The Philosophical Review, Vol. XCVIII, No. 4 (October 1989)

SHOULD ETHICS BE MORE IMPERSONAL? A CRITICAL NOTICE OF DEREK PARFIT, *REASONS AND PERSONS**

Robert Merrihew Adams

When Derek Parfit comes to articulate "one common feature" of most of the manifold conclusions of his book, he says,

> I have argued that, in various ways, our reasons for acting should become *more impersonal*. Greater impersonality may seem threatening. But it would often be better for everyone" (p. 443).

As this statement intimates, *Reasons and Persons* is an ambitious book, morally as well as theoretically, by an avowed "revisionist" in philosophy (p. x). The work is animated by a manifest concern to find rationally warranted changes in our beliefs that will help us to deal in a morally and humanly more satisfactory way with such problems as the inevitability of our own deaths and the impact of our actions on future generations. It is a book that most philosophers in the English-speaking world, and many others with an educated interest in moral and social problems, will want to read —not only because of its boldness and its wealth of ingenious and fascinating examples and puzzles, but also because of the author's architectonic sensibility, which forces us, time and again, to think about issues in a new way.

The present reviewer disagrees strongly with Parfit's case for greater impersonality; and much of this article will be devoted to criticism of it. This is particularly true of Sections I and II, which are about Parfit's claim that Common-Sense Morality "fails in its own terms," and about the practical implications he draws from his Reductionism about personal identity. In Section III, I will argue that some of his ethical paradoxes about the further future can be solved in ways that he does not consider. These are among the

*Derek Parfit, *Reasons and Persons*. Oxford, England: Clarendon Press of Oxford University Press, 1984. Pp. xv, 543. Parenthetical page or section references in the text are to this book.

main topics of Parts One, Three, and Four, respectively, of *Reasons and Persons*.

The refutation of the Self-Interest Theory (S), which is not only the central topic of Part Two, but a major unifying theme that runs through the first three parts of the book, will not receive such extended discussion here; but something must be said about it at the outset. I agree that S is wrong, and have no interest in defending it against Parfit's critique. S is a theory of practical rationality. Its "central claim" is

> (S1) For each person, there is one supremely rational ultimate aim: that his life go, for him, as well as possible (p. 4).

By an "ultimate" (as opposed to an "instrumental") aim Parfit means a goal that is sought for its own sake and not merely as a means to some other end.

Parfit argues persuasively (in Section 46) for the controversial assumption that ultimate aims, in this sense, can be irrational, or more and less rational. He gives us less help, I think, in seeing whether there are intuitively plausible and interestingly general criteria for assessing the rationality of such aims. His first, and what he entitles "the best objection to the Self-Interest Theory," is so simple an appeal to intuition that he has more recently commented that it "hardly deserves the name" of "argument."[1] It is merely that there are several ultimate aims or desires (for example, for achievement, and the good of other people) that are (as we judge intuitively) "no less rational than the bias in one's own favor," and in pursuit of them it can be rational to do what is worse for oneself (pp. 131–133, 192). Like other reviewers, I find this argument, simple as it is, convincing, and indeed "the best" of Parfit's objections to S.

His more elaborate arguments against S, in Part Two, seem to me much less persuasive. They have to do with *time*. The concern for one's own good demanded by S, as Parfit interprets it, is temporally neutral, at least as regards the future. For this reason S condemns as irrational the very common "bias towards the near," in which one is willing to accept somewhat lesser pleasures, if they

[1] Derek Parfit, "Comments," *Ethics* 96 (1986), p. 843n.

will happen sooner, and somewhat greater pains, if they will happen later. Parfit uses this to show that S allows some types of relation but not others to have rational importance, and he argues that S should be rejected because no good argument can be given to justify this asymmetry. Much that is interesting and illuminating emerges in the course of his argument; but I do not see why those who find the Self-Interest Theory intuitively plausible (as Parfit and I do not) should admit that they stand in any special need of argumentative support here for granting to one sort of relation a rational importance that they deny to another. If they did need it, why wouldn't Parfit need an argument (which he does not give) for granting (as he does) a rational importance to one type of sensation (pain) that he would not grant to others (such as, perhaps, the smell of eucalyptus)? Of course he thinks that the latter asymmetry is obviously right in a way that S's asymmetries are not; but that just carries us back to the appeal to intuition.

I. What Must a Moral Theory Do to Succeed?

Part One of *Reasons and Persons* is about "Self-Defeating Theories." Parfit thinks of moral theories and theories of rationality as "giving" us "aims." For instance, S gives me the aim "that [my] life go, for [me], as well as possible" (p. 4). He classifies a theory, T, as *self-defeating* if following, or trying to follow, T is not the most effective way of achieving "T-given aims." The self-defeat is *direct* if it results from actually doing what T recommends; *indirect* if it results rather from *trying* to follow T, or from a desire or disposition to do so. The self-defeat is *individual* if individuals' T-given aims are frustrated through individual adherence to T; *collective*, if everyone's T-given aims would be frustrated if everyone adhered to T. The two most important kinds of self-defeat for our discussion are defined by Parfit as follows. He says T is

> *indirectly individually self-defeating* when it is true that, if someone tries to achieve his T-given aims, these aims will be, on the whole, worse achieved (p. 5);
> *directly collectively self-defeating* when it is certain that, if we all successfully follow T, we will thereby cause the T-given aims *of each* to be worse achieved than they would have been if none of us had successfully followed T (p. 55).

Parfit compares S with Consequentialism (C), according to which the "one ultimate moral aim" is "that outcomes be as good as possible" (p. 24), and with Common-Sense Morality (M). Each of these three theories, he argues, is self-defeating in at least one of the four possible ways, but only M thereby "fails in its own terms." He holds that S is indirectly individually self-defeating, because it is often worse for individuals if they are disposed to be never self-denying, and that probably C is indirectly collectively self-defeating, because everyone's having the dispositions of "pure do-gooders" would likely have worse results than our having some Non-Consequentialist dispositions (for example, "not to kill, even when we believe that this would make the outcome better" (p. 28)). But S and C do not thereby fail in their own terms, because such failure results only from direct, not from indirect, self-defeat, according to Parfit.

There are cases (for example, "Prisoner's Dilemmas") in which S is directly collectively self-defeating. And Parfit contends that there are analogous cases in which Common-Sense Morality (M) is directly collectively self-defeating. Though direct, this sort of self-defeat does not show that S fails in its own terms. For it is of the essence of a Prisoner's Dilemma that I will do better for myself in it by making the self-interested choice, whatever the others do, because it is not my own altruism but the altruism of others that would benefit me (p. 89). Here it would be better for me if others did not follow S. But that is no embarrassment for the S-Theorist, who (like an investment adviser or a coach in a competitive sport) makes no promise that her advice to me will work out better, from my point of view, if *other people* follow it too. For "S is not a collective code. It is a theory about individual rationality" (p. 92). Parfit argues, however, that M cannot get off the hook in this way, because M, unlike S, is a moral theory, and "morality is essentially a collective code" (p. 106). Therefore, he holds, M does fail in its own terms, and must be revised, if not rejected.

I shall argue that M has not been shown to be directly collectively self-defeating, and that in any event such self-defeat would not amount to failure in M's own terms. I shall also argue that indirect self-defeat at the individual level is more damaging to S than Parfit maintains; and that some features which C might need to have in view of its indirect self-defeat at the collective level, and which Parfit thinks acceptable, are really more unacceptable in a moral theory than direct collective self-defeat. I begin with the point about S.

Parfit distinguishes two ways in which a theory, T, could be indirectly individually self-defeating. The first way is by the agent's trying but failing to do what T recommends. If T is "too difficult to follow," my T-given aims might be better achieved if I pursued some other goal instead of trying to achieve them. "But this is not true of S," Parfit claims (p. 5).

This claim is not obviously correct. It does seem difficult for most of us to know what will be best for us. It is not difficult to recognize some things that are obviously bad for us, or obviously necessary for us. But S demands much more discrimination than that. Might there not be, say, a moral code (L) that guided us to our own good more reliably than our own judgment would? Perhaps Parfit would reply that in that case we could (and according to S, should) try to do what is best for us *by* following that code. This might be psychologically impossible, however, because there might be circumstances in which it would be psychologically impossible for us to believe that following L would be better for us than doing something else—even though, in fact, in most such circumstances, following L would be better for us. If we have reason to believe that something of this sort is true, that would seem to me to be a potentially serious criticism of S.

The other way in which a theory can be indirectly individually self-defeating is through the agent's dispositions rather than through his or her actions; and Parfit speculates that this is true of S for most people most of the time (p. 7). Suppose I always try to do the action that will be best for me, and always succeed in doing so. I will not make this strenuous attempt unless I have the disposition never to be self-denying; and it is a thesis (S6) of S that this is "the supremely rational disposition" (p. 8). But having this disposition is likely not to be best for me. It may be incompatible with other desires that would contribute much to my happiness—strong desires for achievement, for example, or for the good of others (p. 6). As Butler remarked, "that character we call selfish is not the most promising for happiness."[2] And if I have the disposition to be

[2] Joseph Butler, *Fifteen Sermons Preached at the Rolls Chapel,* ed. T. A. Roberts (London, England: SPCK, 1970), p. 102 (Sermon 11, par. 9).

never self-denying, I am unlikely to be completely successful in concealing this from others, and the consequent loss of their trust may cost me dearly (cf. p. 7). (Given that this disposition is likely not to be best for me, it follows of course, according to S, that it would be rational for me to try not to have it.)

Even though it would be worse for many people if they had the disposition S praises as supremely rational, Parfit argues, "S does not fail in its own terms." Why not? "[T]his bad effect," he says, "is not the result either of their doing what S tells them to do, or of their having a disposition that S tells them to have. Since this is so, S is not failing in its own terms" (p. 11). This seems to imply that a theory T fails in its own terms only if doing what T *tells* one to do leads to frustration of one's T-given aims.

The most obvious difficulty in applying this criterion to S is that S does not explicitly *tell* anyone to do anything. Parfit's formulations of S are statements about what is rational or irrational. And they give criteria of (supreme) rationality not only for actions, but also for aims, desires, and dispositions. Taking all those criteria together, it is misleading to say that "S does *not* tell these people to be never self-denying, and it tells them, if they can, *not* to be" (p. 11). It would be much more accurate to say that S tells us that one would be most rational in *having* the disposition and in *trying* not to have it.

This suggests an argument for the conclusion that S does fail in its own terms. S is a theory of rationality. It tells us (among other things),

> (S6) The supremely rational disposition is that of someone who is never self-denying (p. 8).

Parfit grants it would be better for some people if they did not have this disposition. In other words, he grants there are people whose S-given aims would be worse fulfilled if they are supremely rational according to S (in one of the respects in which it is possible to be supremely rational). And their S-given aims would be worse fulfilled *because of* their S-certified supreme rationality. Why doesn't that show that, as a theory of rationality, S fails in its own terms?

Consider the case in which it is not psychologically possible for these people to cause themselves not to have a never self-denying disposition. This is an important case for S because it is the one in which these people will have the disposition that is supremely rational according to S even if they act in the most rational way according to S (that is, even if they try not to have it). Thus they will be as rational as they could be according to S. And precisely because of that they will achieve their S-given aims less well than they might if they were less rational (with respect to their disposition) according to S. This would seem to tell strongly against S, if leading to failure to achieve theory-given aims matters to the success of a theory.

I suspect Parfit would reply that the success or failure of a theory in its own terms depends exclusively on the consequences of the *actions* it commends. This is a claim that deserves to be controversial. It is not only actions but also (as Parfit recognizes) desires and dispositions that we evaluate as to their rationality and morality; and we cannot safely assume that only the evaluation of actions is of practical importance.

We might ask, for example, why we should care about S or any other theory of rationality. More precisely, why should we care what ultimate aim, if any, is supremely rational? That this is not an idle question can be attested by any who can remember attempting, in an existential crisis, to reason themselves into caring about anything at all. But suppose I do in fact care about being rational, and therefore care about the claims of S; and suppose further I am persuaded of the truth of S. Now I discover that this commits me to the conclusion that it would be rational to try not to have the most rational sort of disposition. May this not undermine the aspiration for rationality that led me to be interested in S in the first place? And is that an unimportant failing for S?

Let us turn to Parfit's argument that Common-Sense Morality (M) fails in its own terms. He does not offer any comprehensive formulation of M, presumably because common sense is sensitive to such a variety of independent moral considerations that a complete statement of M would have to be unmanageably complex. What he does tell us about M is that it tells each of us that we ought to give some priority (though not an absolute priority) to the interests of "the people to whom we stand in certain relations—such as our children, parents, friends, benefactors, pupils, patients, clients, colleagues, members of our own trade union, those whom we represent, or our fellow citizens" (p. 95)—our "M-related people," for short.

Given this feature of M, it is possible to devise analogues of the Prisoner's Dilemma in which the good of each agent's M-related people takes the place of the agent's own good. A good example is Parfit's Fisherman's Dilemma, in which we are poor fishermen and, because of overfishing and declining stocks of fish,

> it is true of each that if he does not restrict his catch, this will be slightly better for his own children. They will be slightly better fed. This is so whatever others do. But if none of us restricts his catch this will be much worse for all our children than if we all restrict our catches (p. 100).

In this case, Parfit argues, M is collectively self-defeating; for if all of us follow M, giving priority to the interests of our M-related people, those interests, which are our M-given aims in this situation,[3] will be advanced less well than if we had all followed a different policy. And this self-defeat is direct; for the bad effect results from *actions* in which we successfully follow M. Because the defeat is direct rather than indirect, and because M, as a moral theory, must be "a collective code," Parfit concludes that M fails in its own terms.

He proposes that M should therefore be revised. The most radical revision would convert M into an *agent-neutral* theory. Parfit rightly observes that problems like the Fisherman's Dilemma arise for theories that are *agent-relative*, in the sense that they assign different aims to different agents. N is the agent-neutral theory into

[3] It is an over-simplification to treat these, as Parfit does, as our only M-given aims in the situation, a simplification made inevitable, perhaps, by the (inevitably) incomplete presentation of M. One complication, neglected in Parfit's discussion, which might affect the soundness of his argument, is that an aim may owe its status as M-given to another aim more fundamentally given by M. Thus if I am a lawyer, M gives me the interests of my client as an aim that is specially mine; but that is largely because M gives me more fundamentally the aim of keeping faith with my client. M also gives me such aims as respecting my client's autonomy. It is therefore far from obvious that my M-given aims regarding my client will always be better achieved if my client's interests are better advanced, or even if they are better advanced *by me*. The achievement of my M-given aims may well depend on *how* I advance my client's interests.

which M can be transformed by saying that "each of us should always try to do what will best achieve everyone's M-given aims" (p. 108)—that is, I take it, the aims that had been assigned to each by M, counted in accordance with a principle of aggregation not mentioned here by Parfit. How oppressively conformist a unanimity of purpose would be required under N as the price of avoiding competition and conflict is not spelled out. In any event Parfit acknowledges that his argument cannot compel believers in M to move all the way to N (pp. 109f.).

He argues instead that M-believers ought rationally to be converted at least to a more modestly revised theory, R. R holds in effect that we ought to do the more impartially benevolent thing, rather than giving priority to our own M-related people, *if* we are in a situation where at least *k*-1 others will make the impartial choice if we do, and where everyone's M-related people will be benefited more if at least *k* of us make the impartial choice than if each gives priority to her own M-related people (pp. 100–103). According to Parfit, this revision moves some way from M towards an agent-neutral Consequentialism (C), and C's indirect self-defeat shows that C in turn (without being revised) gives its adherents reason to foster some dispositions that are more characteristic of M than of C. He concludes that C and M need to be enlarged and revised "in ways that bring them closer together" (p. 112).

It may be doubted how far Parfit's proposals would move M toward C. Arthur Kuflik, in a fine paper, has argued, in effect, that Common-Sense Morality already includes what R is supposed to add to M. Common-Sense Morality needs no revision to tell Parfit's fishermen that they certainly ought to make a concerted effort to restrict their catch, if they can; and it tells each of them that it would be unfair, and hence immoral, for him to make his children free riders by fishing without restraint when enough others are restricting their catch.[4] Parfit now grants this point.[5]

Even if R were a real revision of M, moreover, it would not touch the cases about which M and C disagree most deeply. In the course of arguing for R, Parfit claims that "[t]here are countless cases where, if each gives priority to his M-related people, this

[4] Arthur Kuflik, "A Defense of Common Sense Morality," *Ethics* 96 (1986), pp. 784–803; especially pp. 801f.

[5] Derek Parfit, "Comments," *Ethics* 96 (1986), pp. 852f.

would be worse for all these people than if no one gave priority to his M-related people" (p. 102). What are these countless cases? The condition supposedly satisfied in them is a very strong one: that *all* the affected people would be better off under one condition than under the other. This seems to be true in the case of Parfit's fishermen, but is probably not true in most of the cases in which it is morally plausible to give priority to one's M-related people. Where it is true, there is no deep conflict between personal loyalties and impartial benevolence; for when all are benefiting more from general cooperation than they would from unbridled competition, none can reasonably feel betrayed by their special protectors. These facts make it easier to believe that R is already included in M—and harder to see this as bringing M much closer to C.

The difference between moral partiality and impartiality is deeper, more intractable, and of greater importance in those (probably numerous) cases in which, if we all succeed in following M, the result of our giving priority to our M-related people will be that the M-related people of some of us will be better off, and the M-related people of others of us will be worse off, than if we all had successfully followed C or N. In these cases M is *not* directly collectively self-defeating in Parfit's sense. If the results of M in such cases are sufficiently bad from the point of view of fairness or utility or both, that may be reason to revise M; but Parfit's argument is not cast in such a form as to show that M needs revision to deal with such cases.[6]

For these reasons I think Parfit has not established that M is directly collectively self-defeating. And even if he could establish that, there are at least two grounds on which it may be doubted whether it follows that M fails in its own terms. One is that Parfit's argument may be vitiated by consequentialist presuppositions.[7] Why would bad consequences of acting on a theory tend to show that the theory fails in its own terms? The answer is easy if the theory is consequentialist; but M is pretty clearly not a consequen-

[6] As he acknowledges, in effect, in Section 40, where he pulls back from arguing for N.

[7] As Lanning Sowden suggests in his review of *Reasons and Persons, Philosophical Quarterly* 36 (1986), pp. 514–535; see p. 527.

tialist theory. The M-Theorist's obvious response is, "We always knew that acting on moral principle might cost some 'utility'."

Parfit will doubtless reply that he does not conceive of self-defeat in consequentialist terms. He explains self-defeat in terms of non-achievement or inferior achievement of the "aims" endorsed by a theory. And his description of theories in terms of their aims is explicitly meant to avoid consequentialist presuppositions. His "use of aim is broad."

> It can describe moral duties that are concerned, not with moral goals, but with rights, or duties. Suppose that, on some theory, five kinds of act are totally forbidden. This theory gives to each of us the aim that he never acts in these five ways (p. 3).

And he offers us an analogue of the Fisherman's Dilemma in which I must choose between giving my child some benefit, and enabling you to give your child a greater benefit; and you face a symmetrical choice. Here, Parfit argues, if each has the M-given aim that *he* benefits *his* child, both of us will better achieve our M-given aims if both choose to enable the other (p. 97).[8]

This is not enough to escape consequentialist presuppositions. Describing my aim as "that I benefit my child," or "that I never act" in a certain way is still describing it as an outcome, and one which you could take as an aim of yours with reference to me. Many non-consequentialists (Kantians, for example) will deem it important to have aims that are not outcomes but actions, describable as "to benefit my child" and "not to act" in that certain way. These you could not intelligibly take as aims of yours, with reference to my action.

This may be a verbal difference, but it corresponds conveniently to a difference that is important for moral theory, the difference between acting on a principle and trying to bring about the optimization of one's whole future course of action from the point of view of the principle.[9] We can say that I have doing the best for my patients as an *action-aim* insofar as I am disposed to do (now) what I think is best for my (present) patients. I have it as an *outcome-aim*

[8] Cf. Sowden, *op. cit.*, pp. 527ff.

[9] I am much indebted to John Giuliano for help in understanding this distinction.

insofar as I am disposed to try (now) to bring it about that I do (in the rest of my career) the best for my (present and future) patients.[10] If I have it as an action- but not an outcome-aim, or if the action-aim takes precedence, I will be disposed to do what I think is best for the patient I am treating now even if I foresee that that will prevent me from making the money to buy equipment that would enable me to do even better for other patients in the future. And that is what many non-consequentialists would say I morally ought to do, if the stakes are high enough for my present patient. Given that they self-consciously take this sort of stand, why should they think their theory fails in its own terms if following it will not result in the best outcomes on the whole, from their point of view?

The other ground for doubting whether M must fail on its own terms if it is directly collectively self-defeating is that it is not clear that M is, as Parfit claims, "a collective code." What is a collective code? Parfit says, "Call a theory . . . *collective* if it claims success at the collective level" (p. 92). If this is a definition, it will be analytic that collective theories fail in their own terms if they are directly collectively self-defeating. As usual, however, analyticity comes at a price. Non-consequentialist moral theories need not be collective on this account, if I am right in thinking that they need not claim "success" at all.

Waiving this point, we may observe that the apparent definition seems tailor-made for theories like Kantianism (p. 92) and Rule Utilitarianism (RU), which subject policies to the test, "What if everybody did the same?" Lanning Sowden has made the interesting point that this test does not have the importance for Act Utilitarianism (AU) that it has for RU. For AU may be seen as requiring "that an individual maximize social utility subject to several constraints the most important of which is that he regard the strategies of all other agents as given or constant."[11] Given this

[10] I do not know whether this distinction has anything in common with Parfit's distinction between "formal" and "substantive" aims (p. 3).

[11] Sowden, *op. cit.*, p. 526. Sowden goes on to say, "Roughly, RU tells me to select an action on the assumption that we all perform that action, whereas AU does not. Thus RU, but not AU, is a collective code in Parfit's sense." This may be too swift. I imagine Parfit would reply that while AU often commends individual "strategies" (Sowden's term) that would not be successful if they were followed (as they will not be) by everyone, AU's one fundamental principle of maximizing utility would (necessarily) be suc-

feature of AU, one might argue that AU's acceptibility does not depend on whether AU's aims would be achieved if everyone always acted in accordance with AU—since there is manifestly no danger that such utilitarian virtue will actually become universal.

Similar considerations apply to M. Parfit says that when deciding on a moral theory, "we should first consider our Ideal Act Theory," which says "what we should all ideally do, when we know that we shall all succeed." Such a theory is "ideal" rather than "practical," because it is a fact that "[w]e are often uncertain what the effects of our acts will be" (and often mistaken about such matters, I would add), and "some of us will act wrongly" (pp. 99f.). In saying that a moral theory "claims success at the collective level," Parfit seems to mean that the theory is committed to the thesis that its aims would be better achieved by universal following of its Ideal Act Theory than by universal following of any other Ideal Act Theory. This strikes me as an implausibly romantic constraint on moral theories. Why should it be an objection to a moral theory that, if universally followed with perfect success (as it will not be), it would yield somewhat worse results than would be obtained by everyone following with perfect success an alternative theory *that certainly will not and probably could not be so followed by all?* What is the relevance of this impossible alternative?

In discussing why S is "not a collective code," Parfit says something that suggests a different understanding of "collective."

> Suppose that we are choosing what code of conduct will be publicly encouraged, and taught in schools. S would here tell us to vote against itself. If we are choosing a collective code, the self-interested choice would be some version of morality (p. 92).

If a collective code is simply a code that is meant to be publicly adopted, in the sense of being publicly encouraged, inculcated as part of moral education, and widely practiced, then it is plausible to hold that moral theories must be collective codes. So if it could

cessful if universally followed. Even so, as I argue in the text, it is hard to see why this collective success would matter to the tenability of AU. It is interesting also to note that Parfit himself distinguishes Consequentialism, as a moral theory that is "*individualistic* and concerned with *actual* effects," from what he calls "Collective Consequentialism," which "is both *collective* and concerned with *ideal* effects" (p. 30).

be shown that such public adoption of M would result in a worse outcome, from the point of the concerns that support M, than some alternative that really could be adopted in this way, that might be a serious criticism of M. But this could not be established by proving that M is directly collectively self-defeating in Parfit's sense, because the results of public adoption of a moral theory do not coincide with the results of universal perfect compliance with it. (I am enough of a believer in original sin to suspect that they are not even very similar.)

The thesis that a moral theory must be collective in the sense of being meant to be publicly adopted is actually denied by Parfit (though without using the word "collective") in connection with Consequentialism (C), which claims,

> (C3) If someone does what he believes will make the outcome worse, he is acting wrongly (p. 24).

Should Consequentialists think that C ought to be publicly adopted? Sidgwick famously inclined to the negative on this question.[12] Parfit inclines to the affirmative, but argues that the tenability of C need not depend on this answer. Consequentialists should prefer to avoid such dependence, because there is at least some reason to think that in view of the indirect self-defeat of C the public adoption of some other code would lead to a better outcome than the public adoption of C.

Parfit calls a theory "self-effacing" if it implies that one ought to try to bring it about that it is not believed. He notes that on some views, "a moral theory cannot be self-effacing," but "must fulfill what Rawls calls 'the publicity condition'; it must be a theory that everyone ought to accept, and publicly acknowledge to each other." Parfit claims, however, that this view is tenable only for those who "regard morality as a social product." "If a moral theory can be quite straightforwardly *true,* it is clear that, if it is self-effacing, this does not show that it cannot be true" (p. 43). Thus he seems to imply that requiring moral theories to satisfy the publicity condition commits one to some sort of subjectivism or anti-realism. But that is surely wrong.

[12] Henry Sidgwick, *The Methods of Ethics,* seventh edition (London, England: Macmillan, 1907), p. 490, cited by Parfit (p. 41).

The publicity condition does connect morality with society; for it says that a moral theory, as such, must be meant to be publicly adopted. By their very meaning, it may be argued, moral claims have implications about how certain social practices ought to be related to the types of behavior discussed in the claims. Part of what is meant by saying that a certain type of conduct is morally wrong is that it ought in general to be publicly discouraged as wrong.[13] In this moral claims differ, no doubt, from scientific and mathematical claims. A proposition of nuclear physics or molecular biology can be objectively true even if the danger of abuse is so great that it ought not to be divulged. But that is because the rightness or wrongness of publicizing them is extraneous to the content of scientific statements. If the publicity condition is not extraneous to the content of moral claims, it is hard to see how this compromises their objectivity, since it has not been shown that it cannot be "straightforwardly true" that a type of conduct ought to be publicly discouraged as wrong.

I shall mention a second point at which Parfit seems to me to offer an unconvincing defense of a feature that Consequentialism might need to have in order to avoid bad consequences of Consequentialist dispositions (and thus to deal with its indirect self-defeat). This second point has to do with blame and remorse. Parfit holds that Consequentialists should think that there are cases in which a morally wrong action should not be an object of blame and remorse because it flows from a disposition that is morally advantageous and hence should not be discouraged. He comments, "Consequentialism does not in general break the link between the belief that an act is wrong, and blame and remorse. This link is broken only in special cases," such as "those in which someone acts on a motive that it would be wrong for him to cause himself to lose" (p. 35).

This apology is not convincing. Morality doubtless allows for mitigation of blame and remorse when a wrong action is done from a good motive. But if we say that an action, for which the agent is fully responsible, ought not to be an object of blame or

[13] This is not to say that the opinion that it is *occasionally* right, in special circumstances, to lie about moral principles is logically inadmissible. To affirm a principle of conduct, however, while denying that it ought *in general* to be inculcated, is not to affirm it as a moral principle.

remorse at all, what can we mean in calling the action "morally wrong"? The obvious Consequentialist answer is that we can mean that from the point of view of achieving the ends that are important to morality, it would have been more advantageous not to have done the action. This answer makes a Consequentialist account of moral wrongness true by definition. It is not a plausible definition. There is an important difference between saying that an action was not likely to result in the best outcome from a moral point of view, and saying that it was morally wrong. And no small part of the difference is in what the charge of moral wrongness implies about the appropriateness of blame and remorse.

These are not trivial problems about Consequentialism. If C is self-effacing, or if it has the implications Parfit admits it to have regarding blame and remorse, that is a reason for thinking it is not really a theory about *moral* right and wrong. It is a much more serious failure for a proposal in moral theory, in my opinion, than it would be for Common-Sense Morality to have less than optimal consequences at the level of Ideal Act Theory.

II. THE IMPORTANCE OF PERSONAL IDENTITY FOR ETHICS

In Part Three of *Reasons and Persons* Parfit defends a Reductionistic conception of persons and their identity through time. Using a fascinating array of examples (of a predominantly science fiction character), he tries to show that "what matters" practically, where we care about our transtemporal identity, is not identity as such, but the relations of psychological connectedness and continuity in which, he thinks, it mainly consists. He argues that this has important implications for morality and practical rationality. It provides a final argument against the Self-Interest Theory, whose insistence that rationality requires equal concern for all periods of our future life is "defeated" by the consideration that our farther future will be much less connected, psychologically, to our present than our nearer future will be. In addition to other possible implications for morality, Parfit claims that Reductionism increases the plausibility of views, like Utilitarianism, for which distributive considerations have no intrinsic moral importance. His reason for this is that facts of personal identity and non-identity, on which distributive principles "are often held to be founded," are less "deep," metaphysically, on the Reductionist view, and therefore it "becomes more plausible to be more concerned about the quality of experiences, and less concerned about whose experiences they are" (p. 346). I shall not enter here into discussion of Parfit's arguments for Reductionism,[14] as I prefer to focus on the practical inferences he draws from it.

Let us begin with his claim that if personal identity is less "deep," it becomes more plausible to care less about it. I think he means it becomes more plausible to think it irrational to care as much as we do about personal identity.[15] This is in some ways a puzzling claim. It is not obvious what it means to say that personal identity is less "deep" on the Reductionist view,[16] or what is the connection between metaphysical depth and practical importance. Perhaps the argument that Parfit proposes is that our valuing personal identity as most of us do rests on a belief about its nature that is mistaken if Reductionism is right.

That would not be a very good argument, if it means that we explicitly infer the value of personal identity from Non-Reductionist beliefs about its nature. For (with the possible exception of a few philosophers) who does that? And who needs to? We care *who* does what, and what happens to *whom*, because we care in a special way about ourselves and about people we love. And who needs reasons for that? Parfit agrees that love need not be based on reasons.[17] Few would suppose that our special concern for ourselves and our own future needs reasons any more than our love for other individuals does.

Can anything be said in defense of the rationality of the way we ordinarily care about personal identity? Here we do well "to consider the way in which our identification of and concern for ourselves and each other as *persons* essentially contributes to, if you'll pardon the expression, our form of life," as Susan Wolf urges in a

[14] They are the subject of Sydney Shoemaker's illuminating review of *Reasons and Persons* in *Mind* 94 (1985), pp. 443–453.

[15] More recently he has claimed to have argued that our "car[ing] a great deal about personal identity . . . is irrational" (Parfit, "Comments," p. 833).

[16] Parfit has conceded ("Comments," p. 838) the justice of a charge of vagueness at this point, made by Bart Schultz, "Persons, Selves, and Utilitarianism," *Ethics* 96 (1986), pp. 721–745; p. 732.

[17] Parfit, "Comments," p. 834, responding to Susan Wolf, who was making essentially the same point I am making here.

richly rewarding essay about this aspect of Parfit's work.[18] I take this remark not as an appeal to a peculiarly Wittgensteinian philosophy of language, but as a claim that the way we care about persons is and should be affected by the deep embedding of the concept of personal identity in a complex web of social practices.

One of these practices is child-rearing. According to Parfit it would be rational to care less about one's own farther future if it will be only weakly connected, psychologically, to one's present, because such connectedness is a major part of "what matters" in personal identity. If this applies to self-interest, it would seem to apply also to one's concern for the futures of people one loves. But small children, as Wolf points out, are only weakly connected, psychologically, to the adults they will become. So if we cared little about the farther, weakly connected futures of people we love, she argues, love would not motivate parents to discipline children for the sake of their adult development. "Why should a parent reduce the happiness of the child she loves so much for the sake of an adult she loves so little?"[19] In fact, however, it is a normal part of our practice of child-rearing that one takes the whole life of one's child as a project[20] about which one cares greatly.

I regard my own life in that way too. No doubt I learned that from my parents, whose project it was before it was mine. Had I been neglected as a child, I might have found it natural to live more "for the moment," and might never have learned to "postpone gratification." Having adopted my own (whole) life as a project, however, I have access to further projects and practices that would otherwise be inaccessible to me. I can enter into long-term commitments, such as marriage. I can apply for a thirty-year mortgage. I can undertake a scholarly project that will take twenty years to complete. I can care about the moral significance and consistency of my life as a whole. I can aspire to grow and change in ways that I cannot fully foresee but that will surely involve some loss of psychological connectedness. I am committed to my future

[18] Susan Wolf, "Self-Interest and Interest in Selves," *Ethics* 96 (1986), pp. 704–720; p. 708.

[19] Wolf, *op. cit.*, p. 711.

[20] In this use of "project" I am mindful of, but probably not in complete conformity with, John Perry's use of it in "The Importance of Being Identical," in Amélie Oksenberg Rorty, ed., *The Identities of Persons* (Berkeley, Calif.: University of California Press, 1976), pp. 67–90.

without regard to variations of connectedness within the range that normally occurs in human life, though a sufficiently radical loss of connectedness might place a possible future outside the bounds of my project altogether.[21]

Much of what we care most about in human lives cannot, of its very nature, be found in an experience or a short period of life. Should Reductionists care more about experiences, and less about persons? "If the reason we care about persons is that persons are able to live interesting, admirable, and rewarding lives," Wolf argues,"we may answer that time slices of persons, much less experiences of time slices, are incapable of living lives at all."[22] For this reason, indeed, I think that most of our caring about the quality of experiences, except perhaps those that we are having right now, is due to our caring about the people to whom they belong. Many of Parfit's arguments seem to presuppose that our concern for the quality of experiences would, or rationally should, continue undi-

[21] Parfit considers a position like this on the practical importance of psychological connectedness, but rejects it in favor of the belief that such connectedness is one of those "relations which can be rationally believed to be less important when they hold to reduced degrees." He states without argument that this belief *"cannot be defensibly denied"* (p. 314). I think common sense would suppose that I have as much reason from the connectedness between my present state and my state a week hence to care about what will happen to me then as I have from the even greater connectedness that obtains from day to day to care about what will happen to me tomorrow. I do not really wish, however, to dispute Parfit's rejection of S's contention that it would be *irrational* to be less concerned about my more distant, and less connected, future. What I want to say is that I *can* reasonably have my life as a project in which my level of concern for my future does not vary with any normally expected variation in psychological connectedness.

[22] Wolf, *op. cit.*, p. 709. Thus we have a *reason* for caring about persons that does not apply to experiences. Parfit alludes to this argument in a footnote ("Comments," p. 833), but does not respond to it. He focuses on Wolf's claim that the reason for caring about persons as such "is that life, or, if one prefers, the world, is better that way" (Wolf, *op. cit.*, p. 713). His main response is that "if some desire has good effects, this fact cannot show that this desire is rational; it can at most show that we have a reason to try to have, or to keep, this desire." "Whatever the effects," for example, "it would be irrational not to care about future Tuesdays" (Parfit, "Comments," pp. 832f.). With this rejoinder in mind, I am trying, in the text, to address the issue of rationality (though I believe the relation between the salutariness of a motivational pattern and its rationality is probably quite complex).

minished if we became less interested in whole lives as such; this seems to me very questionable.

Another way in which our conception of personal identity is woven into our form of life is that it marks the boundaries of the past and the future that *belong* to me. My life belongs to me retrospectively, in the sense that I am responsible for it; prospectively, in the sense that it is, in a special way, mine to shape. This belonging has a subjective and voluntary aspect: I *take* responsibility for my past (and in a different way for my future); I have *intentions* for my future. It has a social and voluntary aspect which is at least as important: I am *held* responsible for my past; "It's your life," others say with regard to many decisions I make or can make about my future. There is also a normative aspect to this belonging: I have a *right* to a certain control over my future; I *ought* to be held responsible, and accept responsibility, for certain things.

The rationality of caring about personal identity in this complex network of ways, which I have only begun to sketch, is established, *within* a form of life to which they belong, by our finding that they *make sense*. The concepts involved in these projects "work." We are able to interpret our lives in terms of them. Using them, we can commonly make plausible judgments, which often enough seem to us illuminating. There is nothing we care more about than some of the projects that are inextricably intertwined with our conception of personal identity. We can commonly pursue these projects with some hope of success, but it would hardly make sense to pursue them at all if we did not care about our lives as wholes in the way we normally do. These considerations do not show that it would be *ir*rational to adopt a radically different form of life in which we would not think about our lives in this way, if we could do so (or even imagine doing so, in any detail). But they establish a strong presumption that we are not irrational in continuing to treat the whole lives of particular individuals as specially important projects, and to regard our pasts and futures as belonging to us in a special way.[23]

[23] This is only a presumption. I have not proved that it could not be overridden. If we came to believe, for example, that there is no causal influence, direct or indirect, of earlier on later stages of the same person, then, perhaps, it would be irrational for us to care as we do about personal identity. The point of invoking a presumption in this context is that we are

None of this implies that there are not imaginable circumstances in which it would become irrational (cease to make sense) to care about personal identity in these accustomed ways. To draw on Parfit's examples, it is imaginable that incidents of fusion and fission of persons (or at any rate, events empirically indistinguishable from such) would often happen. In that context it might well not make sense to regard pasts and futures as belonging to us in the way that we now do. If we value our present form of life (as most of us do), we have reason to prevent such incidents if we can. Parfit sometimes seems to be arguing on the assumption that if we care about identity in our actual circumstances in a way that we would not if certain science fictions became reality, we ought to have a quite general and theoretically interesting rationale for the difference. In view of the subtle complexity of human forms of life, I think it is quite unrealistic to expect that we should have such a rationale for the differences in what concerns would seem reasonable to us in vastly different physical, and especially social, environments.[24]

So the value we set on personal identity needs no justification from Non-Reductionist arguments. But Parfit may think it rests on Non-Reductionist beliefs in a less explicit way. He might argue that

considering whether Parfit can support his position by arguing that the value we are accustomed to set on personal identity must be justified by appeal to Non-Reductionist beliefs. And the answer to this question is, "No, we need no such Non-Reductionist justification," if the sort of presumption I have described favors the belief that our valuation is rational.

[24] One might try to construct a different sort of argument from Parfit's cases of fission and fusion, using them first to try (as Parfit does) to establish that it is psychological connectedness and continuity, rather than personal identity, that it would be reasonable for us to care about in these cases where they may be supposed to diverge. This in turn might be taken as some reason to think that in ordinary cases, where (according to this argument) they coincide, it is the connectedness and continuity, rather than personal identity as such, that matters to us. This argument, however, proposes a reinterpretation rather than a revision of our interest in our lives. It is crucial to Parfit's revisionist project that he sees the interest in continuity and connectedness as diverging in the actual case too from the interest in identity as such, inasmuch as connectedness, unlike identity, comes in degrees. What I argue in the text is that science fiction examples provide dubious support, at best, for the revisionist project, because there is little reason to suppose that what it is rational to care about under actual circumstances must be the same as what we think it would be reasonable for us to care about under the very different imaginary circumstances.

if we became Reductionists, and reflected adequately on the significance of Reductionism, we would find that personal identity no longer seemed so important to us. He testifies that something like that has happened to him since he became a Reductionist; and he thinks it is a good thing. Because facts of personal identity and non-identity seem less important, "I am less concerned about the rest of my own life, and more concerned about the lives of others"; and "my death seems to me less bad" (p. 281).[25]

Not all who become Reductionists will react in this way, however. Wolf does not, for one. "Parfit has convinced me of reductionism with respect to persons," she says. "But I find that this conviction does not lessen the degree of my interest in persons a bit."[26] If Parfit charges that Wolf's caring about personal identity rests on a belief in Non-Reductionism, she will reply that she does not hold that belief. He could argue that if she had reflected adequately on Reductionism, she would no longer care as much about personal identity, and that therefore the degree of her caring about it still rests on the Non-Reductionism she no longer holds. But that kind of attack on opposing intuitions would not settle anything, as it is a game that any number can play.

This does not show that Parfit could not reasonably have a deep conviction that Reductionism warrants diminished concern for facts of personal identity and non-identity. It might be a religious conviction. Parfit calls attention to the fact that Buddha was a Reductionist about persons and their identity (p. 273). And Buddhist texts do use Reductionistic metaphysical arguments about identity in an effort to weaken such attitudes as self-centeredness and anxiety about death.[27] But that is not the only possible broadly reli-

[25] Parfit might not want to rely on this argument. He reports that he is still "much more concerned" about his own future than he "would be about the future of a mere stranger." And he thinks that this would be generally true of Reductionists. In saying this, however, he distinguishes sharply between the questions whether Reductionists would have this attitude and whether it would be rationally justified (p. 308). This suggests that the claims he discusses about what it would be rational for Reductionists to care about are not meant to depend on what Reductionists would in fact care about.

[26] Wolf, *op. cit.*, p. 705.

[27] This is far from the whole story about Buddhism, of course. The third of its Four Noble Truths is that cessation of suffering comes from cessation of craving; and the craving that must cease is "the craving for sensual

gious response to Reductionistic conclusions. I doubt that Kierkegaard was committed to metaphysical Reductionism about personal identity. But he certainly agreed with Parfit that being the same self over time in a way that is morally and humanly significant requires psychological connectedness. In fact he thought it requires repentant taking of responsibility for one's past, followed by constantly repeated affirmation of (the same) ethical or religious commitment. The persistence of significant selfhood is assured not by ontology but by will-power.[28] But this certainly did not lead Kierkegaard to think personal identity less important. On the contrary, it was for him the most precious of all achievements. It would be hard to think of anyone who cared more about it.[29]

Besides claiming that the lesser metaphysical "depth" ascribed to personal identity by Reductionism makes it more plausible to care less to *whom* things happen, Parfit offers arguments for other theses about morality, of which the most important, perhaps, are about *desert* and *guilt* and about *compensation.* Parfit views with increasing favor arguments for the Extreme Claims "that, if the Reductionist View is true, we cannot deserve to be punished for our crimes," and cannot be compensated at any time for anything that happens at another time. In *Reasons and Persons* he wrote that both the Extreme Claims and their denials are defensible (pp. 324f.,

pleasure," the concern about the quality of present experience, paradigmatically rational for Parfit, that seeks "satisfaction now here now there," just as much as "the craving for continuing existence." And this cessation is promised as the fruit, not of mere philosophical reflection, but of persistent following of the Noble Eightfold Path of moral and spiritual discipline, at the end of which one may hope for an *enlightenment* that seems to be held out as something rather more interesting and enticing than merely coming not to care so much about one's own inevitable death. I quote from *Majjhima-nikaya,* iii. 248–252, as translated in Sarvepalli Radhakrishnan and Charles Moore, eds., *A Source Book in Indian Philosophy* (Princeton, N.J.: Princeton University Press, 1957), pp. 276f. An alternative translation can be found in Edward Conze's collection of *Buddhist Scriptures* (London, England: Penguin Books, 1959), pp. 186f.

[28] At this point Parfit (p. 446) may come close to Kierkegaard.

[29] See Søren Kierkegaard, *Concluding Unscientific Postscript,* trans. David F. Swenson and Walter Lowrie (Princeton, N.J.: Princeton University Press, 1941), for example, pp. 152–158, 276–282; and *Either/Or,* vol. 2, trans. Walter Lowrie, revised by Howard A. Johnson (Garden City, N.J.: Doubleday Anchor, 1959). Much of the discussion of "existence" in the *Postscript* should be classified by analytic philosophers as being about personal identity through time.

342f.); in more recent "Comments" he doubts that they can be defensibly denied.[30]

I will discuss first the argument he suggests in the book for the Extreme Claim about desert and guilt. The argument for the Extreme Claim about compensation (pp. 342f.) is similar, and my response to it would be similar. The arguments are based on a case of "my imagined division," in which my brain has been divided into left and right halves, which are transplanted into the bodies of my two brothers, resulting in two persons, Lefty and Righty, somewhat similar, physically, to each other and to me as I was, and strongly connected, psychologically, with my presurgical state. In this case, Parfit claims, Non-Reductionists must say it will be true either that I am Lefty, or that I am Righty, or that I am neither.

Suppose I am Righty. On that assumption, Parfit thinks a Non-Reductionist could defensibly deny that Lefty would "deserve to be punished for the crimes that I committed before the division. . . . How can he deserve to be punished for crimes that someone else committed, at a time when he himself did not exist? Only the deep further fact of personal identity [as the Non-Reductionist believes it to be] carries with it responsibility for past crimes." From this he infers that a Non-Reductionist, if converted to Reductionism (and hence any Reductionist as well), can defensibly deny that we ever have desert or guilt, since (according to Reductionism) the "deep further fact" is never present (pp. 324f.).

Non sequitur. For the fact of personal identity, whether deep or not, is assuredly present in the cases in which we ordinarily assign desert and guilt. Suppose it is true that when we think that personal identity requires something more than physical and psychological continuity, it seems reasonable to deny desert and guilt where that something more is not present. How can that fact show that it is defensible to deny desert and guilt where only (unduplicated) physical and psychological continuity are present when we think that is all that is required for personal identity?

Parfit seems to be assuming that whether we count a condition as sufficient for personal identity cannot affect whether it is reasonable to regard it as sufficient for assigning desert and guilt. This assumption disregards the role of the concept of personal identity in our form of life. An important part of the concept's role

[30] Parfit, "Comments," p. 843, n. 26.

is juridical (a point that is implicit in Locke's pioneering discussion of personal identity[31]). Part of what we decide when we decide that a certain condition is sufficient for personal identity is that it will be sufficient for a past and/or future to *belong* to me in a way that is linked with such concepts as those of responsibility, desert, guilt, and compensation.[32] So when we shift from thinking physical and/ or psychological continuity insufficient for personal identity to thinking it sufficient for personal identity, why shouldn't we be expected also to shift from regarding it as insufficient to regarding it as sufficient for desert and guilt?

To be sure, Parfit has argued in Sections 90–91 that identity is not "what matters when I divide." But his arguments there are not clearly relevant to questions about desert and guilt. The question mainly addressed there is whether a development that preserves Relation R (as he calls psychological continuity and connectedness) but not personal identity would be (as Parfit claims) "about as good as ordinary survival" (p. 264). And he answers that for Reductionists, "what fundamentally matters is whether I shall be R-related to at least one future person" (p. 268). Let us grant him, for the sake of argument, that he has described imaginable cases of division in which Relation R would be preserved but identity would not; and that such division, if not as good as ordinary survival, would at least be much better than annihilation. This gives him a way in which identity, as such, matters much less than Relation R. But it certainly does not follow that personal identity does not matter much for questions of desert and guilt, which are quite different from questions about the satisfactoriness of alternatives to straightforward survival.

[31] John Locke, *An Essay Concerning Human Understanding,* II, xxvii, 18–19, 26.

[32] Perhaps the link between the concept of personal identity and these other concepts could be broken by a sufficiently radical change in our beliefs about causality. This would entail a significant change in our form of life, and I am not sure how well the concept of personal identity would retain its own identity through the change. But I cannot see that a compelling reason has been given for thinking that the conceptual link would be broken by a change from a Non-Reductionist to a Reductionist theory of personal identity. Indeed the view I am sketching in the text is particularly well adapted to Reductionism, inasmuch as the juridical role of the concept of personal identity can be seen as a constraint on the form of an acceptable reduction or construction of the concept.

It is also part of Parfit's argument in Sections 90–91 that all that keeps identity from being preserved in cases of my division, on a Reductionist view, is the fact that there is a plurality of fairly equally matched contenders for the status of being me. He argues that this fact is too *trivial,* and too *extrinsic* to the relation between successive person stages, to determine "what matters." In relation to juridical issues, however, it is not a trivial fact at all. Our practices of assigning rights and responsibilities to persons rest on the assumption that personal pasts and futures will have at most one "owner" at any time. Since some of these rights and responsibilities are not easily or conveniently divisible, the existence of two evenly matched claimants is decidedly an important fact. Similarly, if we are using the concept of personal identity to give shape to decisions about such matters as desert and guilt, it is hard to see how a reasonable objection could arise from the fact that identity involves a uniqueness condition that is "extrinsic" to the relation between successive person-stages. These juridical issues are undeniably social; why shouldn't they depend on relations between persons as well as between person-stages?

Parfit's arguments for the Extreme Claims in his more recent "Comments" are in one way harder to dismiss, because he begins, not with what it may plausibly be claimed we *would* say if we were Non-Reductionists, but with what we *are* intuitively inclined to say about another of his examples. Again I will discuss the argument for the Extreme Claim about desert and guilt; but the argument for the Extreme Claim about compensation[33] is similar, using the same example, and my response to it would be similar. In teletransportation my brain and body are destroyed here on Earth, while "the exact state of all my cells" is recorded and beamed by radio to Mars, where the Replicator "create[s], out of new matter, a brain and body exactly like mine" (p. 199). Relation R is perfectly preserved in teletransportation. The Branch Line Case is a variant of teletransportation in which, for reasons too complicated to explain here, the Scanner on Earth does not destroy my body but damages it, so that it will die of cardiac failure within a few days. Though Parfit thinks we might reasonably decide to say that in simple teletransportation *I* get to Mars, he believes we

[33] Parfit, "Comments," pp. 839–842.

clearly must describe the Branch Line Case as one in which I remain on Earth, doomed to early death, while a replica of me ("Backup") begins a more promising life on the red planet (pp. 201f.).

"Suppose that, in the Branch Line Case, I had earlier committed some crime." Could Backup justly be punished for it? "Most of us," Parfit claims, would say no. We would hold that because Backup would not be me, he would not be guilty, or even responsible, for my crime, even though Backup would be as strongly connected and continuous, psychologically, with my earlier history as we normally are with ourselves from day to day. On the Reductionist View, Parfit claims:

> Backup is not me only because . . . these [psychological] connections do not have their normal cause: the continued existence of my brain. Is it the absence of this normal cause which makes Backup innocent? Most of us would answer no. We would think him innocent because he is not me.

Parfit goes on to argue that "[t]his reply would show that we are not Reductionists" and believe that only the Non-Reductionistic "further fact" of identity "carries with it desert and guilt," and that therefore if we become Reductionists, we ought to conclude that "[n]o one ever deserves to be punished for anything they did."[34]

This is an outrageous argument. When we say that Backup is innocent because he is not me, we are not committing ourselves to any metaphysical view about the nature or grounds of the non-identity; we are simply saying that it is the non-identity, and not its grounds, whatever they may be, that is the ground of the innocence. If asked whether it is because of the metaphysical irreducibility of personal identity that I am guilty of my own offenses "most of us," I think, would equally answer no—that I am guilty of them simply because I am still myself.

Moreover it is extremely plausible to think that normality of the causal basis of action is relevant to questions of moral responsibility. Philosophical discussions of free will have emphasized this point. If I act under the influence of drugs, hypnosis, electrodes in my brain, or even insanity, my responsibility for the action is di-

[34] Parfit, "Comments," pp. 838f.

minished or eliminated. We often explain this by saying that "I was not myself" when I did it. We are concerned with quite a different sort of abnormality in Parfit's science fiction examples, of course; but there is no clear reason why it should not "matter" for moral responsibility, especially when it supports a judgment of non-identity.[35]

Both versions of Parfit's arguments for the Extreme Claims fail, at bottom, for the same reason. He argues that the Non-Reductionist "deep further fact" of identity is intuitively required for moral responsibility and for the possibility of compensation. But all that his examples will support is that personal identity, whatever its metaphysical basis may be, is required. He begins with intuitions that I think draw their power (though he might deny it) from the difference between these examples and normal, clear cases of personal identity, and then tries to use these intuitions to talk us into assimilating the normal to the abnormal case. But we cannot pull ourselves up by our own intuitive bootstraps.

Positive arguments also can be offered against the Extreme Claims. The analogy between the nature of persons and the nature of nations, to which Parfit appeals, not only fails to help his argument for the unimportance of personal identity,[36] but actually tells against the Extreme Claims about desert, guilt, and compensation. For hardly anyone holds a Non-Reductionist view about nations or other institutions, but most of us think that they can be deserving, guilty, and compensated. Our legal system reflects this belief. Perhaps it is erroneous; but if so, the error cannot be explained by (mistaken) belief in a Non-Reductionist theory of nations or institutions, for we hold no such belief. This consideration adds to the implausibility of holding that it would be unreasonable for Reductionists to believe that persons can be deserving, guilty, or compensated.

There may also be a reasonable *ad hominem* objection to the Extreme Claim about compensation, at least as it applies to benefits

[35] Parfit argues that "we cannot rationally . . . claim that [it] matters much" whether Relation R has its normal cause. But in saying this he seems to be thinking mainly about what matters for the question whether teletransportation is as bad as annihilation. And I cannot see much more in his argument than a bare appeal to the intuition that "[i]t is the *effect* which matters" here (p. 286).

[36] As Parfit concedes to Wolf in his "Comments," p. 837, n. 14.

received *after* the burden is borne. It is plausible to suppose that *I* can be personally compensated for present burdens by a benefit enjoyed by someone existing at a future time *t* if the existence of that person at *t* will be about as good, for *me*, as my surviving to *t*. But then, since Parfit claims that the existence at *t* of someone R-related to my present person-stage is about as good, for me, as surviving to *t* (pp. 263f.), he must also grant that I can be compensated by benefits to such a person.

Parfit says an Extreme Claim about *commitments*, that "we can never be bound by past commitments" if Reductionism is true, can be defended by arguments similar to those offered for the other Extreme Claims (p. 326). He also argues that one might become so weakly connected, psychologically, to one's earlier self as to be unable to release another person from a commitment made to that earlier self. He builds this argument on an example that involves no science fiction, The Nineteenth-Century Russian. A young nobleman of socialist ideals plans to give to his peasants the vast estates he will inherit in several years. Knowing that he might lose his ideals, however, he "signs a legal document, which will automatically give away the land, and which can be revoked only with his wife's consent." And he obtains from her a promise that she will never give her consent, even if he changes his mind and asks for it. He says,

> I regard my ideals as essential to me. If I lose these ideals, I want you to think that I cease to exist. I want you to regard your husband then, not as me, the man who asks you for this promise, but only as his corrupted later self. Promise me that you would not do what he asks.

"This plea," as Parfit says, "seems both understandable and natural." He also claims that if the Russian nobleman in middle age did ask his wife "to revoke the document, she might plausibly regard herself as not released from her commitment" (p. 327).

I agree she might plausibly think she had some sort of obligation not to sign the revocation, but I am not convinced by Parfit's analysis in terms of a commitment from which she cannot be released because the "self" to whom she made it no longer exists. The example turns on at least two factors extraneous to Parfit's theories: the moral value of the contemplated actions, and the wife's judgment of their value. A pair of variations on the story may help us to see the importance of this point.

Case (*A*): Suppose the direction of change reversed, and suppose the wife to change with the husband. That is, suppose that when young they are arch-conservatives, and hear with horror and disgust of other landowners freeing their serfs and giving land to them. The husband therefore executes a document that would require his wife's signature for any gift of land, and gets her to promise that if he should ever be corrupted by the spread of liberalism and wish to give land to his peasants, she will refuse to sign. In this connection he says things about selfhood similar to those that Parfit's young liberal landowner says. Some years later, however, both the landowner and his wife have been won over to a more liberal persuasion, and wish to give land to the peasants. Will she or should she feel that she must not sign because she is bound by a promise to her husband's earlier self that his present self cannot revoke? No; I think she probably will not, and certainly should not, feel that. If that is right, it shows that if in Parfit's case there is some obligation from which the husband cannot now release the wife, it is not because an earlier self no longer exists. For there is as much reason in Case (A) to think the husband no longer the same self, but in Case (A) either there is no obligation or the husband is still able to release his wife from it.

There are two alternative explanations for this reaction to Case (A), by which Parfit might try to defend himself against my argument. (i) He might claim that even in Case (A) the husband cannot release his wife from a promise made to an earlier self, but that the wife rightly regards the promise as void because it was an evil promise. (ii) Alternatively he might claim that in Case (A) the husband cannot release from the promise made to his earlier self, but that the wife is not bound by it because she too is changed and cannot be bound by a promise made by her earlier self. I can respond to both of these replies with a single variation on my counterexample.

Case (*B*) is like Case (A) except that the wife's promise (not to consent to a gift of land to the peasants) was made, not to her husband, but to her late father-in-law, a generous man to his family and friends, but extremely conservative politically, who bequeathed the land to his son. In this case, it seems to me, she probably would and should feel the force of an obligation from which her husband cannot release her—especially if she feels that her father-in-law is someone who deserves their gratitude. It may be that, having become more liberal, she would and should conclude that the moral desirability of giving the land to the peasants is great enough to override this obligation. But at least there seems to be something to override—whereas my intuition about Case (A), in which the promise was to her husband (or her husband's earlier self) is that there is no obligation there to be overridden, once her husband has made his present preference clear. This tells against both the replies I suggested for Parfit; for (i) the original promise was as bad in Case (B) as in Case (A), and (ii) the wife is as changed in Case (B) as in Case (A), but there is an obligation remaining in Case (B) as there is not in Case (A).

This suggests to me a different view of Parfit's example. I think our sympathy with his verdict on it depends heavily on our sympathy (and the wife's continuing sympathy) with the husband's earlier ideals. The consideration that does and perhaps should hold her back from signing the paper, in Parfit's case, is less like ordinary promise-keeping than like a solemn vow. The appropriate reason for her refusing to sign, and for her thinking that her husband cannot release her from her commitment, is not that he is not the same person, but that signing would be playing false to something that is important to the significance of her and his individual and shared lives—more important than her respecting his wishes now. I doubt that Reductionism about personal identity has any implications for the morally binding character of commitments.

III. Future Generations

The fourth and final part of *Reasons and Persons* seems likely to provide, for some years to come, one of the main frameworks for discussion of ethical issues about future generations. It is linked, somewhat loosely, to the rest of the book by the fact that it begins with a point about personal identity. No child not conceived by my parents within a month of my actual conception would have been me (p. 352). (Whether any such child *could*, metaphysically, have been me, is another and more controversial question.) Any public policy or feature of collective behavior sufficient to have a globally significant effect on the environment in which future human generations will live will have a pervasive impact on who conceives children with whom, and when. After at most three centuries, Parfit (p. 361) suggests—after much less time, I would guess—

there will be no one living who would have existed if our policy or practice were significantly different. If none of them is so miserable that it would be better for them if they had never existed, it follows that after that period of time there will be no one living for whom it would have been better if we had followed a different policy or practice—no one who will have been harmed, on the whole, by what we have actually done.

Should we conclude that we have no moral obligation to restrain, for example, our consumption of natural resources, for the benefit of such distant generations (so long as we can be confident that their lives will be at least minimally worth living)?[37] Such a permissive conclusion, as Parfit heartily agrees, seems obviously wrong. But what sort of ethical theory will provide a home for our rejection of it? Our need to answer this sort of question Parfit calls "the Non-Identity Problem" (p. 363).

This problem shows, as Parfit argues, that our obligations regarding future generations cannot be adequately accounted for by his first attempt at formulating a "person-affecting" ethical principle of beneficence (V): "It is bad if people are affected for the worse" (p. 370). At the end of a very interesting and clarifying discussion, he is able to formulate principles of beneficence that are recognizably person-affecting (concerned with benefits or harms to individual persons as such) and that do solve the Non-Identity Problem, on the assumption[38] that a person can be benefited by being caused to exist (Section 136). But there is little practical difference between these principles and more impersonally stated principles about the sum, or average level, of human well-being.

It is worth noting, because Parfit singles it out as a main contribution of Part Four to the impersonality theme of the book, that he also argues that we should not invoke principle V even when it could apply, when we are thinking about beneficence that will have its effect in the near future. He proposes the example of Two Medical Programs. One would screen pregnant women for condition J. By curing J doctors could prevent the children to be born

[37] In practice this is a major proviso which could not be ignored. But assuming that it is satisfied will help us to focus on the issues that Parfit wishes to discuss.

[38] Which seems defensible to Parfit, obviously correct to me.

from having handicap H (which impairs life but leaves it well worth living). The other program would screen for condition K in women intending to conceive. By waiting to conceive until condition K has disappeared, after two months, women can avoid having children with handicap H. It is true of the first program, but not of the second, that identifiable individuals will be worse off if it is not adopted, because the children who would be born healthy because of it would be the same persons as children who would otherwise have had handicap H. Is that a morally good reason for favoring the first program, if only one of them can be funded? Parfit thinks that intuitively it is not.

I am not sure that I agree with him about that. But if we do, that shows at most that our intuitions about beneficence to persons yet unborn are not person-affecting.[39] I doubt that many of us will have similar intuitions about beneficence to persons already clearly recognizable as part of our moral community. Varying the example, let's suppose that handicap H does not become evident until after puberty, and that condition J is not a condition of the pregnant mother, but a disease of prepubescent children, which, if untreated, causes them later to have handicap H, but which has no harmful effects if treated by the age of eleven. I think most of us would feel that a program of testing eleven-year-olds for condition J should be preferred to the program of testing intending mothers for condition K, unless the latter would be *much* more efficient than the former.

Parfit also argues that our obligations regarding future people cannot be accounted for entirely in terms of *rights*. I think this conclusion is probably right; but I will not discuss here his arguments for it, which are only partly based on the Non-Identity Problem.[40]

[39] Perhaps Parfit did not mean to show more than this. But his claim of "very wide theoretical implications, of an impersonal kind" for this argument (p. 447) suggests that he did.

[40] James Woodward argues for according a larger role to *rights* of future people, in a fine paper on "The Non-Identity Problem," *Ethics* 96 (1986), pp. 804–831. He criticizes my use of a similar problem, in papers about the problem of evil, as well as Parfit's use of the Non-Identity Problem. I am not convinced that the concepts of rights, fairness, and wronging people can be given the larger scope that Woodward claims for them in these matters. But I am persuaded by his examples that room must nonetheless be found in intergenerational ethics for such a rights-related con-

Parfit concludes that we need a theory of beneficence to deal with this area of ethics. Indeed he thinks "we need a new theory about beneficence" (p. 443) which he has not yet found if we are to solve all of the problems and paradoxes that set the agenda for Part Four of his book. All the accounts of beneficence he discusses contain or presuppose the principle, "If other things are equal, it is wrong knowingly to make some choice that would make the outcome worse" (pp. 394, 396). They differ from one another chiefly in what they say about what would make an outcome better or worse. They presuppose that (at least in many cases) alternative futures for the world, in which entirely different individual persons, and vastly different numbers of them, would live under widely diverse physical and social conditions, can be compared as globally better and worse, either impersonally, or for the people who would live in them, considered collectively.

This presupposition seems to me very questionable, though I will not launch a systematic attack on it here.[41] I believe a better basis for ethical theory in this area can be found in quite a different direction—in a commitment to the future of humanity as a vast project, or network of overlapping projects, that is generally shared by the human race. The aspiration for a better society—more just, more rewarding, and more peaceful—is a part of this project. So are the potentially endless quests for scientific knowledge and philosophical understanding, and the development of artistic and other cultural traditions. This includes the particular cultural traditions to which we belong, in all their accidental historic and ethnic diversity. It also includes our interest in the lives of our children and grandchildren, and the hope that they will be able, in turn to have the lives of their children and grandchildren as projects. To the extent that a policy or practice seems likely to be favorable or unfavorable to the carrying out of this complex of projects in the nearer or further future, we have reason to pursue or avoid it.

ception as that of compensation owed to a person for harm arising from an action without which her (worthwhile) life would never have begun.

[41] Some of the arguments of James Woodward (*op. cit.*, pp. 828–831), and of Philippa Foot, "Utilitarianism and the Virtues," *Proceedings and Addresses of the American Philosophical Association* 57 (1983), pp. 273–283, are relevant here.

The concept of "quality of life," which dominates Parfit's discussion of the evaluation of alternative futures, may have some role to play in thinking about what is "favorable or unfavorable" here. But it is too abstract to represent adequately the concrete concerns that are bound up in our commitment to the human project. And it focuses attention too much on the quality of experience at particular moments in the future, as opposed to how we get there. Continuity is as important to our commitment to the project of the future of humanity as it is to our commitment to the projects of our own personal futures. Just as the shape of my whole life, and its connection with my present and past, have an interest that goes beyond that of any isolated experience, so too the shape of human history over an extended period of the future, and its connection with the human present and past, have an interest that goes beyond that of the (total or average) quality of life of a population-at-a-time, considered in isolation from how it got that way.

We owe, I think, some loyalty to this project of the human future. We also owe it a respect that we would owe it even if we were not of the human race ourselves, but beings from another planet who had some understanding of it. But this is not the place to enter into a discussion of why it is not morally optional to care about this project.[42] For in what follows I will mostly not follow my own preferred line, but will discuss Parfit's paradoxes in his own terms of global outcome and quality of life. I think that even in those terms something can be done toward realizing his hope of more adequate solutions than those discussed in his book.

One test he proposes, to be passed by an acceptable theory, is that it should avoid

> *The Repugnant Conclusion*: For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things

[42] Considerations that I think highly relevant to this issue are developed in my paper on "Common Projects and Moral Virtue," *Midwest Studies in Philosophy* 13 (1988), pp. 297–307. Religious considerations are likely also to bear on it. Jonathan Bennett, in one of the best essays I have read on this subject, appeals to much the same sort of commitment to a project or "great adventure" as I have been discussing, but quite explicitly does not regard it as a matter of moral obligation; his paper, "On Maximizing Happiness," is in R. I. Sikora and Brian Barry, eds., *Obligations to Future Generations* (Philadelphia, Penn.: Temple University Press, 1978), pp. 61–73.

> are equal, would be better, even though its members have lives that are barely worth living (p. 388).

This absurdity clearly is implied by both impersonal and person-affecting forms of the Total Principle, which measures the value of outcomes by summing the net quantities that different persons enjoy in them of "whatever makes life worth living" (p. 387). This view allows the value of an outcome to be improved by the addition of sheer numbers of persons, so long as their lives are at least minimally worth living.

The most discussed alternative to the Total Principle for evaluating outcomes in terms of quality of life has been the Average Principle, which awards the prize to the outcome in which the average quality of life is highest. The Average Principle does indeed avoid the Repugnant Conclusion, for it awards no points for the addition of happy people, unless they are average-raisers. But Parfit quite rightly dismisses the Average Principle as implausible because it has such consequences as that the addition to the world of a person who will have a very good life can make the outcome worse just because other people have lives that are even better (pp. 420–422).

Another view considered by Parfit is that "[t]he value of quantity has an upper limit, and in the world today this limit has been reached" (p. 403). (By "quantity" here is meant the quantity of good that can be increased by increasing the sheer number of people, so long as their lives are worth living.) This view is plausible—though I think it would be even more plausible to think of the limit as a size of population rather than a sum of good. If the human race numbered only a few tens of thousands, we should probably think it a good thing to increase our numbers, so long as the newcomers would have lives worth living. Now that we number four billion, it is hard to see that it would make a better outcome to have more people just to be vessels for additional happiness—and I think that would be true even if the addition would not be burdensome to the rest of us.

Parfit argues, however, that this view is not tenable, and no limit can be set to the value of "quantity." His argument begins with the intuition (which I share) that a hell containing ten million innocent people would be a worse outcome than a hell containing just ten innocent people, even if the average level of misery were a little worse in the smaller than in the larger hell. From this he infers that "[i]n the case of suffering, there is no upper limit to the disvalue of quantity" (p. 406). But, he argues, if we still maintain that there is an upper limit to the positive value of the sum total of good enjoyed, we will be led to absurdly incongruous conclusions.

Imagine a population of many billions, living on a number of planets. Almost all of them have a much higher quality of life than is enjoyed by most of even the more fortunate people on Earth today. There is one in ten billion, however, who "through sheer bad luck" suffers so much as to have a life "not worth living." Now imagine another state of affairs, in which there is a population several times as large, living on proportionately more planets, with the same very high prevailing quality of life, and the same proportion of unfortunates. Since there is no upper limit to the disvalue of quantity of suffering, Parfit argues, the second state of affairs will be worse than the first, if there is an upper limit to the positive value of quantity of good lives.

> And, if this population was sufficiently large, its bad feature would outweigh its good features. Badness that could be unlimited must be able to outweigh limited goodness. If this population was sufficiently large, its existence would be intrinsically bad. It would be better if, during this period, no one existed (pp. 409f.).

I agree with Parfit that these consequences are implausible. But they do not show that there must be no upper limit to the value of "quantity." He has not canvassed enough possibilities. Consider the following principles:

> *Positive Threshold Principle*: If the number of people living at any time is at least *N*, the existence of a larger number of people with the same (or worse) average levels and distribution of happiness, suffering, and other goods and evils would not be better.
>
> *Negative Threshold Principle*: If the average levels and distribution of happiness, suffering, and other goods and evils among the people living at any time are *not too bad*, the existence of a larger number of people with the same average levels and distribution of happiness, suffering, and other goods and evils would not be worse.

In other words, there is a *quantitative* threshold beyond which mere quantity of good does not count, and a *qualitative* threshold beyond (better than) which mere quantity of suffering does not count in determining the overall value of states of affairs. These principles both seem plausible to me (if we are going to assign values to these global outcomes at all). Similar principles can be added, if necessary, to deal with other problems. It would probably be fruitless to try to quantify "not too bad" in the Negative Threshold Principle, but for present purposes we don't need to. The average levels in Parfit's Two Hells obviously are too bad, as would be the case in any population that had, on average a life not worth living (and perhaps in some other sorts of situation too). The average levels and distribution in Parfit's imagined populations with one wretched person in ten billion are clearly not too bad, in his judgment. Thus adopting these Threshold Principles would enable Parfit to avoid both the paradoxes that have driven him to abandon an upper limit on the value of "quantity."

This view also has the virtue of accounting for the asymmetry (Section 132) between the cases of the Happy Child and the Wretched Child. The addition of the Happy Child would merely increase the quantity of happiness beyond the quantitative threshold, and so would not result in a better state of affairs. (Other good effects, on other people, or on the *average* quality of life, are assumed not to be in view in this example.) But the addition of the Wretched Child would either make the average quality of life, and its distribution, worse, or, if the situation is already very bad, would increase the quantity of suffering in a population that is below the qualitative threshold, and so would result in a worse state of affairs.

Parfit sees another obstacle in his path as he seeks to avoid the Repugnant Conclusion. In the Mere Addition Paradox we start with possible state of affairs A: a large population, all with an extremely high quality of life. "The quality of life in B [with a population twice as large] is about four-fifths as high as the quality of life in A." Parfit shares the intuition that "B is worse than A," because of the lower (average and maximum) quality of life, although the total quantity of human good enjoyed is about 60% greater in B than in A.

A can be compared, however, with another state of affairs, A+, in which in addition to the A-people, enjoying their A-quality of life, there are "the extra people." There are as many of them as of the A-people, and their quality of life is only about half as high, but their lives are worth living. No issue of social injustice arises here, and the two groups do not harm each other, because they live on different continents (or it could be different planets) and do not even know of each other's existence (Section 142). Parfit argues that A+ is not worse than A. The "mere addition" of happy people does not make the outcome worse, even though it lowers the average quality of life and introduces (without social injustice) an inequality (Section 144). This point may be debated, but I will grant it.

The plot thickens as we turn to another possible state of affairs. In Divided B, as in B and A+, there are twice as many people as in A. As in B, their quality of life is roughly uniform, and about four-fifths as high as in A. But as in A+ (and unlike B), Divided B's people live in two groups of roughly equal size that do not know of each other's existence and have no influence on each other. Parfit argues that Divided B is better than A+ on several counts. In Divided B the average quality of life is higher, there is less inequality, and the worse off are better off. "In a change from A+ to Divided B, the worse-off half would gain more than the better-off half would lose" (pp. 425f.).

One more premise is needed to generate the paradox. "Clearly," Parfit states, "B is as good as Divided B" (p. 425). Given that "Divided B is better than A+," then, "[s]ince B is as good as Divided B, B is better than A+." But Parfit has argued that "A+ is not worse than A. We now believe that B is better than A+. These beliefs together imply that B is not worse than A"—contrary to our initial intuition (p. 426).

They might be thought to imply something worse—namely, the Repugnant Conclusion. "It may seem that, if B would be better than A+, which is not worse than A, B must be better than A" (p. 430). If so, the argument can be iterated, showing that C, which has twice as large a population as B, with a quality of life about 80% as high, is better than B, and hence better than A—and so forth until we reach the vast population of the Repugnant Conclusion, whose lives are marginally worth living, proving by this sorites that that state of affairs is better than A.

But this extension of the argument rests on a mistake, according to Parfit. So long as we claim no more for A+ than that it is *not* *worse than* A, we do not have to agree that if B is better than A+ then B is better than A. He suggests that A+ might be thought only roughly comparable to A; and "[w]hen there is only rough comparability, *not worse than* is not a transitive relation," and does not imply *at least as good as*. Because it does not imply *at least as good as*, we can hold that A+ is not worse than A, B is better than A+, and still B is not better than A, but only not worse than A. Because *not worse than* is not transitive, we can hold that B is not worse than A, and C is not worse than B, but nonetheless C is worse than A; and thus the argument will not carry us to even a "not worse than" version of the Repugnant Conclusion (Section 146).

This point about *not worse than* seems correct to me, and it can be used to solve the Mere Addition Paradox by attacking the premise, "Clearly B is as good as Divided B." "Is as good as" in this context must be intended by Parfit to express a transitive relation. Otherwise the argument, "We would thus believe that Divided B is better than A+. Since B is as good as Divided B, B is better than A+," would be fallacious. Thus the argument presupposes that B and Divided B are comparable, not only roughly, but with some precision.

Why would there be only rough comparability between A and A+? The most obvious reason is that in evaluating A we are evaluating a situation for a single population, whereas in evaluating A+ we are evaluating a situation in which there are two separate populations with no morally interesting relation between them. These are two quite different kinds of evaluation.

But there is exactly the same reason for thinking that there is only rough comparability between B and Divided B as there is for thinking that A and A+ are only roughly comparable. In evaluating B we are evaluating a situation for a single population, whereas in evaluating Divided B we are evaluating a situation in which there are two separate populations with no morally interesting relation between them. So why should we assume that "B is [at least] as good as Divided B" in a sense intended to be transitive? It was not the difference in welfare levels between A and A+, but their mutual isolation, that kept us from making a more precise value comparison between them. Without the isolation, I take it, Parfit would agree that it is initially most plausible to say that A+ is worse than A. That being so, he is not entitled to the assumption that simply eliminating the welfare disparity makes precise comparison possible between B and Divided B.

This conclusion can be reinforced by reflecting on the ways in which A+ might change into B or Divided B. In trying to persuade us that Divided B is better than A+, Parfit helps himself to evaluations of processes, explicitly envisaging a change from A+ to Divided B as a gradual change over two centuries, as "the result of natural events, affecting the environment" (p. 425). I think we may fairly infer that he envisages it as *not* involving any change in the mutual isolation. But how would A+ evolve into B? This might happen by one population discovering the other, and the richer population then voluntarily making some sacrifice to effect a larger improvement in the welfare of the other. This would be a morally attractive history; and we might want to say that if B developed in that way, it would indeed be a better state of affairs than A+ (and *at least* as good as Divided B). But if we stick with Parfit's assumption of no interaction between the two populations, then the only way A+ could evolve into B would be by one population developing into B and the other dying out.[43]

Suppose it is the richer population that develops into B. Considered in itself, this is equivalent to the development of B from A, which Parfit admits would be most plausibly regarded as a change for the worse. And it's hard to see how he could think of the dying out of the other population as a good thing, even if it happened relatively painlessly (perhaps through universal but voluntary adoption of celibacy). So B would hardly be an improvement on A+ if it developed in this way. In fact, it would be plausible to think of this development as a deterioration in the situation. Perhaps this would be a judgment mainly on the process of change, and would not give a clear verdict on the comparative

[43] I ignore here the possibility that B might evolve out of A+ by way of Divided B, the two populations discovering each other, meeting, and mingling *after* they had come to the same quality of life in Divided B. This may be the scenario in which it is most plausible to think that B is precisely as good as Divided B. But throwing it into the hopper as yet another alternative only underlines the impossibility of getting a plausible precise comparison of the value of these states of affairs independently of the history by which they would arise.

merits of the initial and terminal states as distinct from the process. But then I think we are likely to be left unsure how to make the latter comparison. We have (as well as need) much clearer intuitions about the the value of possible *changes* than about the comparative value of states of affairs considered in abstraction from any possible story about how one would get to them.

Suppose it is the poorer population that develops into B (certainly an improvement) while the richer dies out. It is very hard to say whether this would be a change for the better (or, more cautiously, a good change), because it is so hard to evaluate the dying out of either population. The main upshot of all these considerations about how A+ could change into B, in my opinion, is that after reflecting on them we are not likely to have confidence in any precise assessment of the comparative value of A+ and B as such, independently of how they would have arisen. In other words, these considerations strongly confirm the judgment that A+ and B are no more than roughly comparable. And if we grant that Divided B can be judged more precisely to be superior to A+ (on the assumption that there has never been any contact between the populations), the judgment that B and Divided B are only roughly comparable is also confirmed.

The most plausible comparative evaluation of B and Divided B, therefore, is that *neither is worse than* the other, and that this relationship is *not* transitive. But now (the first version of) the Mere Addition Paradox collapses. We have the following relationships:

(i) A+ is not worse than A.
(ii) Divided B is better than A+.
(iii) B is not worse than Divided B.

"Is better than" is a tight enough relationship so that from (i) and (ii) we can infer

(iv) Divided B is not worse than A.

But from (iii) and (iv) we *cannot* infer

(v) B is not worse than A,

because transitivity fails.

Parfit presents a second version of the paradox, which is more threatening, inasmuch as it does lead, in his opinion, to the Repugnant Conclusion (Section 148). We begin with the two separate populations of A+; but in this version "even the worse-off group have an *extremely* high quality of life" (p. 434). There are ten billion people in each group, and they live on different planets. In all the other states of affairs that we will consider as alternatives to A+, there are a vast number of inhabited planets, each with a population of ten billion persons; as in A+, none have knowledge of the people on other planets. In New A the people on two of the planets enjoy a quality of life even higher than that of the better-off group in A+; but the people on the remaining planets (the overwhelming majority of New A's total population) "are not much above the Bad Level" (the quality of life below which "it would in itself have been better if they had never existed").

Parfit argues that New A is better than A+. "There is at least one way in which New A is better than A+. In New A there are twenty billion people, all of whom have a higher quality of life than [any of the twenty billion people] in A+." And the inequality in New A is not worse than that in A+, because inequality "produced by Mere Addition . . . does not make the outcome worse" (p. 434).[44]

Now consider New B, which is like New A except that the people on four rather than two of the planets are better off than the others. The forty billion fortunates in New B all enjoy a quality of life about four-fifths as good as that of the twenty billion privileged in New A. The people on the remaining planets still subsist just above the Bad Level. Parfit argues that "New B is better [than

[44] Parfit tries to score an extra point here by arguing that the inequality in New A is actually *better* than in A+, on the ground that "[t]here is no longer inequality between the two best-off groups." This is a bad argument. It depends on Parfit's thinking of the best-off groups in New A as the two groups from A+, and the many worse-off groups in New A as the "mere additions." But at this point in his argument, the identities of people are supposed to make no difference to the comparative values of outcomes. This type of flaw in his argument will be discussed more fully below, in another connection.

New A] on any plausible principle of beneficence." For "[i]f there was a change from New A to New B, worse-off groups would gain *very much more* than better-off groups would lose" (p. 435).

Better than is a transitive relation. So if New B is better than New A, and New A is better than A+, then New B is better than A+. The path of sorites to the Repugnant Conclusion is clear. We will arrive eventually at a state of affairs in which there are very many tens of billions of people, none of them much elevated above the Bad Level, but which must be judged a better state of affairs than A+, in which all of the twenty billion people enjoy an extremely high quality of life.

I believe that this argument is unsound, and in particular that there is as much reason to think A+ is better than New A as to think New B is better than New A. For consider the morally relevant differences between New A and New B. New A has the advantage that the quality of life that is ever achieved by a significant number of people is significantly higher there than in New B. On the other hand, in New B there is (i) double the number, and (ii) a (perhaps significantly) higher proportion, of people enjoying a very high quality of life; and there is (iii) a (perhaps significantly) higher average quality of life for the aggregate of populations, and (iv) the overall distribution of quality of life is somewhat more egalitarian.

Reflection on this comparison might lead us in more than one direction. The first point to which I wish to call attention is that except for (i), all the advantages of New B over New A are also advantages that A+ enjoys over New A[45]—and (i) is a very dubious advantage, since the total population in all these cases is above the threshold beyond which it is implausible to think that mere quantity matters. So if it is clear to us that New B is better than New A, why shouldn't we also conclude that A+ is definitely better than New A, contrary to Parfit's claims?

But perhaps New B is not better than New A. If all the people in New A and New B were socially related to each other (in a broad sense), the advantages of New B over New A would be morally

[45] Indeed they also seem to be advantages that A+ enjoys over New B, but I will not develop that point here.

decisive. But since they are not so related, the comparison is not so clear. It is not clear, for example, that a more egalitarian distribution between socially unrelated populations is a moral advantage. Nor is it clear that averages of quality of life across socially unrelated populations are morally important. These considerations might lead us to conclude that New B and New A are only roughly comparable, and neither is worse than the other. This relation, being intransitive, will not lead to any version of the Repugnant Conclusion.

Parfit's argument largely ignores these considerations, and relies heavily on one-to-one matchings of the welfare of the isolated subgroups, which causes the comparison between New A and A+ to come out quite differently from that between New A and New B, because of the much smaller number of subgroups in A+. There is much potentiality for misleading argument here. Parfit says, for example, that in New B, in comparison with New A, "it would be true that, though the better-off group would lose, a worse-off group would gain *several times as much*" (p. 435). The same could be said about A+ in comparison with New A, if the worse-off subgroup in A+ were identified with one of the worst-off subgroups in New A. When making his comparison arguments, Parfit speaks as if the two subgroups in A+ are assumed to be identical with the two best-off groups in New A; but officially, as far as I can see, the identity of the subgroups is not supposed to play a part in the argument.

Suppose it is meant to play a part, however, and that A+ is composed of the two best-off groups of New A. Then we are certainly entitled to consider also A*, which is just like A+ except that the less well-off group in A* is identical with one of the worst-off groups in New A. Now if New B is better than New A, A* must surely be better than New A with respect to this pair of groups. And if mere quantity does not matter in these cases, so that its additional populated planets are not a countervailing advantage of New A over A*, we seem to be led to the conclusion that A* is better on the whole than New A. But since A+ and A* are qualitatively identical, it is hard to deny that they have exactly the same value, and that if A* is better than New A, so is A+.

By judicious choice of perspective, now one, now the other of many such pairs of "outcomes" can be made to seem the better. I think the wisest conclusion to draw is that it is a very dubious enterprise to assign comparative values to outcomes in abstraction from our moral assessment of processes by which they might arise. But even if we insist on pursuing that enterprise, I do not see how the Mere Addition Paradox can be made to stick, or to lead to the Repugnant Conclusion.[46]

University of California, Los Angeles

[46] I wish to thank Marilyn McCord Adams for discussion of ideas for this article, and Warren Quinn and the editors of *The Philosophical Review* for comments on drafts of it.