may be able to clarify them when they are unclear. The results in this case violate property α (Sen 1969), but once again, this is exactly what we need.

However, the next example reveals a genuine problem:

Example 3

$$w(a') = (4, 4, 6, \Omega, 1, \Omega, \dots)$$

$$w(b) = (4, 4, 5, \Omega, \Omega, \Omega, \dots)$$

$$w(c) = (4, 4, 4, 4, \Omega, \Omega, \dots)$$

Example 3 is the same as example 2, except that a', a permutation of a, replaces a. The second child in c is the same person as the second child in a, but not in a'. Continue to take the decision-makers to be the first three people, and $t < \frac{1}{4}$. In example 2, the two-stage procedure recommends b. In example 3, it recommends a'. So it treats a and a' differently. The reason is that in example 2, the presence of c knocks out a at the first stage of the procedure, whereas in example 3, a' stays in the reckoning because c and a' have different populations. I doubt the procedure is giving the right answer in this case. Since a and a' differ only in the identity of their populations, surely they ought to be treated the same. Parfit (1984, pp. 366–79) argues strongly for the view that the identities of the population should not matter.

Between a' and c, the people making up the population are different, but the number of people is the same. Dasgupta does not himself consider examples with the same number of people but different people, and I am not sure I have correctly described the procedure he would favour in these cases. In stage one, he might want the classes of options to be selected according to the number of the population rather than the identities of the population. For each number m, he might want us to define a class as $S(m) = \{x \in S \mid n(x) = m\}$. If we took this definition, a' and c would fall into the same class. Then c would knock out a' at stage one, just as it knocks out a. Histories a and a' would be treated similarly, as it seems they ought to be. So this approach gives satisfactory results in this example.

However, it does have implications for Dasgupta's relativist notion of value. Originally, in presenting the theory, I suggested that betterness might be relative to the point of view of a person or a group of people. For instance, it might be relative to a population. If we are to take the line I suggested in the previous paragraph, we will have to make betterness relative to a number of people rather than to the specific people making up a population. Number-relative betterness is less plausible than population-relative betterness. I shall say more about this in the next section.

6. The basis of relative betterness

The examples suggest that relativity theory could be used to reproduce the conclusions of intuition, given some suitable assumptions. However, I am doubtful about the foundations of the theory itself. An ethical theory needs to provide a convincing ground for the conclusions it arrives at. Dasgupta's notion of relative goodness needs to be justified, and so does his two-stage decision procedure. When it comes to justification, I am doubtful about both. This section discusses the first, and the next section the second.

Why does Dasgupta think betterness is relative to the population, and why does he think betterness relative to a population gives more weight to the population's own well-being than to other people's? He treats population-relativity as a type of community-relativity. He thinks members of a community have special claims on each other that outsiders do not have. A population forms a community, and from its point of view, people not yet born are outsiders. Like me, Dasgupta uses examples with small numbers of people, and the community he is particularly thinking of is a family. 'Family members', he says, 'have a special claim upon another. Potential persons do not have this claim. "They" are not members of our community' (Dasgupta 1994, 119).

I am not concerned with the idea of community-relativity in general. Let us take it for granted that a community should give more weight to its members' well-being than to outsiders'. But I do not think community-relativity can be used to justify Dasgupta's theory of population.

Take example 2 first. Option *b* is better thân *c* from the point of view of the existing population, and the reason on Dasgupta's account is this. Option *c* has a greater total of well-being, but much of it belongs to the second child, who is an outsider to the existing population. Provided outsiders get a weight less than a quarter, *b* comes out ahead of *c*. But this account misrepresents the real moral considerations in the example. It suggests it is in the interest of the second child to have *c* rather than *b*, but her interest is overridden by the more heavily weighted interest of the first child. However, the truth is not that the second child has an interest in the choice that is outweighed, but that she has no interest in the choice at all. It is not in a person's interest to be born, because being born does not make the person better off than she would otherwise have been. Dasgupta says a 'potential person' does not have the same claim as a family member on other family members. But a potential person has no claim to being created, obviously, since until she is created she does not exist. It is not that she has some claim, but a weaker one. The second child, who might or might not be brought into existence, is not at all like an outsider to a community, who might or might not be allowed to join the community.

In the latest statement of his theory, Dasgupta himself denies it is in the second child's interest to be created. He says that adding a person to the population is a good thing, 'not because the added person's *interests* are served, but because good lives are part of the good' (Dasgupta 1994, 118). But I do not think he is entitled to this denial, because it is inconsistent with his appeal to community-relativity. Community-relativity just is the view that the interests of the community should count more for the community than the interests of outsiders. Dasgupta suggests in this remark that a good life is worth creating, not because it is good for the person who lives it, but because it is somehow good in itself. If that is so, he has not explained why the community of existing people should count its own good above the good of creating people. He says it is because the people who might be created do not have the same claim on the community as members of the community itself. But the reason he offers for creating people is nothing to do with the claims of the people who are created; it is that creating people is good in itself.

Now look again at Examples 2 and 3 together. Histories a and a' ought to be treated similarly. But they will only be treated similarly if betterness relations are made relative to the number of the population, rather than to the set of people making up the population. But whereas a population can be considered a community, a number is not a community. So a community-relative theory of betterness cannot support a number-relative theory.

I conclude that the idea of community-relativity cannot justify Dasgupta's theory of population. If we are to have a relativist theory, it will need some other foundation. I do not know what that could be.

7. The basis of the two-stage procedure

Now I come to question of justifying Dasgupta's two-stage decision procedure. Let us now take it for granted that there is a relative betterness ordering for each population. Given that, when a population is making a decision, what ought it to do? If relative betterness means anything, it should surely reflect what morality requires of the population. So it would be natural to expect that the population ought to do the best it can according to its own betterness ordering. But that is not Dasgupta's theory. His theory says that the betterness orderings of other populations also come into determining what it should do, through the two-stage procedure.

One way this could happen is that one population's betterness ordering could affect the constraints under which another population acts. One population's betterness ordering helps to determine that population's actions, and its actions help to determine what another population ought to do, by constraining this other population's options. By this route, one population's betterness ordering affects what another population ought to do only in so far as it affects what the first population actually does, or would do if it had a choice. One population's ordering has no direct moral bearing on what another ought to do.

Dasgupta seems sometimes to have this sort of constraint in mind, but I do not think his ethical theory fits it very well, and there are indications that he himself thinks one population's betterness ordering has a direct moral influence on what another population ought to do. One indication is that Dasgupta simply does not discuss what each population will do—only what it ought to do. It would be naive to assume each population will act as it ought. Yet it is a population's actual acts that should determine the constraint.

A second indication is obscured by the fact that Dasgupta assumes each population has control over all the options that make up its own class of options: all the options that have it as their population, that is. This is an implausible assumption. For instance, in example 2, if the parents decide to have a second child, they might well be able to fix up the choice between *a* and *c* before the second child is even conceived. However, although Dasgupta assumes the parents do not have this choice, at one place in his writings he does raise the question of what they should choose if they did:

> We have to decide ... whether [the parents] can make binding commitments at the first instance, or whether they cannot. The deeper question however is whether binding commitments by [the parents] are ethically defensible even if they were feasible (Dasgupta 1988, 120).

(The parents' fixing up the choice between *a* and *c* before their second child is conceived, on the basis of their own betterness ordering, would be a 'binding commitment'.) Dasgupta does not explicitly answer his 'deeper question'. But in adopting the assumption that binding commitments are ruled out, he gives the impression that the answer 'no' is one of his reasons for doing so. In any case, 'no' is the most plausible answer. In the example, if the final result of the decision-making process was either *a* or *c*, the population in existence would be the later expanded population, which includes the second child. Surely, therefore, the right choice between *a* and *c* is determined by what is better from the point of view of the expanded population. This surely must determine what is the morally right decision for the existing population to make. That is to say, in the choice between *a* and *c*, it seems that betterness from the point of view of the later population should have moral force over the existing population. It is not simply that the potential decisions of the later population constrain the options of the earlier one.

This seems particularly pausible when we bear in mind that two populations normally share many of the same members. In the example, the parents belong to both populations: the original three and the expanded four. So what determines what the parents ought to do: betterness relative to the existing population, or relative to the expanded population? Should they pursue the first sort of betterness up to the moment the second child is conceived, and then start pursuing the second? This would mean that, up to the moment of conception, they should do their best to bring about *a* rather than *c*, and after that moment they should do their best to bring about *c* rather than *a*. That seems implausible.

It certainly seems that if the existing population had a choice between *a* and *c*, it ought to choose *c*. I think Dasgupta would agree. Yet, according to its relative betterness ordering, *a* is better than *c*. So what it ought to do here seems to be determined directly by another population's betterness relation, rather than its own.

This is puzzling. If it is so, what does it mean to say *a* is better than *c* relative to the existing population? Certainly, *a* is more in the existing population's interest than *c*, but that is not in question. Dasgupta's relative betterness relations are not meant to express the interests of particular populations; if they were, they would give no weight at all to other people's well-being. Yet we have just seen they do not tell us what the population ought morally to do, either, since this population ought to choose *c* over *a* even though *a* is better according to its betterness ordering. So I do not know what Dasgupta's idea of relative betterness really amounts to. It would make sense if each population ought to do the best it can according to its own betterness ordering. But since the two-stage procedure implies that is not so, the idea of relative betterness is mysterious.

I conclude that Dasgupta's relativist theory is not at present well founded. If it is to be a successful response to the basic intuition, it needs better justification.

8. Critical-level utilitarianism

Critical-level utilitarianism was originally proposed within economics by Charles Blackorby and David Donaldson (1984). It has recently been developed by the same two authors together with Walter Bossert (1995). It has this value function

$$v(x) = \sum_{i \in J(x)} (w_i(x) - \alpha)$$

where α is the critical level of well-being. For each person who exists, take the amount (positive or negative) by which her well-being exceeds the critical level, and then add up all these amounts. The aim of critical-level utilitarianism is to maximize the total.

Adding a person to the population increases v if her well-being will be above α and decreases v if it will be below α. Only if her well-being will be exactly α does adding a person leave v unchanged. So critical-level utilitarianism is neutral about adding a person only at the critical level. If therefore conflicts with the basic intuition, which is always neutral about adding a person, provided her life will be good.

The conflict with intuition is very severe. If a person might be created with a well-being above the critical level, the critical-level theory says it is better to create her than not. It is strictly better, which means it would be worth some sacrifice on the part of existing people to bring this person into the world: there is some small reduction in existing people's well-being such that it would be better to create the new person and have existing people suffer the reduction. But this is just what our basic intuition opposes. Creating a person seems not to be valuable in itself and would not merit sacrifices on the part of existing people. Perhaps if the new person would be supremely happy we might modify our intuition: we might accept that bringing a supremely happy person into existence would be worth some sacrifices by the rest of us. So the idea of a critical level might possibly be reconciled with intuition this way, provided the critical level represented a very high level of life.

However, if the criticial level is high, there is an opposite difficulty. Critical-level utilitarianism claims that if a person might be created with a well-being below her critical level, then it is better not to create her. Once again, this means it would be worth some sacrifice on the part of existing people to prevent this person's creation. If the critical level is high enough to represent a modestly good life, this, too, is very much against our intuition. True, we would be opposed to creating a person whose life would be bad, and would be willing to make sacrifices to avoid that happening. But if a person's life would be modestly good, we should surely not be positively opposed to her existence, and make sacrifices to prevent it.

In sum, if a person's critical level is anything less than a very good life, there is a serious conflict with intuition on one side, but if it is anything more than a modestly good life, there is a serious conflict on the other side. I think this is a sufficient reason to reject the critical-level theory. There cannot be a critical level such that adding a person above it is good and adding a person below it is bad.

However, the theory does have a solid argument on its side. Take any distribution $w(x) = (w_1, w_2, \ldots, \Omega, \ldots)$ with Ω in the ith place. For a variable u, let $y(u)$ be the history whose distribution $w(y(u))$ is the same as $w(x)$, except that u appears in the ith place instead of Ω. In moving from x to $y(u)$, person i is added to the population at a level of well-being u. Provided the betterness ordering is complete—an assumption I shall discuss in a moment—there must be one and only one value of u for which x and $y(u)$ are equally good. Here is why. First, it is most implausible that x is better than $y(u)$ for every value of u, or that x is worse than $y(u)$ for every value of u. So, given that the betterness ordering is complete, x and $y(u)$ must be equally good for some value of u. However, they cannot be equally good for more than one value of u. Suppose they were; suppose $x \approx y(\bar{u})$ and also $x \approx y(\tilde{u})$, where $\bar{u} > \tilde{u}$. Then by the transitivity of the betterness relation, $y(\bar{u}) \approx y(\tilde{u})$. But $y(\bar{u})$ and $y(\tilde{u})$ have exactly the same distribution of well-being, except that person i has more well-being in $y(\bar{u})$ than in $y(\tilde{u})$. By the principle of personal good, $y(\bar{u}) \succ y(\tilde{u})$, which contradicts that $y(\bar{u}) \approx y(\tilde{u})$.

The unique value of u is a critical level: adding a person above this level is good and adding one below it is bad. For all we know at this stage, the critical level may vary with x and i. Most theories of the value of population imply the existence of a critical level that may vary. Average utilitarianism, for instance, implies a critical level for adding any new person that is equal to the average well-being of the people who already exist. Blackorby and Donaldson's (1984) innovation was to insist that the critical level is a constant, independent of the present population and its well-being. This rules out average utilitarianism, amongst other theories. The chief purpose of Blackorby *et al.*'s new paper (1995) is to present new arguments in support of this claim.

I have no need to pursue the new arguments, because the very existence of a critical level, whether constant or not, already comes up against the severe conflict with intuition that I mentioned. I said the conflict was enough to show there could not be a critical level. This means we must reject the argument I have just given, because it implies a critical level exists. There is only one lacuna in the argument: it assumes the betterness ordering among histories is complete. We must reject that assumption.

9. Critical-band utilitarianism

In a more recent paper, Blackorby *et al.* (1996) have exlored this route. Taking up a suggestion of Derek Parfit's (1984, 430–2), they have extended critical-level utilitarianism by dropping the assumption of completeness. I shall call the extended theory 'critical-band utilitarianism'. It assumes there is a critical band or interval $[\bar{\alpha}, \bar{\alpha}]$ of well-beings, where $\bar{\alpha} > \bar{\alpha}$, rather than a single critical level. It says one history x is better than another y if and only if

$$\sum_{i \in J(x)} (w_i(x) - \alpha) > \sum_{i \in J(y)} (w_i(y) - \alpha)$$

for every value of α in the band $[\bar{\alpha}, \bar{\alpha}]$. If neither x is better than y, nor y better than x, then x and y are 'not ranked'. In particular, x and y are not equally good. The idea is that betterness is not a fully determinate relation, and sometimes there is no determinate answer to the question which of two options is better than which.

According to the critical-band theory, adding a person with a well-being above the band is good and adding one with a well-being below the band is bad. Adding a person with a well-being within the band is neutral. However, 'neutral' has to be understood in a new way. When adding a person is neutral, up to now I have assumed that means the history where she is added is equally as good as the one where she is not. But in the new theory, it means that these histories are not ranked against each other.

Unlike the relation 'equally as good as', the relation 'not ranked against' is not necessarily transitive. Take Example 1, for instance, and suppose the critical band includes both levels of well-being 1 and 2. Then according to the critical-band theory, *a* is not ranked against *b*, and *b* is not ranked against *c*, but *a* is ranked against *c*—it is better than *c*. By understanding neutrality as non-ranking rather than equality, the theory can avoid a contradiction whilst preserving the basic intuition that adding a person is neutral. The critical band may be as wide as we like. For instance, it may extend from infinity down to lowest level of well-being that is not definitely bad. In that case, the theory says that adding a person is neutral provided only that her life is not bad, just as the basic intuition suggests.

So on the face of it, critical-band theory fits the intuition well. However, it has its problems. One is that it owes us an explanation of the indeterminacy in the betterness relation. Just what does it mean to say that one option is not ranked against another? It does not mean they are equally good, but what alternative is there? For instance, what ought one to do if faced with a choice between options that are not ranked? If they were equally good, it would not matter which is chosen, so one could simply choose randomly. But if they are not ranked, presumably choosing randomly would not be appropriate. Until the meaning is properly explained, critical-band theory can be suspected of cheating. To say that adding a person is equally as good as not adding her leads to contradiction, because 'equally as good as' is constrained by transitivity. So instead the theory says these alternatives are 'not ranked', and that frees it from the constraint. But wriggling out of transitivity this way seems like cheating unless a proper explanation is provided.

This worry about neutrality reveals itself concretely in something I call 'the package problem'. Take this example:

Example 4

$$w(a) = (1, 1, \Omega, \Omega, \Omega, \ldots)$$

$$w(b) = (1, 1, 3, -1, \Omega, \ldots)$$

Here the question is whether it is good or bad to add two people as a package. The first of the two will have a good life, so the basic intuition is that adding her is neutral. However, the second will have a bad life (if we assume a well-being below zero is bad), so our intuition is that it is bad to add this second person. The package, then, consists of one neutral addition and one bad addition. This suggests it is a bad package: *b* is worse than *a*.

If we assume the critical band extends from zero to infinity, it is easy to check that critical-band utilitarianism says *a* and *b* are not ranked. If the first additional person were added on her own, at level 3, that would be neutral (not ranked). Adding the second, at level -1, would be bad. But adding the two together is supposed to be neutral. The theory, then, allows a neutral addition to cancel out a bad addition, to produce a neutral package. This is strongly against intuition. It could not happen if neutrality were simply equality of goodness, as we originally assumed. It results from this new unexplained sort of neutrality.

I am not sure critical-band utilitarianism should really be blamed for the package problem. It may be a further difficulty in the basic intuition itself. (David Donaldson made this point to me.) Compare a third history *c*, where

$$w(c) = (1, 1, 1, 1, \Omega, \ldots)$$

This has the same population as *b*. How does it compare with *b* in goodness? For simplicity, let us assume that histories with the same population should be compared on straightforward utilitarian grounds. Since *b* and *c* have the same total of well-being, they are equally good, therefore. Now let us compare them both with *a*. The basic intuition implies that *b* is worse then *a* for the reason I explained: moving from *a* to *b* involves one bad addition and one neutral one. The intuition also implies that *c* is not worse than *a* because moving from *a* to *c* involves two neutral additions to *a*. So *b* is worse than *a* and *c* is not worse than *a*, which contradicts our earlier conclusion that *b* and *c* are equally good. This contradiction arises directly from the intuition. Critical-band utilitarianism avoids it by denying that *c* is worse than *a*. It may be unfair to criticize this conclusion for being counterintuitive, since our intuition is itself self-contradictory here.

10. Conclusion

That is as far as I shall pursue the dialogue between theory and intuition. In this paper, I have examined three theories about the value of population, to see how well they can accommodate the basic intuition. (I have examined a few more in Broome 1994, and there is a very comprehensive examination of many others in Parfit 1984, Part IV.) Average utilitarianism fails badly. Relativist utilitarianism is attractive but has some foundational problems of its own. I think critical-band utilitarianism holds the best hope for progress. Its claim that betterness is not fully determinate in the area of population seems plausible to me, but it needs to give a proper account of the nature of this indeterminacy. Even critical-band utilitarianism cannot fully accommodate the basic intuition. I am sure this intuition will have to be modified, at least, if it is to fit into a coherent theory of the value of population.

ACKNOWLEDGEMENTS

This paper was presented at the Latin American Meeting of the Econometric Society, Tucuman, August 1993. I have had many useful discussions with many people on the subject of this paper, particularly at a series of seminars I led at the Center for Applied Ethics at the University of British Columbia in 1993. I must particularly thank Charles Blackorby, Partha Dasgupta, and David Donaldson for their comments on this paper, and Derek Parfit for a great deal of instruction on population theory generally.

REFERENCES

Arthur, W B (1981). 'The economics of risk to life', *American Economic Review*, **71**, 54–64.

Blackorby, C., Bossert, W., and Donaldson, D. (1995). 'Intertemporal population ethics. critical-level utilitarian principles', *Econometrica*, **65**, 1303–20.

Blackorby, C., Bossert, W., and Donaldson, D. (1996). 'Quasi-orderings and population ethics', *Social Choice and Welfare*, in press.

Blackorby, C. and Donaldson, D. (1984). 'Social criteria for evaluating population change', *Journal of Public Economics*, **25**, 13–33.

Broome, J. (1992). *Counting the Cost of Global Warming*, White Horse Press, Cambridge.

Broome, J. (1994). 'The value of a person', *Proceedings of the Aristotelian Society*, Supplementary Volume, **68**, 167–85.

Dasgupta, P. (1988). 'Lives and well-being', *Social Choice and Welfare*, **5**, 103–26.

Dasgupta, P. (1994). 'Savings and fertility: ethical issues', *Philosophy and Public Affairs*, **23**, 99–127

Narveson, J. (1967). 'Utilitarianism and new generations', *Mind*, **76**, 62–72.

Parfit, D. (1984). *Reasons and Persons*, Oxford University Press, Oxford.

Sen, A. (1969). 'Quasi-transitivity, rational choice and collective decisions', *Review of Economic Studies*, **36**, 381–93.

Sen, A. (1982). 'Rights and agency', *Philosophy and Public Affairs*, **11**, 3–38.