Intransitivity and the Mere Addition Paradox

Larry S. Temkin

Philosophy and Public Affairs, Vol. 16, No. 2 (Spring, 1987), 138-187.

Stable URL:
http://links.jstor.org/sici?sici=0048-3915%28198721%2916%3A2%3C138%3AIATMAP%3E2.0.CO%3B2-Q

Philosophy and Public Affairs is currently published by Princeton University Press.

LARRY S. TEMKIN

Intransitivity and the Mere Addition Paradox

In "Future Generations: Further Problems,"[1] and Part Four of *Reasons and Persons*,[2] Derek Parfit raises many perplexing questions. Although some think his ingenious arguments little more than delightful puzzles, I believe they challenge some of our deepest beliefs. In this article, I examine some of Parfit's arguments, focusing mainly on "The Mere Addition Paradox." If my analysis is correct, Parfit's arguments have extremely interesting and important implications that not even Parfit realized.

In Part I, I present Parfit's argument for the Mere Addition Paradox, and show that given Parfit's assumptions, a radical conclusion seems to follow: the notion of "better than"—indeed, even the notion of "all things considered better than"—is *not* a transitive relation. In Part II, I argue that Parfit's results won't stand if his assumptions are revised so as to avoid this conclusion. Most people believe "all things considered better than" *must* be transitive. In Part III, I discuss the plausibility and cost of retaining this view. Among other things, I shall argue that important positions besides Parfit's raise the specter of intransitivity, including Rawls's.

I am grateful to the National Humanities Center, whose generous support enabled me to complete this article. I am also grateful to Baruch Brody, John Broome, David Gauthier, Richard Grandy, Carl Hoefer, Shelly Kagan, Evelyn Keyes, Gregory Trianosky, Jennifer Whiting, members of the Center for Philosophy and Public Policy, and the Editors of *Philosophy & Public Affairs* for their comments. Derek Parfit opened my eyes to the issues discussed, and his written and oral comments were invaluable. Finally, I would like to thank Richard Dees and Martin Halbert—past students whose relentless worrying about Parfit's conclusions forced me to reconsider the structure and implications of his work.

[1] *Philosophy & Public Affairs* 11, no. 2 (1982):113–72.

[2] Oxford: Oxford University Press, 1984.

A word about terminology. In what follows, I shall often use *preferability* and its cognates as shorthand for "all things considered better than" and its cognates. So, to say that A is preferable to B, or better regarding preferability, is to say that all things considered A is better than B; it is not to say that any person, or group, actually prefers A to B (though perhaps, on certain interpretations, it corresponds to what an "ideal observer" or "rational impartial spectator" would prefer).

I

A

Parfit's argument for the Mere Addition Paradox is set up by a discussion of "the Repugnant Conclusion," which runs roughly as follows. Consider Diagram 1, where the width of the blocks represents the number of people living, and the height, their quality of life.

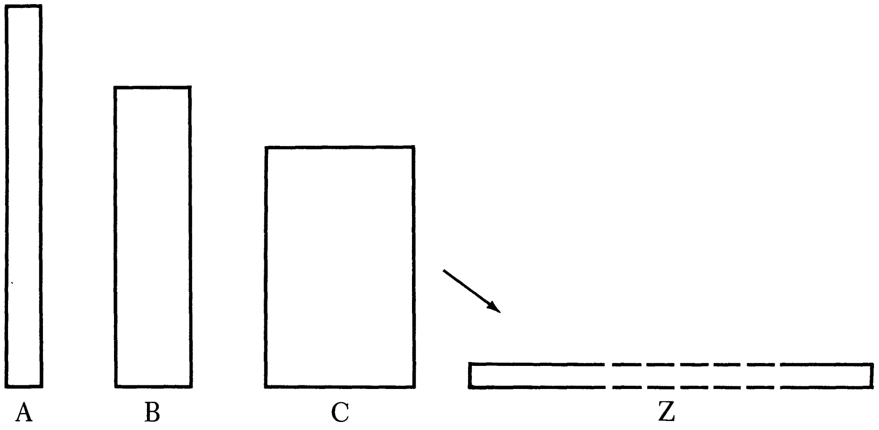

Diagram 1

If we were total utilitarians,[3] we would think B is better than A. While everyone in B is worse off than everyone in A, B is twice as large as A, and those in B are *more* than half as well off as those in A. Similarly, we would think C is better than B, since C stands to B, as B stands to A. By this reasoning, Z would be the best, where "Z is some enormous popu-

[3] Parfit's discussion is put in terms of what he calls "the Wide Total Principle." For our purposes, it will be sufficient to use the terminology of "total utilitarianism."

lation whose members have lives that are not much above the level where life ceases to be worth living" (p. 142).[4] Thus, total utilitarianism implies:

> *The Repugnant Conclusion* [RC]: For any possible and large population, say of eight billion, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, and be what we ought to bring about, even though its members have lives that are barely worth living (p. 142).

Parfit believes RC is *intrinsically* repugnant, and that while a few might accept RC, and others some, but not all, of the steps leading to it, many would hold that *all things considered* A is better than B, B is better than C, C is better than D, and so on. Correspondingly, Parfit contends that, on reflection, most believe that the mere addition of extra lives does not improve a situation. Thus, he writes, "If these [extra] lives are worth living, they have personal value. But the fact that such lives are lived does not make the outcome better."[5]

The Mere Addition Paradox can now be presented.[6] Consider Diagram 2.

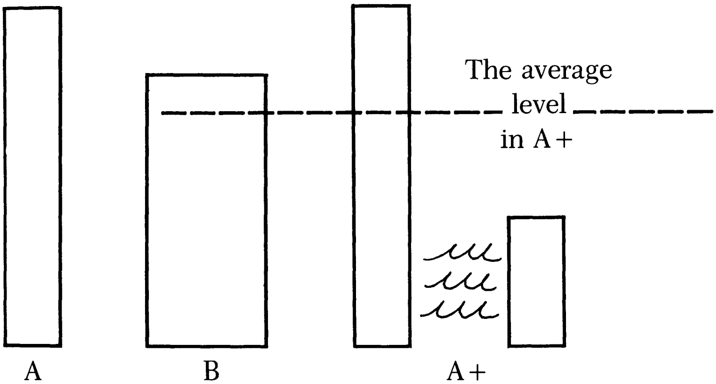

Diagram 2

[4] In this article all page references are to "Future Generations: Further Problems" unless noted otherwise.

[5] *Reasons and Persons*, p. 412.

[6] For simplicity, I omit Parfit's "Divided B." This does not affect my argument.

The move from A to A+ involves the *mere addition* of a group, all of whom have lives worth living, and affect no one else. The move from A+ to B involves the worse-off gaining more than the better-off lose.

Parfit argues that all things considered most would judge that A+ is not worse than A. He writes:

> Is A+ worse than A? Note that I am not asking whether it is *better.* This we have already implicitly denied, since we have denied that extra lives . . . have intrinsic moral value. [Still,] it seems harder . . . to believe that A+ is *worse* than A. This implies that it would have been better if the extra group had never existed. If their lives are worth living, and they affect no one else, why is it bad that these people are alive? (p. 158).

Parfit also argues that most would judge B to be better than A+, since it is better on both utilitarian and egalitarian grounds, and since, by assumption, the worse-off are so through no fault of their own.

According to Parfit, then, most believe that A+ is not worse than A, and that B is better than A+. Together, Parfit claims, these beliefs "imply that B is not worse than A [since] B cannot be worse than A if it is better than something—A+—which is not worse than A" (p. 160). But, as the discussion of the Repugnant Conclusion revealed, most believed that B *is* worse than A. Hence, our beliefs are inconsistent. This is the *Mere Addition Paradox.*

Parfit insists that the Mere Addition Paradox does *not* merely illustrate a tension between different moral principles, which may often conflict and cloud our judgments of preferability. Such cases needn't reflect any inconsistency. "In the Mere Addition Paradox," Parfit claims, "things are different. We are here inclined to believe, all things considered, that B is worse than A, though B is better than A+, which is not worse than A. These three judgments cannot all be consistently believed. They imply contradictions. One of these beliefs must go" (p. 160).

In sum, Parfit believes that rationality requires us to give up one of the three claims, though all three seem very plausible. Hence we face, Parfit thinks, a genuine paradox.

B

One natural response to the Mere Addition Paradox is to reject the move from A to A+, by claiming that A+ *is* worse than A, because it is worse regarding inequality. Against this response one might reason as follows. Typically, when we say the inequality in one situation is worse than another, the same people exist in both situations, and the worse-off in the one situation fare worse relative to the better-off than they do in the other. This, we may agree, is bad. However, in comparing A+ to A, the choice isn't between a situation where the worse-off fare poorly relative to the better-off and one where they fare better; rather, it is between one where they exist—with lives worth living—and one where they don't. Here, it may seem, the inequality is not morally regrettable.

Parfit explicitly adopts this line in *Reasons and Persons.* He writes:

> Whether inequality makes . . . [an] outcome worse depends on how it comes about. It might be true either . . . that some existing people are worse off than others, or . . . that there are extra people living who, though their lives are worth living, are worse off than some existing people. Only . . . [the former] makes the outcome worse (p. 425).

Parfit is *not* denying the obvious fact that A is perfectly equal while A+ is not, nor that A+'s inequality is morally regrettable when compared with B's. His contention is that A+'s inequality is not regrettable *if the alternative* is A.[7] As we shall see later, a crucial implication of this view is that one must assess alternatives directly to compare them regarding inequality. Knowing how two situations compare with a third will not, in general, be helpful in determining how *they* compare.

For similar reasons, Parfit rejects the view that A+ might be worse than A because it is worse regarding Maximin, where, roughly, one situation is worse than another if its worst-off group fares worse. Parfit

[7] In "Future Generations: Further Problems," Parfit suggests that the inequality in A+ may be "morally regrettable" compared with A, but that "it seems hard to believe that this feature is *so* bad as to make A+ worse than A" (pp. 158–59, my emphasis). However, as I indicated to Parfit in conversation during the spring of 1983, and in an earlier draft of this article the following summer, this answer is not adequate if one believes that, as Parfit suggests, the lives of the extra people in A+ have no intrinsic moral value. On the latter view there is no respect in which A+ is better than A. Hence, if the inequality in A+ is bad *at all*, it seems one should reject the claim that A+ is not worse than A, since there would be one respect in which A+ was *worse* than A, and *no* respect in which it was *better.*

Parfit notes the shortcoming of his earlier view together with his considered response concerning A+'s inequality in *Reasons and Persons*, pp. 422–25. Although his considered response did not appear in "Future Generations: Further Problems," I believe, based on our earlier conversations, that it was influencing his thinking—though perhaps not fully worked out—even before the article was published.

contends that Maximin is not plausibly applied in cases of mere addition, arguing strongly that if the only reason the worst-off group is better off in one outcome than another, is that in that "outcome certain people do not exist who, in the other outcome, would have lives that are worth living," then the higher level of the worst-off group "is *not* a good feature . . . [and] does not make . . . [that] outcome better" (*Reasons and Persons*, pp. 423–24).

To sum up, the Mere Addition Argument purports to illustrate an inconsistency in people's judgments. It is directed toward those who, like Parfit, believe:

(a) The mere existence of extra lives worth living does not make an outcome better,

(b) While inequality is normally a bad feature, it does not make an outcome worse if it involves the mere addition of extra people who have lives worth living,

(c) Maximin is not plausibly applied in cases of mere addition.

Not everyone shares these beliefs, but on reflection I suspect many would. As Parfit amply shows, powerful arguments support them, and perplexing difficulties attend their rejection.[8] Still, these beliefs do not yield the results Parfit thinks they do. Instead, they seem to yield an even more troubling and controversial result. Unfortunately, to establish this I must digress and introduce some technical distinctions. This will facilitate discussion throughout the remainder of the article.

C

It is standard to define the notions of "better than," "as good as" (or "equivalent to"), and "at least as good as" in terms of each other. So, x is *at least as good as* y if and only if x is *as good as* y or x is *better than* y. Similarly, x is *as good as* or *equivalent to* y if and only if both (a) x is at least as good as y, and (b) y is at least as good as x. On the other hand, x is *better than* y if and only if (a) obtains but not (b). Analogous definitions relate "worse than," "as bad as," and "at least as bad as," where, of course, x is worse than y if and only if y is better than x.

The above definitions are uncontroversial. So too, people have thought, is the following inference scheme:

[8] See *Reasons and Persons*, chaps. 17–19.

> (1) Given any concept c, for all x, y, and z to which c is appropriately applied, if x is c-better than y, and y is c-better than z, then x is c-better than z. (Here, and in what follows, to say that x is *c-better* than y is to say that x is better than y with respect to c, and similarly for *c-worse than, not c-worse than*, and so on.)

Since a relation, R, is *transitive* if and only if for all x, y, and z, if xRy and yRz then xRz, let us call belief in the validity of (1) the belief that *better than is always transitive.* (Of course, this belief only concerns cases where the *same* concept c is involved. If x is a better teacher than y, and y is a better tennis player than z, nothing follows about how x and z compare either as teachers or tennis players.)

In addition to (1), most people accept what might be called *the principle of substitution* for equivalence, or:

> (2) Given any concept c, for all x, y, and z to which c is appropriately applied, if x is as c-good as, or c-equivalent to y, then however x c-compares to z, that is how y c-compares to z.

This principle parallels the principle of substitution for simple algebra, according to which if x is equal to y, x and y can be interchanged without affecting the value of the formulae in which they occur.

Together with the standard definitions noted above, (1) and (2) entail many inferences people readily accept. For our purposes, let us distinguish three of these. Given any c, for all x, y, and z to which c is appropriately applied:

> (3) If x is at least as c-good as y, and y is at least as c-good as z, then x is at least as c-good as z.
>
> (4) If x is c-better than y, and y is at least as c-good as z, then x is c-better than z.
>
> (5) If x is c-better than y, and y is not c-worse than z, then x is not c-worse than z.

Following our earlier terminology, belief in (3) might be called the belief that *at least as good as is always transitive.* As the reader can easily check, (3) and (4) follow from both (1) and (2), while (5) follows directly from (1).

So far I have been presenting claims which, up to now at least, have been regarded as uncontroversial. (Indeed, in the pre-Quinian era, I think virtually everyone would have agreed that (1)–(5) were true in virtue of their meanings, together with the laws of logic.) However, the situation appears a bit more complicated for certain inferences involving the expressions "not worse than" and "not better than."

At first, it is natural to interpret "at least as good as" to mean "not worse than." On this interpretation, (3) and (4) entail:

> (6) Given any concept c, for all x, y, and z to which c is appropriately applied, if x is not c-worse than y, and y is not c-worse than z, then x is not c-worse than z, and
>
> (7) given any concept c, for all x, y, and z to which c is appropriately applied, if x is c-better than y, and y is not c-worse than z, then x is c-better than z.

Moreover, for many concepts, c, (6) and (7) *are* valid. However, as some logicians and economists have recognized, on one plausible interpretation of "not worse than," inferences like (6) and (7) are at most valid for a special class of concepts. These are concepts which are *complete*, or allow *full comparability*, in the sense that for all x and y to which c is appropriately applied, either x is at least as c-good as y, or y is at least as c-good as x.

For *complete* concepts, "not worse than" is extensionally equivalent to "at least as good as." Correspondingly, insofar as people have thought in terms of such concepts (too often, I fear), they have naturally supposed that, like (3) and (4), (6) and (7) are valid. However, many concepts are not complete.

Parfit writes, "Must it be true, of Plato and Aristotle, either that one was the greater philosopher, or that both were exactly equally as great? This seems absurd. But it is surely true that some philosophers are greater than others, and by more or less" (p. 165, n. 43). Parfit's suggestion is that our notion of "great philosopher" is *incomplete*. It allows *partial comparability* in the sense that it enables us to make *some* comparisons between philosophers as to which is greater, but not others. Moreover, our inability to make precise comparisons in all cases need not be due to any ignorance on our part—either of the abilities or achievements of the philosophers in question, or of what the notion of great philosopher involves. It may be due to the roughness or complexity intrinsically involved in that notion. In such cases, where the notion does not permit precise comparisons, Parfit suggests we can rightly say, as of Plato and

Aristotle, that each is not worse than the other, without their being exactly equally as great.[9]

Reflection suggests that incompleteness may be an intrinsic feature of many notions. To note but one other example: it may be true of Ty Cobb and Joe DiMaggio that each was neither a better nor worse baseball player than the other. This need not mean they were exactly equally as great. Nor would it mean that we couldn't make other comparisons perfectly well, for example, that both were better baseball players than most who ever lived.

Once it is recognized that certain notions are incomplete, one can see how "not worse than" can be used such that it is not a transitive relation. For example, among the great philosophers, one might believe that Descartes is not worse than Hume, who is not worse than Kant, who is not worse than Aristotle, yet believe that Descartes *is* worse than Aristotle—that is, it might be that our notion of great philosopher permits precision in the latter comparison, but not in the former ones. (The failure of the transitivity of "not worse than" may sound like yet another example of the familiar Sorites paradox. As will become clear below, this is *not* all that is going on in the cases that interest us.)

So, (6) is not valid for incomplete notions, and from this it follows that neither is (7).

Two comments about the reasoning presented. First, like Parfit, most who recognize that (6) and (7) are not always valid, conclude that "not worse than" is not transitive. While technically correct, this is perhaps a bit misleading. It directs attention away from the fact that whether the relation "not worse than with respect to c" is transitive or not, depends on the nature of c. For certain concepts, (6) and (7) *are* valid, and, we might say, for those concepts "not worse than" *is* transitive. For other concepts, (6) and (7) are *not* valid, and it is for *those* concepts we should perhaps reserve the claim that "not worse than" isn't transitive.

Second, among those who believe that "not worse than" is not always

[9] In suggesting that a notion may not permit precise comparisons, I am not taking a stand on the ultimate reason for this. Aristotle warns us not to expect finer distinctions in our theories than their subject matters permit. Perhaps, in some cases, an intrinsically incomplete notion is simply an accurate mirror of the world.

Also, some use the locution "not worse than" as a synonym for "at least as good as." And words can be used (more or less) as people like. But some have come to use "not worse than" in the manner Parfit suggests, and if that locution is not used to describe the relation in question, another must be found instead.

transitive, I think most in fact believe this for the kind of reasons presented. However, as we shall see, there may be other, somewhat related, reasons which support not only that conclusion, but stronger ones as well.

Let us next introduce some new terminology. Although technically it is *relations* which are transitive, it will be useful to speak of notions or concepts as being transitive or intransitive to varying degrees. Specifically, let us say that a concept, c, is (a) *fully transitive* if inferences (1)–(7) are each valid for c, (b) *partially transitive* or *partially intransitive* if some of (1)–(7) are valid, but others are not, and (c) *deeply intransitive* if one or more of inferences (1)–(5) are invalid. Using this terminology, most would say that the concept "height" is fully transitive,[10] while the concept "great philosopher" is only partially transitive (or intransitive), though not deeply intransitive.

I believe that people have only recently recognized that some concepts may merely be partially, rather than fully, transitive. I also believe that most people, including Parfit, believe that even if *some* concepts are partially transitive, none to which the schema (1)–(7) might be applied are deeply intransitive. As we shall see next, in returning to the main argument of this part, this latter belief is false if the assumptions underlying the Mere Addition Paradox are correct.[11]

D

In section B, we saw that the Mere Addition Paradox is directed toward those who accept that the inequality in A+ is morally regrettable when the alternative is B, but not when the alternative is A. On this view,

[10] Inference schemes analogous to (1)–(7) have been thought to apply not only to "better than" but, more generally, to ". . . er than," e.g., "taller than," "heavier than," "smarter than," etc. And, traditionally, concepts like height and weight have been thought complete, and hence fully transitive. If quantum mechanics and relativity theory are correct, this may be mistaken. Fortunately, this need not concern us here.

[11] The discussion in this section is somewhat oversimplified and thereby misleading. Even on the usual view, whether a concept is fully transitive will not strictly depend on whether it is complete. So, for example, by constructing artificial concepts, or carefully restricting one's domain, one could generate complete concepts that are only partially transitive, or incomplete concepts that are fully transitive. But these technical complications need not concern us here, as they do not significantly affect the substance of my remarks. The main point is that most have believed that the moral ideals discussed in this article are fully transitive, or at worst, if incomplete, partially transitive. It has been thought that such ideals are not, and could not be, deeply intransitive.

Equality is Comparative, not merely in the ordinary sense—that it involves judgments about how some fare relative to others—but in the sense that our judgment about a situation's inequality depends on the alternative it is being compared to. As I noted in section B, on the view that Equality is Comparative (EC), knowing how two situations compare to a third will not, in general, help determine how *they* compare regarding inequality. Let us expand upon this observation, in light of section C.

First, it is easily seen that on EC, the concept of inequality is not fully transitive. According to EC, A+ is not E-worse than A (here, and elsewhere, *E-worse* stands for "worse with respect to inequality," *E-better* stands for "better with respect to inequality"). Yet clearly, A is not E-worse than B. It follows that (6) is not valid for inequality, since A+ *is* E-worse than B. Similarly, (7) is not valid, since B is E-better than A+, and A+ is not E-worse than A, but B is *not* E-better than A. So, according to EC, the relation "not worse than with respect to inequality" is not transitive, and at best inequality is partially transitive.

So far, we only have reason to believe the concept of inequality is incomplete. But notice, with respect to inequality, A and B are both ideal. They aren't merely not worse than each other, they are exactly equally as good. (Not, of course, "all things considered," but *with respect to inequality*.) Yet, B is E-better than A+, which is not E-worse than A. This entails that (2) and (4) are *not* valid. Thus, in the terminology of section C, inequality is *deeply intransitive* if EC is correct.

The preceding result is startling and significant. As we have seen, some have recognized that a concept can be partially intransitive because incomplete. So, a notion too imprecise to permit a comparison between a and b, or b and c, may yet permit one between a and c. However, the above situation is markedly different. Indeed, they are almost polar opposites. We know precisely how A and B compare, and precisely how B and A+ compare, but we can't conclude from that how A and A+ compare. In fact, our supposed knowledge about how A and A+ compare, seemingly conflicts with how they should compare, given how A compares to B, and how B compares to A+.

How can this be? This much is clear: partial intransitivity stemming from the rejection of (2) and (4) cannot be explained by appeal to a concept's incompleteness, the way it has been thought partial intransitivity stemming from the rejection of (6) and (7) could. Thus, the situation is *not* like that of a Sorites paradox. It requires an entirely different analysis and explanation.

The advocate of EC believes that inequality is not objectionable when it is brought about by the mere addition of extra people all of whom have lives worth living and who affect no one else. This underlies her view that the inequality in A+ *is* morally regrettable when A+ is compared to B, but not when it is compared to A. It appears, then, that on EC the *relevant and significant* factors for comparing A and A+ regarding inequality differ from those for comparing A+ and B in a sense connected with inequality being *essentially pairwise comparative.* The full content and implications of this will become clearer as the article progresses, particularly in Part III. For now, it is sufficient to see that it is this essentially comparative feature of EC which accounts for the startling result that (2) and (4) are not valid for the concept of inequality. Moreover, reflection reveals that it is also this, and not incompleteness, which accounts for the invalidity of (6) and (7) in Parfit's example. Thus, while the invalidity of (6) and (7) may at first seem unsurprising when we consider the Mere Addition Paradox—because instances of (6) and (7) being invalid are, for *other* reasons, more familiar to us—in fact, the reasons for the invalidity of (2), (4), (6), and (7) are the same, for the advocate of EC.

As suggested earlier, then, there are different reasons a relation "not worse than with respect to c" may be intransitive. This may be because c is incomplete. Alternatively, it may be because c is essentially pairwise comparative. Unfortunately, even in the latter case c may often be complex and incomplete. Correspondingly, both sources of intransitivity may obtain, and it may be easy to confuse the different reasons "not worse than with respect to c" may be intransitive. Nevertheless, these reasons *are* different, and they differ in their implications.

When a concept is partially intransitive only for the first reason, it will not be deeply intransitive. But when it is partially intransitive for the second reason, it can be. This is because where a concept is essentially pairwise comparative, it can be true that even precise comparisons between a and b, and b and c, do not reflect how a compares to c. Thus, (4) might be invalid because although A is E-equivalent to B, and B is E-better than A+, A might *not* be E-better than A+. Analogous remarks could apply to each of (1)–(7).

We have seen that EC entails rejecting (2), (4), (6), and (7). It also entails rejecting (1), (3), and (5).[12] While initially this may seem shocking, once it is accepted that inequality is deeply intransitive, and once it is understood what this involves, the rejection of (1), (3), and (5) should be no *more* surprising or controversial than the rejection of (4), which is clearly implied by EC. Thus, on EC, the principle of substitution for equivalence must be rejected, along with the beliefs that "not worse than," "better than," and "at least as good as" are always transitive.

One point cannot be overemphasized. On EC, how bad the inequality of a situation is, is not an *intrinsic* feature of that situation—that is, it depends on factors that are not internal to the situation. There is no fact of the matter as to how bad the inequality in A+ *really* is considered just by itself.[13] How bad it is depends upon the alternative compared to it. Compared to B, A+ is bad; compared to A, it isn't.

A second point worth emphasizing is that just as there may be no fact of the matter as to how bad a situation's inequality is considered by itself, there may be no fact of the matter as to how two situations compare considered from a more or less abstract perspective. Suppose, for example, half of the people in A were lowered to the level of the worse-off in A+, and then the population was doubled. A world like A+ would result, and it appears that brought about *this* way, the inequality in A+ *would* be worse than that in A.[14] Thus, even to make the kind of static judgments Parfit is interested in, it would not be sufficient to compare abstract diagrams of A and A+. In addition, we might need to know the relation, if any, between them. In particular, we might need to know who their members are, or how they've come about.

The above result may seem puzzling, but it has a straightforward explanation. On EC, how good a situation is regarding inequality will depend on the alternative with which it is compared. But, in terms of the relevant

[12] Unfortunately, the proof of these claims cannot be pursued here, as it requires more argument than space permits. But the proof is not too difficult to construct.

[13] The issue here is not one of objectivity versus subjectivity. On EC, there may be a fact of the matter regarding how bad the inequality in A+ *really* is when it is compared to A, and also when it is compared to B. But there are no facts about how good or bad situations are regarding inequality outside the context of essentially pairwise comparisons. Cf. note 48.

[14] All of the economists' measures of inequality would support this point, as does my own work on inequality. See my *Inequality* (forthcoming, Oxford University Press). Also, cf. Part III, section C, and note 33 for considerations relevant to this claim.

and significant factors for applying EC, the alternative with which it is being compared may itself partly depend on the members involved or how it has come about. In other words, despite their abstract structural similarities, A and A+ represent different pairs of alternatives in the cases imagined. This explains why our judgments can vary about those cases.

E

Considerations analogous to the preceding ones might also apply to the view that Maximin is not plausibly applied in cases of mere addition, as well as to the view that the mere addition of extra lives does not make an outcome better. Though not necessary to establish my central claims, let me briefly illustrate this, beginning with the former.

Many have a special concern for how the worst-off fare. They believe it is morally regrettable if the worst-off fare worse than they otherwise might, and this may lead them to accept a principle like Maximin, according to which, other things equal, A is worse than B if its worst-off group fares worse. Still, paralleling EC, they may believe that *Maximin is Comparative* (MC)—that while usually it is morally regrettable if the worst-off fare worse in one situation than another, this is not always so, and, in particular, is not if it results from mere addition. Moving from A to A+, it is not as if the worst-off in A (that is, everyone) fare worse in A+—they fare the same in both. Nor do the worst-off in A+ fare worse than they do in A—*they* don't exist in A, while their lives are worth living in A+. Here, it may seem, our concern for the worst-off would not lead us to judge A+ worse than A. But neither would it lead us to judge A+ better than A. Instead, we may judge A+ as good as A—since the worst-off group in A fares the same in both—or, alternatively, that neither is worse than the other. In Part III, I shall say more about Rawls's version of Maximin, together with other versions one might consider. For now, let me focus on MC.

On MC, as with EC, the relevant and significant factors for comparing alternatives may depend on what those alternatives are, and this, in turn, may depend on who their members are, or how they have come about. Suppose, for example, that the population in A was lowered to the level of the worse-off in A+, and then a new group was added at the level of, and the same size as, the best-off group in A+. From an abstract perspective, the resulting world, call it A′, would be *identical* to A+. But, unlike A+, it seems clear A′ would be worse than A regarding Maximin.

The difference between the two cases is obvious. While the worst-off in A actually do fare worse in the move from A to A′, this isn't so in the case of mere addition. Hence, on MC, only the former move would be condemned.

As one might expect, Maximin will be deeply intransitive according to MC. Compare A to B. A is better regarding Maximin.[15] Next, compare B to A+. Clearly, as Parfit describes the case, B is better regarding Maximin. It follows that if "better than with respect to Maximin" were transitive, A would be better than A+. But, on MC, it is not the case that A+ is worse than A. Hence, on MC, (1) must be rejected, and Maximin is deeply intransitive.

On MC, as with EC, analogous considerations could be presented for rejecting (2)–(7). But let us not pursue this here.

Let me next address the view that the mere addition of extra lives does not make an outcome better. Following our earlier usage, let us call this the view that *Utility is Comparative* (UC). On UC, A+ is not better than A regarding utility.

UC is an example of a *person-affecting* principle (about which I shall say more in Part III).[16] Rejecting familiar impersonal principles, UC maintains (attempts to restore?) an essential connection between the ideal of utility and our concern with how people fare. On UC, it is not important that there merely *be* lots of utility, but that those who exist fare as well as possible. Thus, on UC, just as one doesn't improve a situation merely by adding new people, so one cannot make up for losses to those who exist merely by adding new people. As alluded to in section A, one obvious attraction of UC is that it values utility in a way that does not entail the Repugnant Conclusion.

These considerations suggest the following. First, on UC, A is better than B. Loss in people's utility cannot be made up for merely by adding more people.[17] Second, B is better than A+. Loss in some people's utility *can* be made up for by sufficient increases in the utility of others who

[15] As we shall see in Part III, there are some cases where a situation like A might not be better than one like B regarding Maximin. But this is not so in Parfit's example, as he describes the relations between them.

[16] Strictly speaking this is not quite right. As stated, UC is compatible with both person-affecting and non-person-affecting principles. However, for reasons I shall not pursue here, I think the non-person-affecting interpretations of UC are ultimately untenable. In any event, I think UC is best supported by, and hence best interpreted as, a person-affecting principle, and that is how I shall be interpreting it in this article.

[17] I am here assuming that A is our starting point, and that the move from A to B affects

already exist. Third, A+ is neither better nor worse than A. Mere addition doesn't improve the utility of those who exist, though it doesn't worsen it either. From this it follows that (1) must be rejected, and hence, that utility is deeply intransitive according to UC. Again, let me add, without showing here, that on UC most of the other inferences (2)–(7) must similarly be rejected.

F

We have seen that inequality is deeply intransitive given EC. Maximin and utility will also be deeply intransitive given MC and UC, respectively. Reflection suggests that the intransitivity of any one of these notions will carry over into our judgments of preferability.

Consider, for example, the case of inequality. Inequality isn't all we care about, nor even, perhaps, what we most care about; but, for many, it is *one* important element of our judgments of preferability. Thus, how situations compare regarding inequality may determine how they compare regarding preferability if "other things are equal," or at least "equal enough." But then, if inequality is deeply intransitive, it seems likely there are bound to be *some* situations which are equivalent, or nearly equivalent, in terms of the other ideals we care about such that the deep intransitivity of inequality will be carried over into our judgments of preferability.

An analogous argument could be made in terms of the deep intransitivity of Maximin or utility. Indeed, the point is generalizable. *If an important aspect of a complex notion is deeply intransitive, the notion itself will be deeply intransitive.*

G

Together, the preceding sections suggest the following "solution" to the Mere Addition Paradox. Parfit addressed the Mere Addition Paradox to the many who shared his beliefs that:

(a) The mere existence of extra lives worth living does not make an outcome better (UC),

the A-people adversely. This alters Parfit's example, since he assumes that B would evolve from A over the course of several generations, so those in A would not be affected for the worse. My assumption enables me to straightforwardly apply UC to the case in question, and, most importantly, to clearly show its deeply intransitive nature.

(b) while inequality is normally a bad feature, it does not make an outcome worse if it involves the mere addition of extra people who have lives worth living (EC), and

(c) Maximin is not plausibly applied in cases of mere addition (MC).

But the Mere Addition Paradox is only *paradoxical* if it genuinely involves three inconsistent beliefs. So, the question is, are:

(1) A is better than B, and

(2) B is better than A+, and

(3) A+ is not worse than A

truly inconsistent? Parfit thought they were, because, like most people, he assumed, indeed emphasized, that "better than" is always transitive.[18] However, our argument demonstrates that, if EC is true, "better than" isn't always transitive. Indeed, not even "all things considered better than" is transitive, and similarly, if UC or MC is true. So, if (a)–(c) are acceptable, there is room to believe, as many are inclined to, that (1), (2), and (3) are each true, for, as we have seen, (1) and (2) would not entail the falsity of (3).[19]

So far, our response to the Mere Addition Paradox has been rather abstract and formal. Let me next add some flesh to the argument by noting directly how our theoretical results apply to (1), (2), and (3).

In comparing Parfit's situations, several factors influence our judg-

[18] Cf. pp. 166 and 168 of "Future Generations," and p. 432 of *Reasons and Persons* where the analogous claim is made for "worse than."

[19] Some people think the Mere Addition Paradox trades on a confusion about different situations. On this view, the situation depicted by A+ when A is the only alternative to it, *is a different situation* than the situation depicted by A+ when B is an alternative to it. Hence, there isn't any intransitivity, as A can be better than B, and B better than A+ ', and A+ " not worse than A. Parfit explicitly considers this position in *Reasons and Persons*, pp. 428–29, and he offers powerful arguments against it. However, his arguments depend on the claim that "The relative goodness of . . . two outcomes cannot depend on whether a third outcome, that will never happen, might have happened" (p. 429), and this, in turn, may depend on the *Independence of Irrelevant Alternatives Principle* (IIAP), about which I shall say more in Parts II and III. As we shall see, I think one *can* preserve transitivity in the cases we are discussing if one rejects IIAP, and I think one can reject IIAP if one adopts the view that certain factors are essentially comparative. However, such a move not only has numerous implausible implications, it faces grave practical and theoretical problems which undermine much of the *point* of preserving transitivity. I cannot pursue here the full implications of adopting the position in question, but some of its main costs will be presented in Part III.

ments of preferability. For many, these factors include equality (E), utility (U), and Maximin (M), where these are to be understood in terms of, or consistent with, EC, UC, and MC, respectively. For most, they also include *perfectionism* (P).

Espoused in different forms by Aristotle and Nietzsche, P has received relatively little attention from philosophers of late. Nevertheless, in thinking about Parfit's situations, it is difficult not to be strongly pulled toward some version of P as *one* ideal, among others, deserving of value. For our purposes, let us say that according to P, A is better than B if *some* of A's members are better off, or lead fuller, richer lives, than the members of B.[20]

On reflection I think it is primarily the four factors mentioned that play a significant role in comparing these situations, and throughout the remainder of this article I shall assume this to be so.[21] Those who think other factors relevant must amend my remarks accordingly. However, even if some emendation is in order, my main results will stand.

Comparing A+ with A, we see that according to EC, UC, and MC, A+ is not worse than A regarding E, U, and M, respectively. Moreover, A+ is not worse than A regarding P (neither is it better). Thus, there is good reason to hold that A+ is not worse than A all things considered. Comparing B with A+, we see that B is worse than A+ regarding P, but better regarding E, U, and M. Although we care about P, few think it more important than E, U, and M combined. Hence, it isn't surprising that many will think this a case where the *one* respect in which B is worse than A+ is outweighed by the *three* respects in which it is better.[22] Finally, comparing B with A, we see that B is worse than A regarding P, and is *not* better than A regarding E, M, or even U. To the contrary, B is only as good as A regarding E, and, according to MC and UC, it is worse than A regarding M and U. So, there is good reason to hold that B is worse than A all things considered.

Our earlier argument established that on EC, UC, and MC, (1) and

[20] This is a simplified version of P. But it suffices for my present purposes, and is not too unlike the versions of P many are drawn to in thinking about Parfit's alternatives.

[21] Some invoke a principle like average utilitarianism in comparing Parfit's situations. But I think there are powerful reasons to reject "average" principles. Parfit gives some of these in *Reasons and Persons*, pp. 420–22, and I present others in my *Inequality*.

[22] This "one" to "three" point is slightly misleading. If one situation is *much* worse than another in one respect, and only a *little* better in the other respects, the second might be preferable to the first. But, as they are presented, this isn't the situation obtaining between B and A+.

(2) do not entail the rejection of (3). We now see that not only are these positions *theoretically* compatible, there is every reason for people to hold them, given their beliefs.

It is worth emphasizing that in an important sense the extent to which we care about the different principles remains the same in each comparison. It is not as if we care about Maximin when comparing A with B, but not when comparing A+ with A. Nor do we care more about perfection when comparing A with B, than when comparing B with A+. In each comparison, our commitment to the different principles remains unchanged, but their impact on our judgment varies with the alternatives being compared. To be sure, at one level, it may seem that the relevant and significant factors appealed to by the different principles change with the alternatives being compared. However, there needn't be anything inconsistent about this. Instead, advocates of EC, UC, and MC could contend, this merely expresses what we really are (and should be) concerned with regarding these principles, including the conditions under which that concern is properly evoked.

Let us detail the difference between Parfit's view of the paradox and mine. Parfit saw that (1), (2), and (3) were each very plausible. He also saw they were inconsistent if preferability was transitive. But he believed that preferability *must be* transitive, hence he did not think that the plausibility of (1), (2), and (3) could seriously threaten that position. Correspondingly, he thought one of (1)–(3) had to be given up, and the paradox was that each seemed plausible, and there seemed to be no easy way to give up one rather than another.

On my view, once we see that (a)–(c) underlie the plausibility of (1)–(3), the "paradox" takes on a new light. To preserve the transitivity of preferability it is not simply sufficient to reject one of (1)–(3). One must reject *all* of (a)–(c). Correspondingly, *if* one wants to preserve the transitivity of preferability, there is no longer a question about which of (1)–(3) must be given up. Insofar as UC in fact underlies our judgment about (1), and UC, EC, and MC in fact underlie our judgment about (3), both (1) *and* (3) will have to be given up. Finally, and most importantly, once we see the plausibility and nature of UC, EC, and MC, the transitivity of preferability no longer seems inviolable. It is not as if the threat to transitivity merely arises from the apparent plausibility of three inconsistent beliefs. It arises directly from the plausibility of each of UC, EC, and MC. Thus, significantly, one does not need to consider (1), (2), and (3) to feel serious doubts about transitivity. Consideration of the factors underlying our judgment regarding (3), *alone*, is sufficient to raise such doubts.

Initially, one of the most surprising results of Parfit's "paradox" is the apparent ease with which it is able to call into doubt such a deeply held belief, as the belief that preferability is transitive. Yet, on reflection, perhaps this is not so surprising. In the past, it was difficult to imagine what sort of reason might even count against the transitivity of preferability. However, once we recognize the nature and plausibility of essentially comparative factors, we see not only that there may be good reason to give up our previous belief, but that doing so *may* no longer seem so deeply perplexing.[23]

H

Throughout Part I, I have taken Parfit's own position as my starting point. The Mere Addition Paradox was presented as a paradox for people who share, with Parfit, certain beliefs. We have seen that on the beliefs in question the Mere Addition Paradox is not a paradox. Instead, we are left with a conclusion which, despite the closing remarks of the previous section, will strike many as even more startling and worrisome: "better than," even "all things considered better than," is not always transitive.[24]

Most react strongly to the suggestion that preferability might be intransitive. Some believe that, if true, it threatens to undermine large parts of, not only morality, but all of practical reasoning.[25] Some believe this

[23] As will become apparent in Part III, in saying this, I am not denying that giving up the belief in question may have tremendously important, and radical, implications. My point is rather that once we see what essentially comparative views involve, we see how such views might clearly and straightforwardly undermine the transitivity of preferability.

[24] Although throughout Part I, I have focused on responding to the Mere Addition Paradox, similar considerations could be presented against arguments for the Repugnant Conclusion. *However* A and Z compare to some intermediate world, or set of worlds, this does not entail how *they* compare if preferability is deeply intransitive. And, when one considers how A and Z compare directly, it is evident that Z *is* worse than A given UC, EC, and MC. The full details of this argument are interesting, as are certain variations of it. Unfortunately, they cannot be provided here.

[25] The temptation is to believe that the suggestion *must* be false if the threat to morality and practical reasoning is sufficiently grave. But as Hume warns us, "There is no method of reasoning more common, and yet none more blameable, than, in philosophical disputes, to endeavour the refutation of any hypothesis, by a pretence of its dangerous consequences to religion and morality. When any opinion leads to absurdities it is certainly false; but it is not certain that an opinion is false, because it is of dangerous consequence." (*An Enquiry Concerning Human Understanding*, Section VIII, "Of Liberty and Necessity," Part II, first two sentences.)

because they think our very concept of rationality is intimately bound up with the notion that preferability *must* be transitive.

Some believe the suggestion borders on the self-contradictory. Even in this post-Quinian era, the reaction of many able philosophers is that it is virtually part of the *meanings* of the words that "all things considered better than" is transitive. Some claim they would not understand what was being said if someone claimed that, *all things considered*, A was better than B, and B was better than C, but C was better than A. To even make such a claim is, on this view, sufficient to establish that the words "all things considered better than" are being misused.

Finally, some believe that even if the suggestion is not self-contradictory, or a major threat to practical reasoning, it is wildly implausible. More particularly, they believe that no matter how plausible the positions leading to such a conclusion *may* have appeared, their denial is more plausible than the alternative of accepting that preferability is intransitive.

It should now be clearer why I believe that even if Parfit's arguments don't establish what *he* thought they did, they strike at the core of some deeply held beliefs.

II

In Part I, we saw that certain beliefs, shared by Parfit and others, threaten the transitivity of both "better than" and "all things considered better than." But most, including Parfit, firmly believe "better than" and "all things considered better than" are *not*, and perhaps *cannot* be, intransitive. Thus, if forced to choose between the latter beliefs, and those entailing their rejection, I suspect most would, at least initially, opt for the latter beliefs. In this part, I shall examine how such a choice would affect the Mere Addition Paradox and the Repugnant Conclusion.

A

The most natural way of avoiding the conclusions that "better than" and "all things considered better than" are intransitive, is to deny that concepts can be comparative in the manner suggested by EC, MC, and UC. Drawing on our earlier exposition, let us be clear about what such a position involves, focusing discussion on the normative realm.

On the view that moral concepts are not comparative, how good or bad a situation is regarding some factor, f, will be an *intrinsic* feature of that situation—that is, it will not depend on the alternative that situation is compared with, but solely on features internal to the situation. On this view—henceforth, the *Intrinsic Aspect* view, or "IA" for short—how a situation has come about, or who its members are, will be irrelevant to the abstract, impersonal judgment about how it fares regarding f. Thus, for example, on IA, how good a situation is regarding inequality depends solely on how its members fare relative to *each other*.

One natural way of representing IA is in terms of a numerical model. On such a model, each situation will merit a "score" representing how good that situation is all things considered, where *that* score will be a function of other "scores" for each factor relevant to preferability (for instance, inequality, utility, perfection, and so on). Naturally, scores will be based solely on the internal features of the situation, hence alternatives with the same internal features will be assigned the same scores, whatever their origins, or comparative alternatives. For complete concepts, perhaps a precise number could be assigned, at least in principle, for each situation, such that the better the situation was with respect to the concept, the larger the number it received. For incomplete concepts such precision would be impossible even in principle, but perhaps a rough range of numbers could be assigned along similar lines.

Reflection on the numerical model reveals that IA accords with the beliefs most have shared regarding (1)–(7) of Part I—that is, that complete concepts are fully transitive, and that even if some, incomplete, concepts are only partially transitive, none are deeply intransitive. More particularly, on the numerical model one can easily see that IA can capture the deeply held beliefs that "better than" and "all things considered better than" are transitive.

I believe that IA, or something close to it, may also lie at the heart of another deeply held belief. Let me briefly present it here, and come back to it in Part III.

Arrow's Impossibility Theorem invokes an *Independence of Irrelevant Alternatives Principle* which has been the subject of much scrutiny and criticism. Whatever the merits of the principle as Arrow presents it, one version of such a principle seems almost overwhelmingly compelling. It might be put as follows. For any two situations, A and B, to know how A compares to B all things considered, it is, at least in principle, sufficient to compare them directly in terms of each of the factors we care about. In such circumstances, knowing how A or B compare to other alternatives would be unnecessary, and indeed, completely irrelevant to knowing how A and B compare.

Understood in such a way, an Independence of Irrelevant Alternatives Principle has great plausibility. Moreover, while Arrow himself probably had in mind that the comparison between A and B was in a sense essentially comparative, for many, I suspect, the core of the position's plausibility is IA.[26] The point is that *any* alternative will be irrelevant to how A and B compare, because how they compare will depend solely on how good *they* are all things considered, and, on IA, this will depend solely on the internal features of A and B.

So far, I have been merely explicating what is involved in the rejection of the view that concepts can be comparative in the manner of EC, MC, and UC, and the corresponding adoption of IA. It follows from what has been said that, on IA, we are mistaken if, in considering the Mere Addition Paradox, we allow our comparative judgments about A, A+, and B, to be influenced by Parfit's claims about what would be involved in moving from one to another. What Parfit is presumably interested in is our abstract, impartial, static judgments about how such situations compare all things considered. But, on IA, *those* judgments can only be based on the internal features of the situations.

Consider, for example, the following diagram.

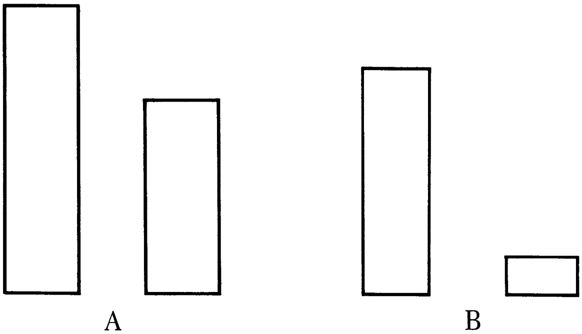

DIAGRAM 3

[26] For Arrow, the comparison between A and B could not be achieved by combining evaluations of A and B arrived at independently of each other. However, he also thought comparisons with other alternatives completely irrelevant to the comparison between A and B. Hence, for Arrow, to know how A compared to B it was both necessary, in an essentially comparative sense, and sufficient to compare them directly.

Assuming those in A and B are equally deserving, and that B is no better than A regarding factors like freedom, virtue, or rights, A would be better than B regarding preferability, as it is better regarding U, E, P, and M. Of course, if one's mother would be in the worse-off group in A, but the better-off group in B, this might give one reason to prefer B to A, perhaps even a duty to promote B rather than A, depending on one's agent-relative duties, if any. But this would not make B better than A from the standpoint Parfit is interested in. Similarly, if the only way of bringing about A would be to lie or cheat, that might be relevant to what ought to be *done*, and even to whether A *together with its history*, would be better than B *together with its history*. But it would not be relevant to how A and B themselves compare, and it is *such* judgments Parfit asks us to make.

Thus, *on IA*, we and Parfit have gravely erred insofar as we have allowed our judgment about A and A+ to be influenced—as most of us assuredly have—by Parfit's claim, indeed emphasis, that the move from A to A+ involves "mere addition."[27] On IA we must close our ears to such information, which may be relevant to the permissibility of such a move, but not to how A and A+ themselves compare.

B

We are now in a position to consider what effect, if any, the acceptance of IA has on the Mere Addition Paradox. Employing the numerical model of IA, I shall argue that the Mere Addition Paradox is much less intractable than it initially appeared, and also, that insofar as it remains problematic, it illustrates familiar problems, rather than new ones previously unrecognized. Let me emphasize that throughout my discussion I am considering what follows if one accepts IA. I am neither contending, nor assuming, that IA is ultimately defensible.

In Part I, we saw that on EC, UC, and MC, preferability would not be transitive, and hence, that it would not only be difficult, but mistaken, to give up any of the three judgments comprising the Mere Addition Paradox. On IA, matters are considerably different. As we have seen, "all things considered better than" would be transitive. So, as Parfit first thought, one could not consistently maintain that all things considered, A+ is not worse than A, though it is worse than B, which is worse than

[27] It does not follow that Parfit was mistaken in emphasizing the mere addition, since his argument, in essence, challenges IA. But if, on reflection, one accepts IA, one cannot be influenced by such information.

A. However, once one accepts IA, it seems there would be much less difficulty giving up one of the judgments. In particular, while accepting IA would not affect our judgment about A+ and B, it could easily cause revision in one of the other judgments.

Consider, for example, the judgment that A+ is not worse than A. Accepting IA would not affect our judgment that A+ is neither worse nor better than A regarding perfection. However, it *would* affect our judgments regarding utility, equality, and Maximin. On IA, A+ would be worse than A regarding E, and also worse regarding M. Now admittedly, A+ might be better than A regarding U. Still, one can see how IA leaves room for the judgment that A+ is worse than A. This will be so if the extent to which A+ is better than A regarding U is outweighed by the extent to which it is worse regarding E and M.

Similarly, IA leaves room for the judgment that B is better than A. As we saw, such a judgment would be impossible on EC, UC, and MC, where, I argued, B would be the same as A regarding E, but worse regarding P, U, and M. On IA, on the other hand, although B would only be the same as A regarding E, and worse regarding P and M, it might be better regarding U. Hence, one could judge B better than A all things considered, if one thought the extent to which it was better regarding U outweighed the extent to which it was worse regarding P and M.

Although IA rules out the consistency of the three judgments in the Mere Addition Paradox, it is misleading to think it *forces* us, by logic alone as it were, to give up one of three judgments we find deeply plausible. To the contrary, someone who accepts IA, will no longer find all three judgments plausible.

Consider, for example, the person who rejects EC, UC, and MC, but retains the view that A+ is not worse than A. Unless she grants *no* weight to E or M, she must think there is one respect in which A+ is better, namely U. And, since it is unlikely that the extent to which A+ is better than A regarding U will be *exactly* equal to the extent to which it is worse regarding E and M, we may assume that such a person believes the former outweighs the latter so that A+ is actually better than A—that is, that all things considered it is *good* the extra people are alive, rather than merely being not bad. (This assumption corresponds to many people's actual reactions to A+ and A.) But if she finds this plausible, why should she find it implausible that B is better than A? Why shouldn't she believe that the extent to which B is better than A regarding U outweighs the extent to which it is worse regarding P and M? Indeed, given that there is even more utility in B than in A+, and that B is better than A+ regarding both equality and Maximin, it would only be plausible to suppose that B was worse than A, if one thought that the extent to which B was worse than A regarding perfection, outweighed all the other respects in which B was even better than A+. But of course, if one thought this, there would be no paradox, as it would not seem that B was better than A+.

I believe then, that if one accepts IA, the Mere Addition Paradox should lose its air of intractability. No longer must it seem deeply implausible to give up one of the three judgments. To the contrary, on IA, the very factors convincing one of the plausibility of two of the judgments, will serve to convince one of the implausibility of the third.

These considerations fit well with reactions many have to variations of the Mere Addition Paradox. Consider Diagram 4, where in each case A+ involves, relative to A, the existence of extra people all of whom have lives worth living and who affect no one else, and B stands to A+ in the manner Parfit described.

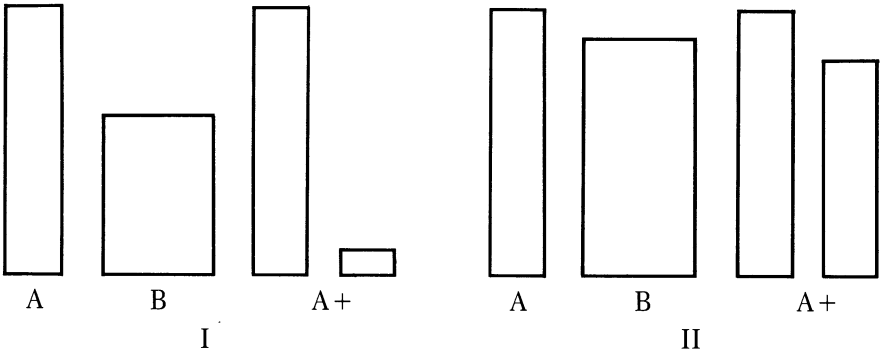

DIAGRAM 4

The difference between Parfit's original case and the above variations is that the worse-off in A+ fare much worse in case I than they do in Parfit's case, while they fare best in case II. Similarly for the B group.

Examining case I, I think most would agree that A is better than B. And many, though certainly not all, would judge that B is better than A+. However, on IA, there is good reason to reject the claim that A+ is not worse than A. A+ is better than A regarding U, but not by much. On the other hand, A+ is *much* worse than A regarding E and M. So, unless one believes that utility matters *so* much more than equality and Maximin that small increases in the one can outweigh large decreases in the others—in which case one would not only have to, but want to, revise one of the other judgments—the judgment that A+ is not worse than A will be implausible.

In case II, I think many, though certainly not all, would continue to judge that B is better than A+. But I also think many would alter their other two judgments. Unlike case I, the extra people in A+ are *very* well off, almost as well off as the very best off. That may seem like a significant improvement regarding U. On the other hand, while A+ is worse than A regarding E and M, it is not much worse. In this case, then, it seems quite plausible that all things considered, A+ is better than A. But, then, on similar reasoning, it no longer seems implausible that B is better than A. To the contrary, in case II, it seems that the small difference between A and B regarding perfection and Maximin, would be outweighed by there being twice as many in B all of whom are *almost* as well off as those in A. Cases I and II suggest the following. IA provides a way of avoiding the Mere Addition Paradox, because, on IA, one can, and indeed must, always reject one of the three judgments comprising the paradox. But this does not mean that in all cases of the sort Parfit has imagined the same judgment must be rejected. Which judgment should be rejected will depend on how the situations actually compare regarding the ideals we value, together with how much these ideals matter in relation to each other. But, of course, this is precisely as it should be.

Now in some cases, like I and II perhaps, it may be obvious which judgment should be rejected. However, in other cases, it may not be possible to simply look at a diagram of the sort we have been considering and "read off" which judgments should be kept, and which rejected. Perhaps by drawing his diagram the way he did, Parfit, in essence, presented such a case. That is, even accepting IA, it may not be evident for Parfit's case which of the three judgments should be given up. But, on IA, there would be nothing surprising, mysterious, or troubling about this, much less deeply inconsistent. Rather, on IA, what Parfit would have given us, despite his protestations to the contrary, is an illustration of a case where our different moral principles so conflict that our all-things-considered view is clouded. Or, more accurately, what Parfit would have given us is a case where we are unsure of how the situations actually compare regarding U, E, P, or M, or of how much these matter in relation to each other.

There are, of course, many problems in deciding how situations compare regarding ideals, and how to weigh ideals against one another. And some of these may strike at the very intelligibility of IA. But these problems are not new with the Mere Addition Paradox. They can arise in comparing just two situations.

To sum up, if one accepts IA, the Mere Addition Paradox will not seem intractable. There will be room to reject at least one of the three judgments. Which one we actually reject will be a function of how we think Parfit's situations compare regarding our ideals, and how much weight we give each ideal. We may feel little difficulty deciding those issues, at least sufficiently to resolve the "paradox." However, if this is not the case, we may be unsure which judgment to give up. This need not reflect inconsistency in our thinking. It may merely reflect the deep difficulties involved in measuring and balancing the things that matter—an important, but familiar problem.

C

Before going on, let me note the following. Some people do not find the Mere Addition Paradox problematic. What they find problematic is iterations of the Paradox which seemingly lead to the Repugnant Conclusion. So, while they have little difficulty accepting the initial moves from A to A+, to B, to B+, they are confident that *somewhere* between A and Z the line should be drawn, even if they are not sure exactly where.

I believe IA can accommodate this view. So, for example, even if A+ is better than A, and B+ is better than B, F+ might be worse than F. This might be so if one believes, as many do, that inequality matters more at low levels than high levels,[28] or if one holds an analogous belief regarding Maximin. (Since Maximin expresses our special concern for how the worst-off fare, this concern may be heightened the worse off the worst-off are.) On such views the move from F to F+ may be worse regarding E and M than the move from A to A+. Similarly, I believe a plausible view of utility can be developed consistent with IA which would enable one to say that the extent to which F+ is greater than F regarding U may be less than the extent to which A+ is better than A. Any of these factors alone could prevent the slide from A to Z. Combined, they make

[28] Amartya Sen advocates this view in *On Economic Inequality* (Oxford: Clarendon Press, 1973). It is defended in my *Inequality*.

it quite plausible that F+ could be worse than F, even if A+ is better than A.

I mention the foregoing partly because of the way it corresponds to some people's reactions to iterations of the Mere Addition Paradox, and partly because I think it has implications which will ultimately prove lasting and significant when fully worked out.[29] Unfortunately, these issues must be dealt with elsewhere, as they lay off our central track.

III

The task of this article would be completed if IA were fully acceptable. Unfortunately, however, the specters raised by the Mere Addition Paradox are not easily laid to rest. Despite the intrinsic appeal of IA, and the unwanted implications of its denial, I believe Parfit was onto something fundamentally important in implicitly recognizing that certain factors relevant to our judgments of preferability are essentially comparative. In this final section, I shall present considerations supporting this view. I shall also note some of the responses and problems such a view generates. Unfortunately, a full exploration of these issues lies beyond the scope of this article. However, I can, at least, begin the task, and perhaps convince the reader of its importance.

A

Shortly, I shall reconsider UC and EC. But first, let me consider another, more general, principle: the *Person-Affecting Principle* (PAP). On PAP, one outcome is worse than another only if it affects people for the worse, so, the relevant question for comparing two alternatives is: would the coming about of the one be worse for people than the coming about of the other? According to PAP:

> (1) One situation is worse (or better) than another if there is *someone* for whom it is worse (or better), and *no one* for whom it is better (or worse), but not vice versa,[30] and

[29] For example, I believe one lesson we learn in thinking about IA and the Repugnant Conclusion, is that perhaps the most natural and prevalent way of regarding utility and its relation to preferability must be rejected. I develop this point in a longer manuscript addressing these topics.

[30] The "not vice versa" clause is necessary on the view, advocated by Parfit in Appendix G of *Reasons and Persons*, that causing someone to exist can *benefit* that person, even

(2) One situation *cannot* be worse (or better) than another if there is *no one* for whom it *is* worse (or better).

Though rarely explicitly acknowledged, PAP underlies many arguments in philosophy and economics, and those appealing to it span a broad range of theoretical positions.[31] Moreover, most believe that PAP expresses a deep and important truth.

Derek Parfit has presented an ingenious argument—the *Non-Identity Problem*—which challenges PAP.[32] Nonphilosophers and philosophers alike find the argument perplexing, and most, at least initially, try to undermine it. Few, if any, conclude that PAP should be rejected outright. Instead, even those accepting Parfit's argument point out, rightly, that the most Parfit *establishes* is that there is a limited and fairly peculiar range of cases where PAP does not apply. These are cases where future generations are involved, and, more particularly, cases where one's choices determine *who* comes to be, such that the same people don't *exist* in the alternative situations *to be affected* for the worse. In most cases of moral concern these conditions do not obtain, and for such cases, most contend, PAP remains plausible.

The reactions to Parfit's argument further illustrate the strength and appeal of PAP. Indeed, Parfit himself seems committed to the view that PAP is plausible in cases other than the Non-Identity Problem.[33]

though failing to cause someone to exist harms (and is worse for) no one. Otherwise, PAP would directly imply that in certain cases each of two situations would be better than the other. Still, on reflection, I believe most advocates of PAP would, and should, accept the claim about not harming those we fail to conceive, but reject the claim that causing someone to exist benefits that person. If this is right, as I shall assume in this article, the "not vice versa" clause is otiose.

[31] Among the many positions deriving some of their force from PAP are Locke's theory of acquisition and property rights, Nozick's medical researcher and Wilt Chamberlain examples, the view of some economists that nonpareto optimal situations are irrational and wrong, the view of some that Rawls's Difference Principle (DP) is more plausible than egalitarianism, and the view of others that DP is too egalitarian to be plausible. In addition, I believe PAP underlies many standard objections to rule-utilitarianism, egalitarianism, rights-based, virtue-based, and deontological theories. I argue these claims in "Harmful Goods, Harmless Bads" (unpublished).

[32] Cf. "Future Generations," and *Reasons and Persons*, chap. 16.

[33] Parfit seems to implicitly appeal to PAP, and person-affecting versions of other principles, in the Mere Addition Paradox and its offshoots. See "Future Generations," pp. 158–59, and *Reasons and Persons*, chap. 19. In addition, see his discussion of The Second Paradox in "Overpopulation and the Quality of Life," in *Practical Ethics*, ed. P. Singer (Oxford: Oxford University Press, 1986).

In discussing PAP with other philosophers, it seems clear that many think it expresses the *essence* of morality, or at least that portion of morality concerned with outcomes. Even if this is too strong, it is hard to deny that PAP is an extremely plausible principle that is at least relevant, if not dominant, in our assessment of (many) outcomes.

Accepting the significance of PAP has important implications. PAP is essentially pairwise comparative. Its content cannot be captured by an Intrinsic Aspect view.

Consider the following diagram.

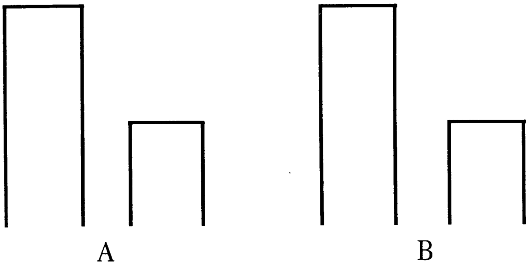

Diagram 5

If the people in A were completely different from those in B, then PAP would yield no reason to prefer one or the other. However, suppose the same people would be in the better-off group in A and the worse-off group in B, while different people would be in the other two groups. On PAP, A would then be better than B, since the coming about of A rather than B would be better for some and worse for no one. On the other hand, if the same people would be in the worse-off group in A and the better-off group in B, while different people would be in the other two groups, then B would be better than A, on PAP. Clearly, then, on PAP, our judgment about how A and B compare will depend crucially on the identities of the people involved. Correspondingly, the desirability of A will depend not, in the relevant sense required by IA, solely on its internal features, but on both the alternative it represents, and the ones with which it is compared.

Next consider Diagram 6. A, B, and C represent three possible outcomes. Each outcome would contain two of the following groups: the p-group, the q-group, and the r-group. The members of each group remain

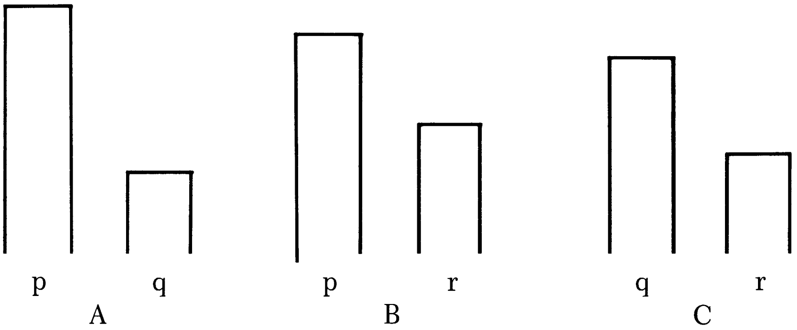

Diagram 6

the same in each alternative in which this group exists. So, for example, the very same people would exist in the p-group in both A and B, though *they* wouldn't exist in C.

On PAP, A would be better than B, as the coming about of A would be better for some (the p-group) and worse for no one. Analogously, B would be better than C. But A is *not* better than C, on PAP. To the contrary, C is better than A, as the coming about of C would be better for some (the q-group) and worse for no one. Clearly, then, "better than" is not transitive on PAP.[34] This in turn threatens the transitivity of preferability, and this is so even if we only regard PAP as *one* important element in our judgments of preferability. After all, as we observed in Part I, if one important aspect of a complex notion is deeply intransitive, the notion itself will be deeply intransitive.

One might attempt to avoid intransitivity by restricting the 'scope of PAP, for example, only allow it to influence our judgments in cases where exactly the same people are involved. However, I believe one cannot plausibly do this and still explain or justify many judgments which seem warranted on person-affecting grounds. More importantly, restricting the scope of PAP opens the possibility that preferability will be intransitive even if none of its aspects are. For example, suppose A, B, and C are three alternatives such that given its restricted scope PAP is only relevant in comparing A and C. It could then be the case that all things consid-

[34] I have learned that Derek Parfit illustrated the possible intransitivity of PAP with a simple three-person case similar to the one noted above in his unpublished "Overpopulation" draft of 1976. I suspect he did not pursue the wider implications of his result, because he thought, on independent grounds, that PAP was implausible for the cases with which he was primarily concerned.

ered—that is, in terms of *all* of the relevant and significant factors for making *each* comparison—A is better than B, and B is better than C, yet C is better than A. After all, even if C is worse than A in terms of the factors relevant for comparing A with B, and B with C, the extent to which this is so might be outweighed by the extent to which C is better than A regarding PAP.

The above point is generalizable and significant. If the scope of a moral principle or ideal is restricted, such that it applies when comparing some situations but not others, then different factors may be relevant and significant in comparing alternative situations. If this is so, then our judgments of preferability may be deeply intransitive *even if none of its aspects are themselves deeply intransitive.* I shall return to this point below in discussing Maximin.

To sum up, PAP expresses an important, widely accepted view whose nature is essentially pairwise comparative. Thus, it appears one must look beyond IA to plausibly avoid the threat of intransitivity.

B

Let us briefly reconsider UC in the context of the foregoing. In Part I, I wrote that UC "maintains (attempts to restore?) an essential connection between the ideal of utility and our concern with how people fare. On UC it is not important that there merely *be* lots of utility, but that those who exist fare as well as possible." Jan Narveson defends UC, though not by that name. He writes:

> Morality has to do with how we treat whatever people there are. . . . [We] do not . . . think that happiness is intrinsically good [in the way required by an impersonal total view]. We are in favor of making people happy, but neutral about making happy people.[35]

On reflection, most people are attracted to Narveson's position. This is not surprising in light of the previous section. UC is a *person-affecting* version of utility, according to which "the principle of utility requires that before we have a moral reason for doing something, it must be because of a change in the happiness [or utility] of some of the affected persons."[36] Thus, UC is able to accommodate people's concern for utility in a way

[35] "Moral Problems of Population," *The Monist* 57, no. 1 (Jan. 1973):73, 80.

[36] This quotation is from Narveson's pioneering work "Utilitarianism and New Generations," *Mind* 76 (Jan. 1967):67.

consistent with their more general, and perhaps fundamental, concern about the extent to which people are affected for better or worse.

In the end, UC may not *fully* capture our views about utility. For example, in the Non-Identity Problem, UC may need to be supplemented, or another principle invoked, to accommodate our beliefs. Still, it seems that for many cases, UC accurately expresses people's concerns regarding utility, and hence will play a significant role in their judgments of preferability. But, as we saw in Part I, UC is essentially pairwise comparative. The threat of intransitivity remains.

C

Let us next reconsider EC. Some people claim that in a world where half are blind, there is *no reason at all* to put out the eyes of those with sight. Although such claims are often adduced by nonegalitarians as "proof" that there is nothing intrinsically valuable about equality, there are various ways an egalitarian might respond. One way is to adopt EC, interpreted as a person-affecting version of egalitarianism. On such a view, to care about inequality is to care about the worse-off faring less well than they otherwise would because of the inequality. So, on EC, concern about inequality is not merely concern that inequality be removed, but that it be removed in a certain way, so as to *benefit* those worse off.

For many, EC best expresses their egalitarian concerns. It enables them to distinguish between (a) cases where the plight of the worse-off is partly a function of the inequality—were the inequality removed in an appropriate way they would be better off, and (b) cases where the plight of the worse-off is not a function of the inequality—were the inequality removed they would not benefit, or perhaps not exist. Such a position can grant the force of many nonegalitarian claims, without succumbing to them. It can also capture the view of Parfit and others about the case of mere addition.

Like PAP and UC, EC is essentially pairwise comparative. For example, whether the inequality in a world where half are blind is morally regrettable, will depend on whether the alternative is a sighted world or a blind one. Correspondingly, like PAP and UC, EC will be deeply intransitive.[37]

[37] This is easily seen. For example, the foregoing consideration directly implies that the principle of substitution for equivalence must be rejected if one believes, as seems plausible, that other things equal, a situation where everyone has sight will be equivalent to one where everyone is blind *regarding inequality*. (They will not, of course, be equivalent "all things considered.")

Let me note another reason some are attracted to EC. Consider Diagram 7.

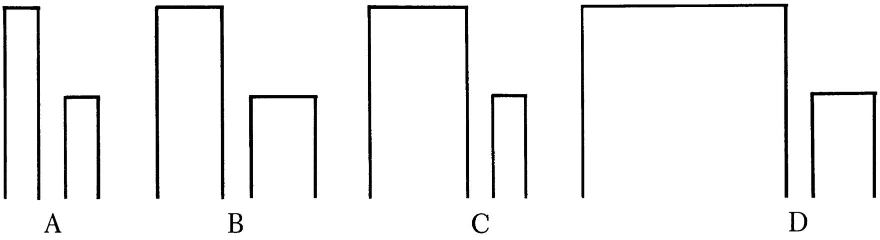

DIAGRAM 7

Most believe proportional changes in population size do not affect inequality. So, most would hold that since the *pattern* of inequality is identical in A and B, they are equivalent regarding inequality. This view is shared by most economists, and, to my knowledge, is yielded by every measure of inequality economists have offered.[38]

Now compare B and C. Elsewhere, I have argued that in some respects B is better than C regarding inequality, while in other respects C is better than A.[39] However, all things considered, I think most would agree that B's inequality is worse than C's, that is, that B's inequality would be improved if most of the worse-off were raised to the level of the better-off.

On an Intrinsic Aspect view, the previous reasoning suggests that D's inequality would be better than B's. After all, since the pattern of inequality is identical in D and C, D would be equivalent to C which is better than B. However, I think most would agree that whether D was better than B regarding inequality might depend on who its members were or how it came about. If B were transformed into C by raising most of the worse-off, and then D resulted from proportional increases in C's populations, then most might agree that D's inequality is better than B's. However, if D resulted from B via mere increases in the population of

[38] Among the measures of inequality yielding this result are the range, the relative mean deviation, the variance, the coefficient of variation, the standard deviation of the logarithm, and the gini coefficient. Also, Dalton's measure, Atkinson's measure, and Theil's entropy measure.

[39] "Inequality," *Philosophy & Public Affairs* 15, no. 2 (Spring 1986).

B's better-off group, I think most would rightly reject the claim that D was better than B regarding inequality.

The previous observation might be usefully compared with Parfit's regarding mere addition. Some egalitarians believe A+ is worse than A regarding inequality, and hence that, contra Parfit, mere addition *can worsen* inequality. My claim is that mere addition cannot *improve* inequality. For the better-off in B to transform their world into one like D by having more children would be for them to do *nothing at all* in terms of improving B's inequality. To the contrary, instead of improving the position of the worse-off relative to the better-off, as the egalitarian desires, such action would only make it the case that there were even *more* people whom the worse-off were worse off than.

An Intrinsic Aspect view cannot plausibly accommodate both the widely accepted view about proportional increases, and the view that mere addition cannot improve inequality. Insofar as these views are hard to give up, there is further support for EC, and hence further reason to look beyond IA if one hopes to preserve the transitivity of preferability.

D

Let us next consider the nature of Maximin. According to Rawls, a situation would be

> (1) "perfectly just" if "no changes in the expectations of those better off . . . [could] improve the situation of those worst off,"
>
> (2) "just throughout, but not the best just arrangement" if "the expectations of all those better off at least contribute to the welfare of the more unfortunate. . . . [so] if their expectations were decreased, the prospects of the least advantaged would likewise fall. . . . [yet] even higher expectations for the more advantaged would raise the expectations of those in the lowest position," and
>
> (3) "unjust" if "the higher expectations . . . [of the better off] are excessive. . . . [such that if] these expectations were decreased, the situation of the least favored would improve."[40]

For Rawls, then, how just a situation is does not depend on the absolute level of the worst-off group, or on how the worst-off fare relative to the

[40] *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), pp. 78–79.

others; it depends on the other alternatives available to it—specifically, on whether a change in its principles and institutions would improve the expectations of the worst-off group. This suggests that Maximin is essentially comparative, and cannot plausibly be captured on an Intrinsic Aspect view.[41]

In discussing Maximin, it is important to distinguish between different kinds of cases. Most Rawlsians believe Maximin can be straightforwardly applied when comparing *direct alternatives*—situations which could be transformed into each other via appropriate changes in their principles and institutions. Direct pairwise comparisons seem sufficient for ranking direct alternatives. The best situation will simply be the one in which the worst-off group fares best.

However, some situations are not direct alternatives in the relevant sense. They cannot be compared directly in terms of how well off their worst-off fare. This point is important, but easily, and too often, overlooked.

If A is a poor society whose principles and institutions are such that "no changes in the expectations of those better off . . . [could] improve the situation of those worst off," and B is a rich society which could greatly improve the expectations of the worst-off by (slightly) reducing those of the best-off, then according to Rawls A would be perfectly just, while B would be (grossly) unjust. Moreover, this would be so even if the worst-off fared better in B than in A. Indeed, even if *everyone* in B was better off than *everyone* in A, while B might be better than A all things considered, it would *not* be better regarding Maximin. After all, Rawls has offered a theory of *justice*, and his theory plausibly recognizes that one society might be less just than another, though in absolute terms its members fare better.

The preceding suggests that although Maximin is essentially comparative, it is not always pairwise comparative in the manner of PAP, UC, and EC. While in some cases it may be sufficient to compare situations directly in terms how well off their worst-off fare, in others it is not.

Next consider Diagram 8.

[41] To be sure, there is a sense in which the alternatives available to a situation partly depend on its internal features. But they also partly depend on outside factors. Moreover, on IA, the fact that a given situation *could be* transformed into some other situation with a different pattern of distribution should be irrelevant to how good or bad *it* is, considered by itself. This is why such a position is clearly unacceptable for capturing Rawls's view.

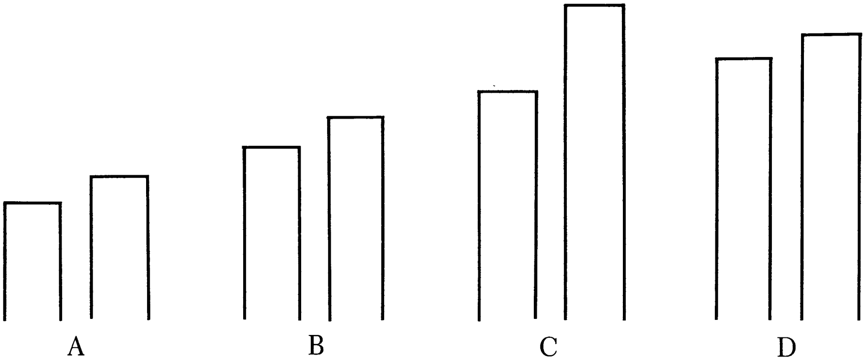

DIAGRAM 8

Suppose that B occupies a sort of midway position between A and C, such that while A and C might each be transformed into B, and vice versa, A cannot be transformed into C.[42] Suppose further that decreases in the expectations of the better off in A would worsen the expectations of the worse-off. Finally, suppose that C could be transformed into D, and vice versa.

Since A and B are direct alternatives in the sense noted above, most Rawlsians would agree that B is better than A regarding Maximin, and similarly, most Rawlsians would agree that C is better than B. However, according to Rawls, C would be *worse* than A, since on the assumptions of the example, C would be unjust, while A would be just throughout, though not perfectly just. (Since A and C are not direct alternatives, it is not sufficient to directly compare how their worse-off fare, instead they must be compared in terms of how their worse-off fare relative to their alternatives. So, while those in A and B could be criticized on grounds of justice for failing to achieve B and C, respectively, those in A could not be similarly criticized for failing to achieve C, since C is not an option for them.)

It appears, then, that given what many Rawlsians believe, Maximin will be deeply intransitive, as the relevant and significant factors for making certain comparisons—for example, those between direct alternatives—will differ from those for making others.

[42] There are various reasons this might be so. The time, effort, and resources required to attain one goal or destination may preclude attaining another, yet leave open the possibility

There are various ways an advocate of Maximin might avoid intransitivity. One way would be to apply Maximin only to situations which are direct alternatives in the sense noted above. Such a move loses touch with the Rawlsian insight that one society might be more unjust than another, though in absolute terms its worst-off group fares better. More significantly, for our purposes, such a move limits Maximin's scope. Hence, for reasons noted in the previous section, such a move won't preserve the transitivity of preferability. At least, this is so if one believes that Maximin is only one aspect of preferability, and that some judgments of preferability would still be meaningful in cases where Maximin does not apply.[43]

Besides contending that Maximin is only plausible when comparing direct alternatives, one might hold that Maximin is only plausible when comparing situations which don't involve future generations, or when comparing situations involving the same people, or when comparing situations involving the same number of people. On any of these views, it will be hard to avoid intransitivity in one's judgments of preferability, since the relevant and significant factors for such judgments will then vary with the situations being compared.

Perhaps I have misinterpreted what many Rawlsians believe. Or perhaps, even if my interpretation is correct, we should adopt an IA version of Maximin, which ranks situations solely on the internal feature of how well off their worst-off fare. As we've seen, such a view is distinctly un-Rawlsian. More importantly, it is implausible. Consider Diagram 9.

Suppose II resulted from I, via groups B, C, and D dying off from disease, and the lives of the A group being adversely affected.[44] Parfit contends it would be *absurd* to say II was better than I, even regarding Maximin. He writes, "How can it be better if all . . . [groups but one] cease to exist, with the result that the survivors would be *much worse off*?" Even if the position is not *absurd*, it seems deeply mistaken. Max-

of returning to one's starting point. Similarly, a society's options can vary with time and place, as well as cultural and technological development.

[43] One way of responding to the purported intransitivity in the above example is to claim that the B which is better than A regarding Maximin is a different alternative than the B which is worse than C. This has a fair amount of appeal, especially given the way we apply Maximin, but it raises the fears and difficulties alluded to in note 16.

[44] This is a simplified version of an ingenious example Parfit presents in *Reasons and Persons*; see "How Only France Survives," p. 421.

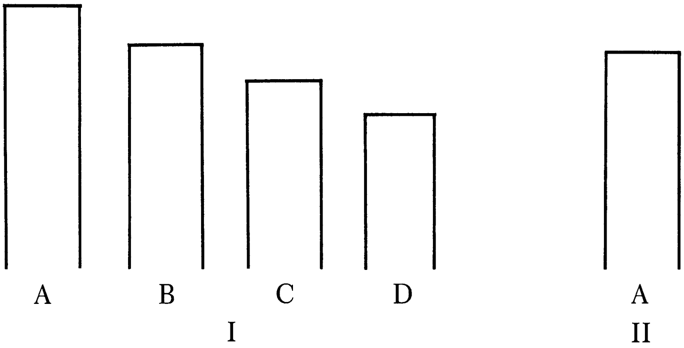

DIAGRAM 9

imin reflects our special concern for those worst-off, and this concern is not alleviated by the prospect of the worst-off *dying* (at least if their lives are worth living).[45]

Next, consider Diagram 10.

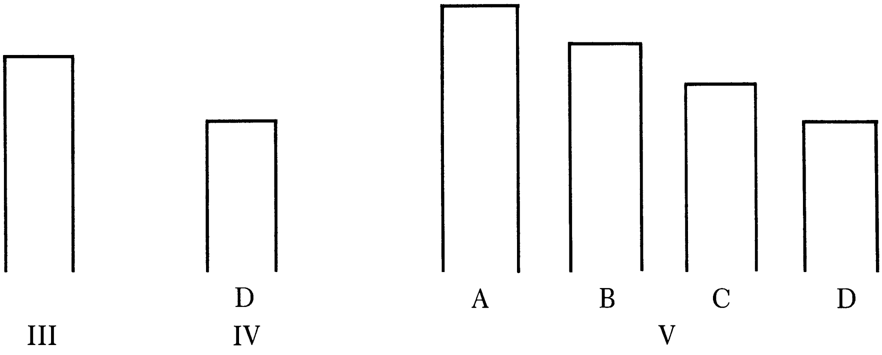

DIAGRAM 10

[45] A holder of both Maximin and IA might claim that II really is better than I regarding Maximin, that perhaps I have been misled by my own example in failing to distinguish between the question of whether II is better than I *regarding Maximin*, and the question of whether II, brought about in a particularly dismaying way, should be regarded as an improvement over I *all things considered*. I deny this. My claim is that such a position does *not* best express what most are concerned about insofar as they care about Maximin. (I am, however, sympathetic with the general tenor of the objection. As I point out in my work on inequality, for *any* ideal it is easy to construct cases where the implications of that ideal seem implausible. This shows *not* that the ideal is implausible, but that morality is complex.)

III is better than IV regarding Maximin.[46] Now suppose V resulted from IV by the *mere addition* of the extra groups A, B, and C. This would not, I think, improve IV regarding Maximin. Just as our concern for the worst-off is not alleviated by the worst-off dying, it is not alleviated by bringing into existence *other* people who are better off than they. But then, if III is better than IV regarding Maximin, it seems that III will also be better than V, if V results from IV in the manner suggested.

Now from the purely abstract perspective II is the same as III, and I is the same as V. So, if IA were appropriate for Maximin, we should think III compared to V as II compared to I. But we don't think this. This is because Maximin expresses a concern for the worst-off which is not fully reflected in a situation's abstract internal features.

In *Reasons and Persons*, Parfit employs an example like Diagram 9 to argue that Maximin is not appropriately applied in different number cases. This entails the crucially important insight that Maximin is essentially comparative, but I think the issue about numbers is a red herring. Consider Diagram 11.

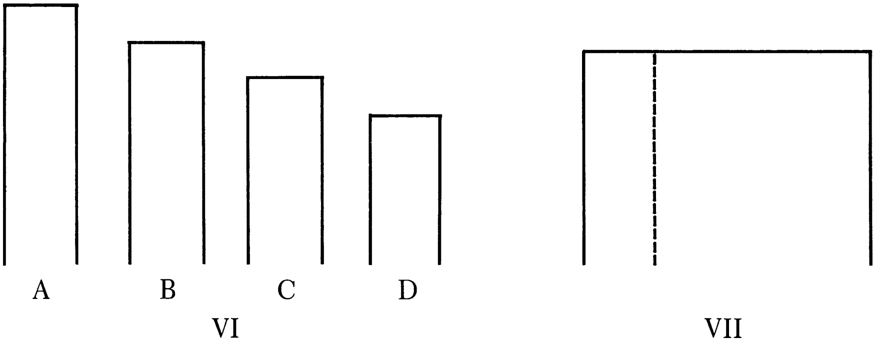

Diagram 11

If VI and VII have the same number of people, most would agree that VII was better than VI regarding Maximin, *if* VII and VI had entirely different populations, or if VII resulted from VI by redistribution. But suppose VII resulted from VI as follows: first, VI was transformed into a

[46] Here, and in what follows, I assume the only relevant alternatives are the ones noted, so we needn't worry about the sort of complications introduced earlier, e.g., that IV might be perfectly just, or just throughout, while III might be unjust.

situation like II, via B, C, and D dying off and A being made worse off,, then, extra people were added so the population was transformed into VII. Brought about *this* way, VII wouldn't be better than VI regarding Maximin, for reasons similar to those presented above. Thus, even where same numbers are involved, judgments about Maximin will not depend solely on the situations' abstract internal features.

Parfit never claims Maximin *is* appropriate in same number cases. Only that it is inappropriate in different number cases. But suppose we could bring about a future state with 8 billion people, or one with 10 billion where the worst-off group would be significantly better off. I think the second situation might be better than the first regarding Maximin, and that that would be *one* reason to prefer it.

I think, then, that the manner and extent to which Maximin applies to two alternatives depends not on their numbers, but on how they are related. But whether I am right about this particular point, or Parfit is, it seems undeniable that Maximin is essentially comparative. Thus, one cannot avoid the threat of intransitivity by simply adopting IA.[47]

[47] Throughout this article, I have mainly focused on general principles and largely theoretical considerations. However, moving from abstract cases to more concrete ones, it may seem there are many situations where the spirit of IA must be rejected—that is, where certain relevant and significant factors for comparing situations are essentially comparative. For example, some believe that for certain positions, greater preference in hiring should be given to blacks over whites, than to blacks over Mexican-Americans, or to Mexican-Americans over whites. Such a policy of affirmative action is justified, they believe, by the particular nature of the relationship between blacks and whites in American society, one which is relevantly and significantly different from the relationships between blacks and Mexican-Americans, and between whites and Mexican-Americans. Similarly, some believe that considerations of familial respect dictate that parents be given some preference over their offspring in certain situations. On either of these views, it seems there could be situations where it would be better all things considered if A were selected rather than B, and B were selected rather than C, yet C were selected rather than A. This might be so if A were white, B were Mexican-American, and C were black, or if C were the parent of A, and B were a stranger. In such cases, a relevant and significant factor which might give C the edge over A, would not apply when comparing A with B, or B with C.

There are, of course, numerous ways one might reject the above claims. Still, it seems that many special relationships give rise to factors that (rightly) influence our judgments of preferability—were the relationships different we would make different judgments. But then, in comparing alternatives affecting different groups, between whom different relationships obtain, it seems the relevant and significant criteria may vary with the alternatives being considered. Hence, the spirit of IA must be rejected.

Some people think examples of the sort presented in this footnote cast doubt on the transitivity of "what we ought to do" but not on the transitivity of "what ought to be the case." My own view is that in some cases, though not all, intransitivity in judgments of

E

Can one avoid intransitivity in one's all-things-considered judgments, once one accepts that certain relevant and significant factors for comparing alternatives are essentially comparative?

Perhaps one could avoid intransitivity via a set of second-order dominance principles, telling us which factors dominate over others when intransitivity threatens. For instance, we might give lexical priority to certain principles, at least for certain cases. I leave this avenue for others to pursue. There are many well-known difficulties with lexical orderings, and I do not see how a set of second-order dominance principles can be arrived at which (a) will not be ad hoc, (b) will plausibly respond to the theoretical difficulties raised by essentially comparative factors, and (c) will not themselves be subject to intransitivity (thus requiring a set of third-order dominance principles, which in turn may be intransitive).

There is another way one might respond to the threat of intransitivity. Let me present it with an analogy from sports. In baseball, it is perfectly possible that the first-place team consistently beats the second-place team, which consistently beats the third-place team, which consistently beats the first-place team. Still, most do not think that "better than" is intransitive regarding baseball teams (perhaps they should?). Instead, most think the better of two teams is the one that wins the most games against *all* the other teams during the season. Thus, for baseball teams, "all things considered better than" remains transitive, notwithstanding the intransitivity of "consistently beats," for, if A has more total wins than B, and B more than C, A will have more than C, regardless of their records against each other.

One might apply a similar model to the judgments we've been dis-

the former sort will carry over into judgments of the latter sort. This is because I regard right and wrong acts as relevant to, though not determinant of, the goodness of outcomes. Either way, I find such examples troubling. Perhaps, if it was better to do A than B, B than C, and C than A, we should rest content in the knowledge that any of the three acts would be (equally) right. However, to me, such cases have the feel of the so-called "moral blind alleys" people like Nagel and Williams worry about. In such a case, each of the three acts may seem seriously wrong, since to choose one would be to choose an act which, by hypothesis, seems morally worse than another available to us. (I owe the parent/offspring example noted above to an unpublished note which Ronald Dworkin sent to Parfit, and which Parfit showed me after reading an earlier version of this work. Frances Myrna Kamm presents two similar examples on p. 137 of "Supererogation and Obligation," *Journal of Philosophy* 82, no. 3, [March 1985].)

cussing. In Part I, we learned that on essentially comparative views, to know how two situations compare all things considered, it is not sufficient to know, even precisely, how they compare to some third situation. This model endorses that lesson, but extends it significantly. To know how two situations compare all things considered, it is also not sufficient to know, even precisely, how they compare directly. Rather, how they compare all things considered will be a function both of how they compare to each other *and* how they compare to other situations.

On this model, then, one can grant that there are essentially comparative factors, and hence, that in terms of the relevant factors for making each *particular* comparison, A is better than B, B better than C, and C better than A, yet deny that these particular judgments are "all things considered," and hence deny that preferability is intransitive.

The position sketched has obvious attractions. But is the analogy with baseball ultimately plausible? In baseball, there is a small, fixed, conventionally agreed upon set of alternatives with which each team is to be compared—that is, the other teams currently in the league. The situation is otherwise regarding most questions of preferability, and this raises both practical and theoretical problems, some of which are noted below.

Consider a simple case, where n people have applied for a job. On the old, standard way of judging candidates, if the first candidate was better than the second considering each of the factors relevant to comparing them directly, the first would be regarded as better all things considered, and the second could be removed from further consideration. Proceeding in this way, one would theoretically only need to make n − 1 judgments to determine the best candidate.

On the baseball analogy, to know how two candidates compare all things considered it is not sufficient to know how they compare directly; one must also know how they compare to each of the other candidates. So, to determine the best candidate, each must be compared to every other. One can easily calculate that this would require (n ÷ 2) × (n − 1) judgments.

On the old way, when a Philosophy search committee is swamped with 200 applications for a single job opening, a "mere" 199 separate comparisons are required to determine the best candidate. On the other hand, to compare each of 200 candidates with every other would require *19,900* separate comparisons. The practical impossibility of this will be evident to anyone who has ever served on a search committee!

There is a further problem exacerbating the one just noted. It may be illustrated as follows. Suppose three baseball teams were up for sale. In assessing which to buy, one would definitely *not* restrict one's attention to how the available teams fared against each other in direct competition. (It might be that while A consistently lost to B and C, A was the very best team while B and C were the two worst.) Instead, one would take into consideration how each team fared against *all* the other teams, *including all the teams that were not on the market.* The point has obvious and important implications for the suggestion in question.

Consider, again, the apparently simple case of job candidates. On the standard way of thinking, it is sufficient to compare the credentials of the applicants *themselves* to determine the best applicant. On the baseball analogy, this is no longer plausible. To the contrary, one would need to compare each applicant not only with the other applicants, but also with those who have not applied! This, of course, would be impossible (practically speaking).

Let me note one other problem along these lines. Suppose we want to know how Parfit's A and A+ compare all things considered. Since it isn't enough to compare them directly, what other alternatives should we consider? Parfit presents some they might be compared with: B, B+, C, C+, . . . Z. But he might have presented others. In fact, there are an infinite number of alternatives Parfit *might* have presented, and it is seemingly arbitrary to compare A and A+ to some, but not others.

It may seem, therefore, that to avoid the conclusion that how A and A+ compare all things considered is an arbitrary matter of convention, one must compare them with each possible alternative, and not merely with some selected set of alternatives that we happen, for one reason or another, to be interested in. However, this raises more than practical difficulties. It raises a familiar problem of infinity.

If, comparing each pair of alternatives directly, A+ is better than A, B better than A+, and A better than B, then there will be an infinite number of possible alternatives sufficiently like those, such that A will be better than an infinite number of B-like worlds, and worse than an infinite number of A+-like worlds, and analogously for A+ and B. It seems, then, that on the baseball model there will be nothing to choose between A, A+, and B. Each will be as good as the others, since for every case where

A, A+, or B is better than an alternative, there will be a case where each of the others is better than an alternative, and similarly, for every case where A, A+, or B is worse than an alternative. But, then, if B+ is better than an infinite number of B-like alternatives, it too will be as good as the others, and, for similar reasons, so will C, C+, D, D+, and so on. More generally, every alternative will be both better and worse than an infinite number of alternatives. Hence, on the baseball analogy, there would seemingly be nothing to choose between any two alternatives regarding preferability.

One might alter the baseball analogy to address these problems. Or one might pursue yet another way of reconciling essentially comparative factors with the transitivity of preferability. However, either way, I think one faces an unavoidable problem of which the shortcomings of the baseball analogy are symptomatic.

I suspect IA lies at the core of both the transitivity of preferability, and the Independence of Irrelevant Alternatives Principle (IIAP) in perhaps its most plausible form. So, if we reject IA we should reject them both. Nevertheless, both positions have enormous appeal, and we continue to find them compelling even if we reject IA. Still, once one accepts essentially comparative factors, as we seemingly must, at least one of these positions must be rejected.

On the view that certain factors are essentially comparative, A can be better than B, B better than C, and C better than A, in terms of the relevant and significant factors for making each comparison. It follows that if, in accordance with IIAP, how two situations compare all things considered depends solely on how *they* compare in terms of each of the relevant and significant factors for making *that* comparison, then the judgments in question will be all things considered, and preferability will be intransitive. On the other hand, if preferability *is* transitive, then the judgments in question are not all things considered, contrary to what is implied by IIAP. Thus, once one rejects IA, while both of the positions in question may be false, they can't both be true.

F

The main reason for rejecting IA, and giving up one or both of the positions in question, is simply that it seems almost undeniable that *certain* special relationships and concerns give rise to essentially comparative factors. A further advantage is that, besides providing a response to esoteric problems like the Mere Addition Paradox, it provides a response to common cases of apparent inconsistency. For example, many seemingly rational people express views like the following. Given the choice between tennis and softball, they prefer softball; between softball and opera, they prefer opera; and between opera and tennis, they prefer tennis. In the past, many philosophers and others have felt compelled to insist—lamely and implausibly?—that, all appearances to the contrary, such preferences were *necessarily* misinformed, muddle-headed, inconsistent, and/or irrational.[48] However, if one rejects IA, many apparently inconsistent orderings may be perfectly understandable and rational, since the relevant and significant factors for comparing certain alternatives (for example, two athletic activities), may differ from those for comparing others (an athletic activity with a nonathletic one).

There is another advantage to rejecting IA. Many philosophers and economists have been deeply troubled by Arrow's Impossibility Theorem, according to which, roughly, there can be no decision procedure for arriving at a social ordering among alternatives which simultaneously satisfies certain extremely plausible assumptions. But, Arrow's Theorem invokes IIAP, which, as we have seen, there is good reason to reject if one rejects IA. Hence, by rejecting IA, one is in a position to reject Arrow's Theorem and its corollaries.

To be sure, rejecting IA and IIAP raises new and significant problems regarding decision procedures for both individual and social orderings, and correspondingly, for both individual and collective rationality. But at least it opens the possibility of there being a decision procedure for arriving at social orderings. And at least the issues, insights, and methods applicable to the individual realm need no longer seem so distinct, much less necessarily irrelevant, to those of the social, or collective realm.

[48] Not everyone has felt this way. Some people have tried to show that purported instances of intransitive preferences are more often apparent than real—hence the holders of such preferences needn't be regarded as inconsistent and irrational after all! Others have suggested that given the actual conditions of choice under which most operate, it might be useful, and therefore rational, to adopt "simplification procedures . . . which approximate one's 'true preference' very well," and hence which *usually* serve one in good stead, but occasionally lead to intransitivities. (The quotation is from Amos Tversky's classic article on this topic, "Intransitivity of Preferences," *Psychological Review* 76 [1969]:31–48.) My suggestion differs from these in supposing that one might rationally hold three genuinely intransitive preferences, and *not* simply as a result of relying on a simplified approximation method which is usually helpful, but steers us wrong in such cases.

In sum, in addition to the independent reasons for rejecting IA, there are certain theoretical advantages to rejecting the transitivity of preferability and/or the Independence of Irrelevant Alternatives Principle. But, of course, there are also enormous problems with such a move. As our discussion of the baseball analogy suggests, giving up IIAP raises substantial practical and theoretical difficulties concerning the alternatives which must be considered to determine how two situations compare all things considered. So, too, will giving up the transitivity of preferability, since presumably, knowing that A is better than B all things considered would no longer be sufficient to prefer A over B, or to remove B from further consideration, if there might be some third alternative which is both worse than B, yet better than A. Moreover, for most, there seems to be an extremely deep, almost conceptual, link between the notion of all things considered better than and *both* the notion of transitivity *and* the Independence of Irrelevant Alternatives Principle.

Conclusion

In Part I, I argued that given the assumptions underlying the Mere Addition Paradox, Parfit's conclusions do not follow. Instead, a more radical conclusion seems to follow: "all things considered better than" is not a transitive relation. In Part II, I showed that one could avoid both the Mere Addition Paradox and the intransitivity of preferability by accepting an Intrinsic Aspect View, according to which how good or bad a situation is depends solely on its internal features. In Part III, I contended that despite IA's advantages, it is extremely difficult to reject the view that at least *some* factors are essentially comparative, and suggested a fundamental incompatibility between that view, the view that "all things considered better than" *must* be transitive, and the view that how two alternatives compare all things considered depends solely on how *they* compare in terms of the relevant and significant factors for making *that* comparison, and not on how one or both compare to some other, independent, alternative(s).

Not everyone will be troubled by my results. Those caring only about total utility, or perfectionism, might simply deny the moral relevance of essentially comparative factors. However narrow and implausible, such positions could at least avoid the deep practical and theoretical problems associated with rejecting either the transitivity of preferability or the Independence of Irrelevant Alternatives Principle.

Others might accept the view that certain notions are essentially comparative, and not be bothered by its implications. In fact, some Aristotelians and Kantians might relish insuperable difficulties with ranking outcomes. Already convinced that the question "what ought to be the case?" receives too much attention, they might welcome its relegation to the scrap heap of the unanswerable and irrelevant, enabling the "genuinely" important questions—that is, "how ought one *to be*?" or "what ought one *to do*?"—to receive more (their rightful?) consideration in the domain of practical reasoning.[49]

Unfortunately, I do not see how to plausibly reject the view that certain morally relevant factors are essentially comparative. Nor am I able to happily embrace the deep difficulties involved in rejecting either the view that preferability is transitive, or the view expressed by the Independence of Irrelevant Alternatives Principle.[50] To the contrary, as important, and perhaps even primary, as the Aristotelian and Kantian concerns are, it seems there are countless moral and practical situations where one either needs or wants to determine which of several outcomes would be best "all things considered." In sum, I am in the disturbing position of failing to see how to reconcile three deeply held views, but am loath to give any of them up. And I believe that, on reflection, many others will also find themselves in this position.

If, in the end, at least one of the three views must be given up, which one(s) should go? I do not know. Much more work needs to be done to answer that question. However, I am confident that, for many, to give up *any* of the three views would require a major shift in their practical and moral reasoning. In fact, I suspect it would fundamentally alter their very conception of practical and moral reasoning—perhaps of rationality itself.[51]

[49] This position should not be taken lightly. Though it is not a position I readily endorse, perhaps the arguments of this article are best interpreted as a frontal assault on the intelligibility of consequentialist reasoning about morality and rationality. Such reasoning may need to be severely limited, if not jettisoned altogether.

[50] There is a sense in which rejecting the transitivity of preferability is not even an *option* for those who think transitivity is part of the *meaning* of "all things considered better than." However, such people may be persuaded that the notion of "all things considered better than" is incoherent—a no less disturbing result.

[51] Some people may interpret the results of this article as (further) reason to be leery

of the subject matter of ethics. But one must be careful here. Intransitivities can occur outside the domain of moral reasoning. One case, which many find counterintuitive, is that of the three, perfectly fair, "intransitive dice." Suppose Die A's six faces are 6, 6, 6, 2, 2, 2; Die B's are 5, 5, 4, 3, 3, 2; and Die C's are 6, 4, 3, 3, 3, 3. It is easily shown that rolling two die at a time, on any given roll, the probability of A beating B will be 6/5, the probability of B beating C will be 14/13, and the probability of C beating A will be 6/5. Hence, in the long run, one would do better to bet on A against B, B against C, and C against A. (I am grateful to Gerald Massey for bringing this example to my attention.)