Rethinking the Person-Affecting Principle

Abstract:

In Rethinking the Good, Larry Temkin argues for a principle that he calls the Narrow Person-Affecting View. In its simplest formulation, this principle states that a first outcome can be better than a second outcome only if there is someone who fares better in the first outcome than in the second. Temkin argues that this kind of principle gives us reason to reject the Transitivity Thesis, according to which, if A is better than B, and B is better than C, then A must be better than C. In this paper, I argue that the various formulations Temkin has offered of the Narrow Person-Affecting View all face serious problems. I then propose an alternative view that captures the spirit of Temkin's formulations while avoiding their difficulties. I conclude by arguing that, even if we accept such a person-affecting view, we needn't reject the Transitivity Thesis.

Keywords: Person-Affecting View, Transitivity, Non-Identity Problem, Mere Addition Paradox, Larry Temkin, Derek Parfit

Larry Temkin's magnum opus, *Rethinking the Good*, is a grand work, both in its philosophical range and in its physical size. It is also weighty work, both in its philosophical significance and in its physical mass. And it is a deep work, both in its philosophical profundity and in its physical thickness. Indeed, it is unquestionably one of the grandest, most significant, and most profound works on value and practical reasoning ever written. For all these reasons, I will be able, in this contribution, only to scratch its surface. I will be focusing my attention on one particular claim that figures in the book, namely the Narrow Person-Affecting View. I believe, however, that this claim merits close attention, not only because it has considerable intuitive appeal and plays an implicit role in much of our thinking (as Temkin shows), but also because it figures in many of Temkin's most powerful arguments for intransitivity.

I will proceed as follows. In order to clarify the motivations for the version of the principle that Temkin proposes in *Rethinking the Good*, I will begin, in the first two sections, by considering an earlier, simpler version of the principle, which he presented in "Intransitivity and the Mere Addition Paradox" (Temkin, 1987). And I will discuss the two main problems facing these principles, namely the Non-Identity Problem and what I will call the Problem of

Symmetric Intransitivity. Then, in the third section, I will argue that, while the new formulation that Temkin presents in *Rethinking the Good* can solve the problems facing his original formulation, it gives rise to new problems that are no less severe. In the fourth section, I will propose an alternative principle which, I argue, does justice to person-affecting considerations while avoiding the various problems facing Temkin's formulations. Then, in the final section, I will consider the implications of person-affecting principles for intransitivity. I will argue that, while Temkin's original formulation of the person-affecting principle had the radical implication that the *better-than* relation is intransitive, more plausible formulations have much more moderate implications. In particular, I will argue that the kind of intransitivity implied by more plausible versions of the principle is best understood in deontic terms, or as concerning what we *ought to do*, rather than in axiological terms, as concerning the *values* of outcomes.

1. The Person-Affecting Principle and the Non-Identity Problem

Sometimes a distinction is drawn between *wide* person-affecting principles, which compare outcomes on the basis of how *people in general* fare, and *narrow* person-affecting principles, which compare outcomes on the basis of how *particular people* fare. In this paper, I will be focusing on the second kind of principle, and I will be using the phrase "person-affecting" to mean *narrow person-affecting*, or concerned with how particular people fare.

An early instance of a principle of this kind is the one Temkin simply calls "the Person-Affecting Principle" (PAP) in his groundbreaking paper "Intransitivity and the Mere Addition Paradox." He states this principle as follows (pp. 166-67):

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

As Temkin is well aware, this principle faces a serious problem, as it often seems to get the wrong results when comparing outcomes whose populations don't overlap. Following Parfit, we

may call this the *Non-Identity Problem*.[1] To illustrate this problem, let's consider the following pair of alternatives, which I borrow from Norcross 1999 (p. 774):

OK: This outcome contains one billion people, each of whom has an okay life, at welfare level 1.

GREAT: This outcome contains two billion people who are distinct from those who exist in OK (that is, there is no overlap among the people who are present in the two outcomes). And each of these two billion people has a great life at welfare level 10.

Since there is no one who exists in both outcomes, it seems to follow that there is no one who fares worse in OK than in GREAT. And hence it seems to follow that OK isn't worse than GREAT for anyone. And if this is right, then PAP will entail that OK isn't worse all things considered. And this result is highly implausible.

Note that we can create a similar problem involving lives that are not worth living. Compare the following two alternatives:

POOR: This outcome contains one billion people, each of whom has a poor life, at welfare level -1.

HORRENDOUS: This outcome contains two billion people who are distinct from those who exist in POOR. And each of these two billion people has a horrendous life at welfare level -10.

Once again, since there is no one who exists in both outcomes, it seems to follow that POOR isn't better for anyone. And so PAP seems to have the implausible implication that POOR isn't better than HORRENDOUS all thing considered.

The derivations of these implausible results from PAP turns on the following assumption:

Neutrality of Nonexistence: A first outcome can only be better for a given person than a second outcome if this person exists in both outcomes and fares better in the first than in the second.

Hence, the defender of PAP might attempt to avoid the implausible implications by rejecting the Neutrality of Nonexistence. For she might claim that an outcome in which someone does not exist is worse for someone than an outcome in which this person has a life that is worth living. Hence, she could maintain that OK is worse than GREAT, since it is worse for all those people who exist in GREAT but not in OK. Similarly, she might claim that an outcome in which

[1] See chapter 16 of Parfit 1984, esp. pp. 122-124.

someone does not exist is *better* for someone than an outcome in which this person has a life that is *worth ending*. Hence, she could maintain that POOR is better than HORRENDOUS, since it is better for all those people who exist in HORRENDOUS but not in POOR.

But Temkin rejects this kind of response.[2] And, within the context of his philosophical project, it's very important that he do so. For, according to Temkin, what makes PAP so significant, philosophically, is that it is *essentially comparative*. That is, it implies that, in order to determine which of two outcomes is better, we can't simply evaluate the two outcomes independently, and then compare the scores that each of these outcomes receives. Rather, our evaluation of each outcome must be guided by the particular alternative to which it is being compared. It is because PAP is essentially comparative, according to Temkin, that it supports intransitivity. But suppose that, in order to avoid the counterexamples just considered, the defender of PAP were to reject the Neutrality of Nonexistence, and maintain instead that non-existence is worse for someone than having a life that is worth living, and better for someone than having a life that is worth ending. On these assumptions, PAP is no longer essentially comparative. Indeed, on these assumptions, PAP is entailed by total utilitarianism, which is clearly an Internal Aspect view rather than an Essentially Comparative view, and which entails that the better-than relation is fully transitive.

Fortunately, Temkin does not take the route of rejecting the Neutrality of Non-Existence. Instead, he proposes an alternative response to the Non-Identity Problem. He writes:

> Derek Parfit has presented an ingenious argument, the *Non-Identity Problem*, which challenges PAP. . . [However,] 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. (Temkin 1987, p. 167.)

[2] See, e.g., Temkin 1987, footnote 30, pp. 166-167.

Here Temkin's suggestion seems to be that, except in unusual cases where there is no overlap between the people who exist in the two outcomes being compared, PAP is highly plausible. However, PAP has implausible implications even in cases where there is such overlap. This can be seen by considering the following pair of alternatives (adapted from Norcross 1999, p. 775):

> OK + Fred: Just like OK, except an additional person, Fred, is present at welfare level 5.
>
> GREAT + Fred: Just like GREAT, except that Fred is present at welfare level 5.

In this case, PAP implies that neither outcome is better than the alternative, since the only person who is present in both outcomes, namely Fred, fares equally well in both. But, clearly, GREAT + Fred is better than OK + Fred. And since this case involves an overlap between the populations of the two alternatives, it shows that PAP has unacceptable implications even when restricted to pairs of alternatives involving such overlap. Temkin responds to this argument in *Rethinking the Good*, where he writes:

> Arguably … the upshot of Norcross's argument is that the Narrow Person-Affecting View is problematic for cases where only one person, or small group, is the focus of the Narrow Person-Affecting View's assessment, in contexts where many other people are also involved (2012, p. 547).

Thus, Temkin's new suggestion seems to be that we need to further limit the scope of PAP: not only must we exclude cases where there is *no* overlap between the people who exist in the two outcomes being compared, but we must also exclude cases where there is only a *very small* overlap, relative to the total populations involved. But, in addition to seeming rather *ad hoc*, this further restriction doesn't really solve the problem. Recall that the problem with the unrestricted version of PAP is that it implies that the GREAT is no better than OK. And the problem with Temkin's first response (the response that involves conceding that GREAT is better than OK, but maintaining that PAP still applies in cases involving overlapping populations) is that it implies that the evaluative difference between GREAT and OK is negated by adding Fred to both outcomes. A very similar problem arises for Temkin's new response (the response that involves conceding that GREAT + Fred is better than OK + Fred, but maintaining that PAP still applies in cases involving *large* overlapping populations). For, while this second response doesn't imply that the evaluative difference between GREAT and OK is negated by adding *small* identical groups to both outcomes, it does imply that this difference is negated by adding *large* identical groups to both outcomes. And this implication isn't much of an improvement.

Moreover, the resulting view still has the unwanted implication that adding *small*, identical groups of people to each of two outcomes can negate enormous evaluative differences between them. To see why this is so, let *n* be the minimum number of people such that the PAP applies to a comparison between OK + *n* people and GREAT + *n* people. Let MINIMAL be a group of people that contains exactly *n* people. And let SUBMININMAL be a group of people that's just like MINIMAL, except that it has a few less people. On the view under consideration, while GREAT + SUBMINIMAL is better than OK + SUBMINIMAL (since, *ex hypothesi*, PAP doesn't apply to this comparison), GREAT + MINIMAL is not better than OK + MINIMAL (since, *ex hypothesi*, PAP *does* apply to this comparison, and the overlapping people fare equally well in the two alternatives). But the only difference between the second pair of outcomes and the first is that, in each of the second pair of outcomes, there are a few extra people. And so the view under consideration implies that adding small identical groups of people to each of two outcomes can abolish the vast evaluative difference between them.

Forturately, I believe there is a much better response to the Non-Identity Problem available to Temkin. Temkin mentions this response in the following characteristically generous footnote (note that the principle of "differential utility" referred to in this passage is a precisification of the Person-Affecting Principle, as it focuses exclusively on the welfare of those who exist in both outcomes being compared):

> In keeping with the spirit of pluralism prevalent throughout this book, [name omitted for blind review] has suggested (in correspondence) that we might want to always attach some weight to both *total utility*, which is best captured by an Internal Aspect View, and what we might call *differential utility*, which is best captured by an Essentially Comparative View. Differential utility reflects how any outcomes being compared differentially affect the welfare levels of those who exist in each outcome. On this view, differential utility matters even in cases where there is only a small overlapping group, such as Norcross's case where there is only a single overlapping person, Fred. But how much it matters [in comparison to total utility] is proportional to the size of the overlapping group, so that where the group is very small relative to the total size of the groups in the different outcomes, as in Norcross's case, the significance of differential utility will be swamped by the significance of total utility. [Reference omitted for blind review.]

I believe that we must adopt the view suggested above, or some other such hybrid view, if we are to endorse a person-affecting principle while avoiding highly counterintuitive implications. For,

in order to avoid the absurd implication that GREAT is no better than OK, we must maintain that non-person-affecting considerations (that is, considerations that do not supervene on the welfare of those who exist in both outcomes being compared) are relevant in comparing OK and GREAT. And, in order to avoid the implausible implication that adding identical groups of people to both outcomes abolishes this evaluative difference, we must maintain that these non-person-affecting considerations are likewise relevant in comparing outcomes whose populations overlap. Hence, if we also want to maintain that, in comparing outcomes with overlapping populations, *person-affecting* considerations are relevant, then we must adopt a hybrid view according to which both person-affecting and non-person-affecting considerations are relevant to such comparisons. The simplest way to revise PAP, so as to make it consistent with such a hybrid view, is as follows:

> **Weak Person-Affecting Principle (WPAP)**: In evaluating outcomes, there *is one important evaluative dimension*, which we may call *the person-affecting dimension*, such that
>
> (i) *On this dimension*, a first outcome is always better than a second if everyone who exists in both outcomes fares better in the first outcome.
>
> (ii) *On this dimension*, a first outcome is never better than a second if there is no one who is present in both outcomes who fares better in the first outcome.

Adopting such a view will enable us to avoid the Non-Identity Problem. It will not, however, allow us to avoid another serious problem, which I discuss in the next section.

2. The Problem of Symmetric Intransitivity

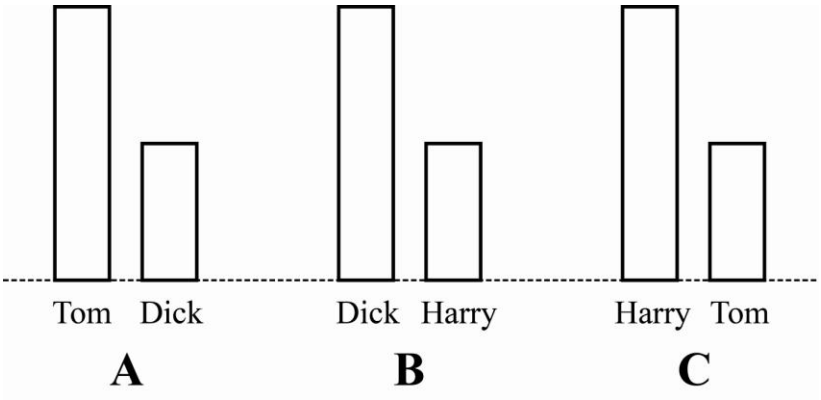
The next problem we must consider is a problem for both PAP and WPAP. Like the Non-Identity Problem, this problem was first raised by Derek Parfit,[3] and can be illustrated by the following trio of outcomes:

A: Only Tom and Dick exist. Tom has welfare level 10 and Dick has welfare level 5.

B: Only Dick and Harry. Dick has welfare level 10 and Harry has welfare level 5.

C: Only Harry and Tom exist. Harry has welfare level 10 and Tom has welfare level 5.

[3]Temkin attributes this problem to Derek Parfit in Temkin 2012, pp. 428-429.



It seems clear that these three outcomes are equally good. But PAP and WPAP each imply otherwise. For, since the only person present in both A and B (namely Dick) fares better in B, PAP implies that B is better than A all things considered, and WPAP implies that B is better than A with respect to the person-affecting dimension of evaluation. And since, setting aside person-affecting considerations, A and B are clearly equally good, it follows from WPAP that B is likewise better than A all things considered. And, by similar reasoning, PAP and WPAP also imply that, all things considered, C is better than B (since Harry fares better in C) and that A is better than C (since Tom fares better in A). Thus, rather than implying that all three outcomes are equally good, as our intuitions indicate, these principles imply that each of the three outcomes is better than one of the alternatives and worse than the other. (Let's call this the *Problem of Symmetric Intransitivity*, since it involves a perfectly symmetrical case where the principles under consideration implausibly imply that there is intransitivity.)

In *Rethinking the Good,* Temkin responds to this objection by offering a revised version of the person-affecting principle, which he introduces in the following passage:

> In any choice situation between possible outcomes, let us call those people who do exist, or have existed, or will exist in each of the outcomes independently of one's choices *independently existing people*. By contrast, let us call those people whose existence in one or more possible outcomes depends on the choices one makes in bringing about an outcome *dependently existing people*. We can now state:
>
> ***Narrow Person-Affecting View***: In assessing possible outcomes, one should (1) focus on the status of independently existing people, with the aim of wanting them to be as well off as possible, and (2) ignore the status of dependently existing people, except that one wants to avoid *harming* them as much as possible. Regarding the second clause, a dependently existing person is harmed only if there is at least one available alternative outcome in which

> that very same person exists and is better off, and the size of the harm will be a function of the extent to which that person would have been better off in the available alternative outcome in which he exists and is best off. (Temkin 2012, p. 417).

Since this is a revision of Temkin's original Person Affecting Principle, I will refer to it as the *Revised Principle*. I note, in passing, that Temkin's formulation of this principle is more complicated than it needs to be. For, with respect to independently existing people, the aim of wanting them to be as well off as possible is equivalent to the aim of wanting them to avoid harm as much as possible (as Temkin defines harm). And so, without affecting the content of the principle, we could eliminate clause (1), and simply state that, in assessing the outcomes in a given set of available alternatives, one should prefer the outcomes in which there is the least amount of harm (where the degree of harm to a given individual in a given outcome is a function of the difference between her level of welfare in that outcome and her level of welfare in the available outcome in which she fares best).

As Temkin shows, his revised formulation avoids the implausible implication that, in a choice among A, B, and C, each of these outcomes would be worse than one of the alternatives. For in such a choice situation, Temkin explains,

> the Narrow Person-Affecting View… generates the judgment… that all three alternatives are equally good … [For in such a situation] there are no independently existing people, so comparisons of the different alternatives will turn on the extent to which, if any, the dependently existing people are harmed by the different alternatives. But, as should be plain, from the standpoint of the Narrow Person-Affecting View there will be a single dependently existing person in each outcome who will be harmed, and to the very same extent, no matter *which* outcome is brought about. (*Ibid*., p. 431.)

Temkin solves the Problem of Symmetric Intransitivity by *relativizing* the better-than relation to choice situations, or to sets of available alternatives. Thus, the Revised Principle implies that B is better than A relative to some choice situations (such as a situation in which A and B are the only two available alternatives), and yet B is not better than A relative to other choice situations (such as a situation in which A, B, and C are the available alternatives).

Note, however, that, while the Revised Principle solves the Problem of Symmetric Intransitivity, it seems no better than his original formulation with respect to the Non-Identity Problem. For, like PAP, his revised formulation appears to imply that person-affecting considerations are the only considerations that are relevant in comparing outcomes. And so it

appears to have the unacceptable implication that, in a choice between OK and GREAT, neither outcome is better than the other, since neither outcome is favored by person-affecting considerations. But this problem can be easily solved by restating the principle so that it concerns only one dimension of evaluation, as follows:

> **The Weakened Revised Principle**: There *is one important evaluative dimension*, which we may call *the person-affecting dimension*, such that, with respect to this evaluative dimension, a first outcome is better than a second outcome, *relative to a given choice situation*, just in case the first outcome involves less harm than the second outcome (where the degree of harm to a given individual in a given outcome is proportional to the difference between her level of welfare in that outcome and her level of welfare in the available outcome in which she fares best).

This formulation solves both the Problem of Symmetric Intransitivity and the Non-Identity Problem. Unfortunately, however, this principle (as well as Temkin's Revised Principle) gives rise to new problems, which we will explore in the next two sections.

3. The Problem of Mere Addition

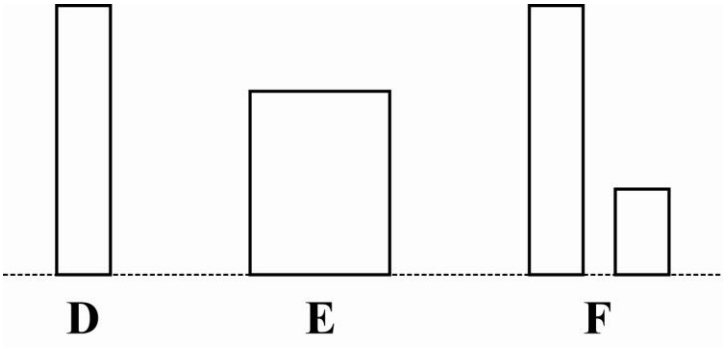
In this section, I will argue that the Revised Principle, and the Weakened Revised Principle, face a particularly acute version of the problem of Mere Addition, which, ironically, is precisely the problem that person-affecting principles were originally introduced to solve.

In order to introduce this problem, let us recall Parfit's Mere Addition Paradox. In its simplest form, it involves the following three alternatives:

D: There are exactly 10 billion people, each of whom has a wonderful life with a welfare level of 10.

E: This outcome contains the 10 billion people outcome D, plus an additional 10 billion people. And everyone is at welfare level 7.

F: This outcome contains the 10 billion people from outcome D, who are at welfare level 10. plus an additional 10 billion people, who are at welfare level 3.



As Temkin indicates in the preface to his book, it was thinking about this very case that led him to question whether the better-than and worse-than relations are transitive. Parfit argues that, in this case, transitivity appears to be violated. His reasoning proceeds thus:[4]

> E is better than D: for doubling a very large population, while significantly decreasing everyone's welfare, is a change for the worse. Furthermore, F is worse than E. For F has the very same people as E, and F is worse than E with respect to total welfare, equality, and maximin. But F *cannot* be worse than D. For F can be derived from D by the mere addition of 10 billion people whose lives are well worth living. And surely the existence of such additional people can't make an outcome worse. Hence, D is better than E, which is better than F, and yet D is not better than E, violating the transitivity of the better-than relation.

Crucial to this paradox is the claim that F is no worse than D. And Temkin argues that what underlies the plausibility of this claim is a person-affecting principle, according to which one outcome can be worse than another only if it is worse for someone. It follows from such a principle that F can't be worse than D, since there is no one who fares worse in F than in D. Moreover, Temkin argues, if we adopt a person-affecting principle of this kind, we can not only explain why F is no worse than D, but we can also explain why we shouldn't expect transitivity to apply in this case. And so the Mere Addition Paradox ceases to appear so paradoxical.

Temkin's revision of the Person-Affecting View is thus a refinement of a principle that was originally introduced in order to explain our intuitive judgment that F is no worse than D. However, his new principle no longer has this implication, at least in the case where all three outcomes are available. For, according to this principle, in evaluating a set of alternatives, our aim should be to minimize comparative harm. And while there is no comparative harm in D (since no one fares better in any other alternative than in D) there is considerable comparative harm in F (since there are 10 billion people who fare much better in E than in F). Hence, Temkin's Revised Principle implies that, when outcomes D, E, and F are available, F is worse than D, in spite of the fact that F can be derived from D by the mere addition of 10 billion people whose lives are worth living.

While there is a certain irony to this result, it certainly doesn't constitute a devastating problem for the Revised Principle. For it's far from obvious that F *isn't* worse than D. After all, F is much less equal than D, and the minimum and average levels of welfare are both much

[4] See the final chapter of Parfit 1984.

lower in E. And it's not so implausible to suppose that these considerations might count against F.

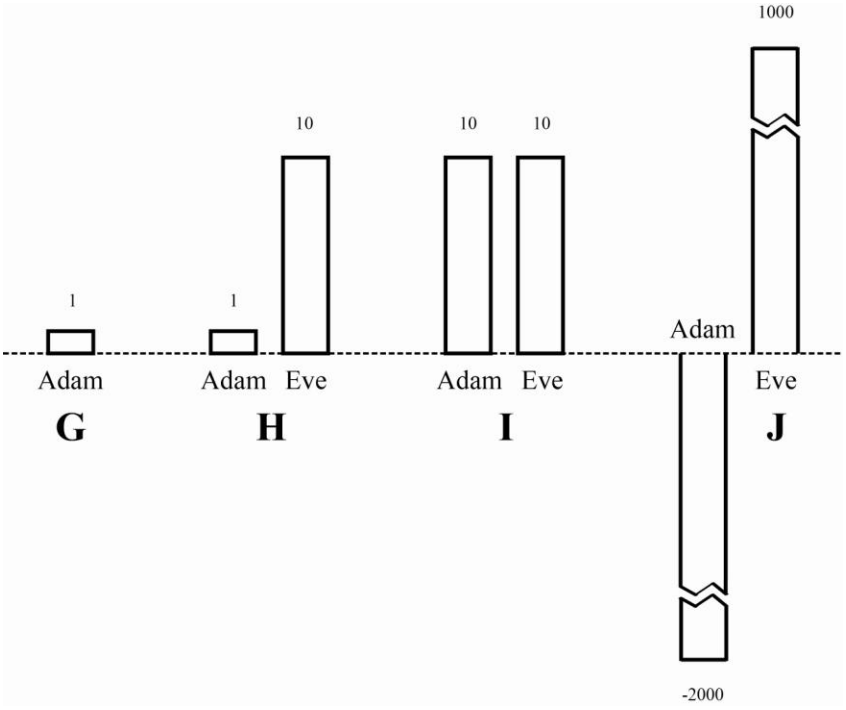
However, the Revised Principle, and the Weakened Revised Principle, face a much worse problem. These principles don't just imply that the mere addition makes the outcome worse in the Mere Addition Paradox (where this implication is controversial); these principles also imply that mere addition can make an outcome worse in a cases where this implication is uncontroversially false. As an illustration, let us suppose that, initially, only Adam exists. And suppose that God is choosing among the following four outcomes:

G: No one else besides Adam is created. He is confined to the wilderness outside of Eden, and has a welfare level of 1.

H: In addition to Adam, Eve is created. While Adam is confined to the wilderness, and has a welfare level of 1, Eve is allowed into Eden, where she has a welfare level of 10.

I: In addition to Adam, Eve is created. Both live in Eden with welfare levels of 10.

J: In addition to Adam, Eve is created. Eve has a blissful life in Heaven for many millennia, with a welfare level of 1000. And Adam is sent to Hell for twice the length of time that Eve spends in Heaven, for a welfare level of -2000.



In this case, H can be derived from G by mere addition. And in H, unlike in F, it should be fairly uncontroversial that this addition is not a change for the worse. For while F was derived from D

by adding a group of people who are much worse off than those in the original group, H is derived from G by adding a group of people who are much better off. And while the first change results in a significant *lowering* in average and minimum welfare, the second change results in a significant *increase* in average and minimum welfare.

There is, however, one respect in which H appears to be worse than G, namely equality. Hence, someone who places a great deal of weight on equality in evaluating outcomes might maintain that H is worse than G all things considered, in spite of being better in several other respects. This inequality, however, is absent in outcome I. For in I Adam's welfare level is raised to Eve's higher level. Thus, I seems to be significantly better than G in several important respects (such as total, average and minimum welfare) and worse in no respects. And so it should be absolutely uncontroversial that I is no worse than G.

Unfortunately, Temkin's Revised Principle is not consistent with these intuitive judgments. For this principle implies that, in comparing the four available alternatives, we should aim to minimize comparative harm. And in outcome G, there are only 9 units of comparative Harm (since that's the margin by which Adam fares better in the best available outcome than in I). By contrast, each of the remaining outcomes contain vastly more comparative harm: 999 in H, 990 in I, and 2010 in J. For Eve fares vastly worse in H and in I than in J, and Adam fares vastly worse in J than in I. And so Temkin's Revised Principle implies that G is the best outcome.

What about the Weakened Revised Principle? While this principle doesn't strictly entail that G is the best outcome, it will have this implication so long as a reasonable amount of weight is given to the person-affecting dimension of evaluation. For this principle implies that, since G contains vastly less comparative harm than any of the alternatives, it is vastly better than these alternatives with respect to the person-affecting dimension of evaluation. By contrast, G appears to be worse than H and I with respect to some important factors (such as total utility), the degree to which G falls short of H and I with respect to these other factors is dwarfed by the degree to which it better than them with respect to comparative harm. Hence, so long as the person-affecting dimension is given significant weight, the Weakened Revised Principle will imply that G is better than H or I.

Nor can the defender of the Weakened Revised Principle solve the problem by simply giving the person-affecting dimension of evaluation a very small weight in determining the overall values of outcomes. For, so long as this factor weighs against the other factors *to any*

degree whatsoever in determining the overall value of outcomes, we can generate a version of the same problem. This can be seen by considering the following variant of the case just considered:

G*: No one else besides Adam is created, and his welfare level is *x*.

H*: In addition to Adam, Eve is created. Adam's welfare level is x, and Eve's is x(1+x).

I*: In addition to Adam, Eve is created. Both have welfare level x(1+x).

J*: In addition to Adam, Eve. Eve has a blissful life in Heaven for many millennia, with a welfare level of 1000. And Adam is sent to Hell for twice the length of time that Eve spends in Heaven, for a welfare level of -2000.

For any positive value of x, it should be fairly uncontroversial that H* is not worse than G* (since H* can be derived from G* by adding someone who is better off than the person in H*), and it should be completely uncontroversial that I* is not worse than G* (since I* can be derived from G* by adding a better off person and raising the original person to this higher level). However, as x approaches zero, the degree to which G* is better than H* and I* with respect to comparative harm remains vast, and yet the degree to which G* falls short of H* and I* with respect to the other seemingly important factors (such as total, minimum, and average utility) approaches zero. Hence, so long as the defender of the Weakened Principle maintains that the person-affecting dimension of evaluation weighs against the other relevant dimensions to any degree whatsoever, she will be committed to the conclusion that, for sufficiently small values of x, H* and I* are both worse than G* all things considered.

We could solve this problem by moving to the following view:

Lexical Revised Principle: The person-affecting dimension of evaluation is relevant only in breaking ties between outcomes that are equally good, overall, in other respects. With respect to the person-affecting dimension of evaluation, a first outcome is better than a second outcome, *relative to a given choice situation*, just in case the first outcome involves less harm than the second outcome.

Such a view may allow us to solve the Mere Addition Problem. In particular, it will allow us to maintain that, regardless of the value of x, H and I are both better than G, since they are better, overall, with respect to non-person-affecting factors. And, given the right kinds of view about what other factors are relevant to evaluating outcomes, the Lexical Revised Principle may allow one to vindicate Parfit's original contention that mere addition can never make an outcome

worse. Suppose, for example, that the defender of the Lexical Revised Principle maintains that the only relevant factor, apart from the person-affecting dimension, is total utility. Since the mere addition of people whose lives are worth living always increases total utility, this view will imply that such mere addition never makes an outcome worse.

However, while the Lexical Revised Principle may solve the Mere Addition Problem, it fails to solve a more fundamental problem of which the Mere Addition Problem is just one manifestation. I will discuss this deeper problem in the next section.

4. The Problem of Improvable Life Avoidance

In any choice situation involving a set S of available alternatives, let us say that a given person has an *improvable life* in one of these alternatives if there is some other available alternative in which she fares better. The fundamental problem with the Revised Principle, as well as with the two weakened version of this principle that we have considered, is that they all imply that we have person-affecting reason to prefer outcomes in which a given person does not exist to outcomes in which this person exists and has an improvable life. For, in any outcome in which a given person does not exist, she experiences no comparative harm, by in any outcome in which she exists and has an improvable life, she does experience comparative harm, since she fares worse than she does in some available alternative. Hence, since all these principles understand the person-affecting dimension of evaluation in terms of the minimization of comparative harm, they all imply that we have reason to prefer for someone not to exist than to exist with an improvable life.

As a result, in most situations in which we are able to bring about a group of people (e.g., when we are planning a future family, or choosing public policies that will affect reproduction), these principles will imply that, at least with respect to the person-affecting dimension of evaluation, we would do best to prevent the group from existing rather than bringing them about. For, setting aside the effects that the new people may have on the welfare levels of independently existing people, failing to bring about a group of people will never increase comparative harm. By contrast, courses of action that involve bringing about a group of people will generally cause comparative harm to at least some of the people created. For, typically, when we are in a position to bring about a group of people, the set of options that are available to us will include options that differentially affect the various people we might create. And we shouldn't expect

any of these options to be best for *everyone*. Rather, in the ordinary case, we should expect that trade-offs will be required among the interests of those that we might create. And so we should expect that, in each of the available options that involves bringing people into existence, there will be someone whom we could have made better off, and hence someone who experiences comparative harm.

Thus, the Revised Principle, as well as the weakened versions thereof, will imply that, setting aside the affects of new people on independently existing people, we normally have most person-affecting reason to avoid bringing new people into existence. Of course, the defender of these principles needn't maintain that, *all things considered*, we should normally avoid bringing others into existence. For she might maintain that procreation is often justified by its benefits for independently existing people, such as the joys it brings to the parents. Or she might maintain that procreation is justified by non-person-affecting considerations, such as an increase in total utility. Still, she must maintain that, *with respect to our concern for those we might create*, we will generally have reason to avoid bringing them into existence, even if, were we to bring them into existence, they would, without exception, have wonderful lives. And this implication is hardly plausible.

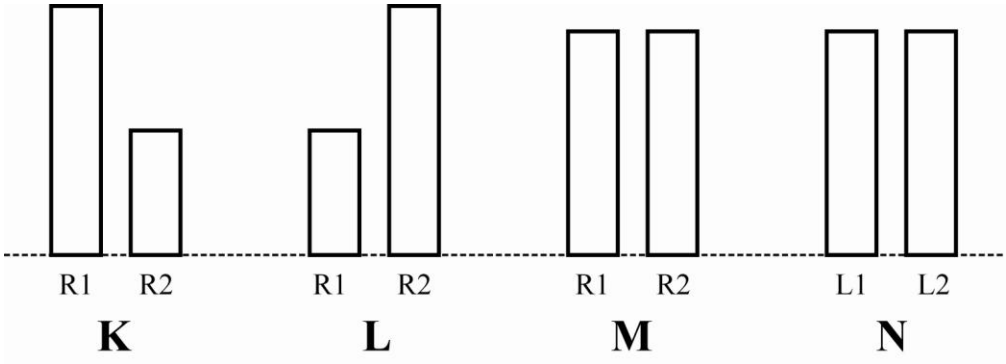
And this implication gives rise to implausible implications even concerning all-things-considered judgments. To see why this is so, consider a case in which Rachel and her Sister Leah are both at a fertility clinic that specializes in producing twins. Not wanting to disappoint their parents, who badly want to become grandparents, they have decided that one or the other of them will have twins. From a prudential point of view, they are each indifferent as to whether to have twins, since they both recognize that the benefits to themselves will counterbalance the harms. Since Leah has no preference one way or the other, she leaves the choice of who will have twins up to Rachel. Rachel is therefore choosing among the following four alternatives:

K: Rachel has two twins, R1 and R2. She favors R1, with the result that R1 has a welfare level of 10 and R2 has a welfare level of 5.

L: Rachel has two twins, R1 and R2. She favors R2, with the result that R2 has a welfare level of 10 and R1 has a welfare level of 5.

M: Rachel has two twins, R1 and R2. She treats them both equally, with the result that they each have a welfare level of 9.

N: Rachel lets Leah do the work. This would result in Leah's having two twins, L1 and L2. And while Leah would have the option of favoring either of her children, she would never do so. Instead, she treats them equally, with the result that they each have a level of welfare of 9.



It seems clear that, in this case, M and N are both better than K and L (since M and N are better with respect equality as well as total, average, and minimum welfare). And it also seems clear that M and N would be equally good. Hence, it seems clear that Rachel could reasonably choose M, and hence choose to have the twins herself. But the Revised Principle, even in its lexical formulation, implies otherwise. For M contains two people, R1 and R2, who each fare worse, by 1 welfare unit, in M than in some other alternative available to Rachel. By contrast, no one fares any worse in N than in any alternative available to Rachel. Consequently, with respect to this choice situation, M contains 2 units of comparative harm, whereas N contains none. Hence, the Revised Principle, as well as the weakened versions thereof, implies that N is better than M with respect to the person-affecting dimension of evaluation. Moreover, apart from person-affecting considerations, N is just as good as M, and N is better than either of the other alternatives. Hence, even the lexical version of the Revised Principle implies that N is the best option. And so it implies that Rachel should refrain from having twins.

Thus, the various versions of the Revised Principle imply that Rachel should leave the twin-rearing to her sister, so that the decision as to whether to favor either child will fall on her sister rather than her. For that's the only way Rachel can avoid doing comparative harm. And this implication is hardly plausible.

So far we have considered a number of versions of the person-affecting principle, as we have seen that they each face one or more problems. The following chart summarizes which problems are faced by which versions:

	Non-Identity	Symmetric Intransitivity	Mere Addition	Improvable Life Avoidance
PAP	X	X		
WPAP		X		
Revised Principle	X		X	X
Weakened Revised Principle			X	
Lexical Revised Principle				X

In the next section, my aim will be to arrive at a view that solves all these problems.

5. Toward a More Plausible View

I will begin by introducing some concepts. I will be assuming that, when certain outcomes are available to an agent, this agent has certain practical reasons in virtue of the availability of these outcomes. If, for example, someone can choose from a set of outcomes that includes [Everyone on Earth is tortured for 1000 years] and [Everyone on Earth is granted 1000 years of bliss], the fact that these are among the outcomes available to the agent will give her reason to avoid [Everyone on Earth is Tortured for 1000 years]. For any set of outcomes, S, let us say that the *S-given reasons* are the practical reasons that anyone would have for whom the outcomes in S are available, purely in virtue of the availability of these outcomes.

For any pair of outcomes, O1 and O2, let us say that O1 *defeats* O2 just in case the reasons given by the set {O1, O2} favor avoiding O2. Thus, [Everyone on Earth is granted 1000 years of bliss] defeats [Everyone on Earth is tortured for 1000 years], since everyone for whom both these outcomes are available would have reason to avoid the latter. Similarly, for any dimension of evaluation D, let us say that O1 defeats O2 *with respect to D* just in case the reasons along dimension D given by the set {O1, O2} favor avoiding O2. Thus, for example, - where A is an outcome with a small number of people who have excellent lives, and Z is an outcome where there is an astronomically vast number of people whose lives are barely worth living, then Z will

defeat A with respect to total utility (since anyone for whom both outcomes are available will have reasons of utility to avoid A) whereas A will defeat Z with respect to perfection (since anyone for whom both outcomes are available will have reasons of perfection to avoid Z).

Given some plausible background assumptions, WPAP and the Weakened Revised Principle each entail that the following is true about the defeat relation:

Person-Affecting Principle of Defeat (PAPD): There is one important evaluative dimension, which we may call the *person-affecting dimension*, such that

(i) With respect to this dimension, a first outcome *always* defeats a second outcome if *everyone* who exists in both outcomes fares better in the first outcome.

(ii) With respect to this dimension, a first outcome *never* defeats a second outcome if *no one* who exists in both outcomes fares better in the first outcome.

Like all the other person-affecting principles we have considered so far, this principle is very weak, whenever there is at least one person who fares better in each of two alternatives, it says nothing about how these alternatives compare. There are various ways in which this principle could be strengthened so as to give a more complete ranking. One simple and natural way to do so is as follows:

Differential Utility Principle of Defeat (DUPD): There is one important evaluative dimension such that, on this dimension, a first outcome defeats a second if and only if, among those who exist in both outcomes, the total utility is greater in the first outcome than in the second.

There are, of course, many alternatives ways in which we might strengthen PAPD, and I will remain neutral as to which of these, if any, we should adopt.

We must now move from the special problem of two-option choices to the more general problem of how we are to choose among arbitrarily many options. Suppose we are faced with a set S of available options, and we know the defeat relations that obtain between all the pairs of options in this set. Is there any way for us to figure out, on the basis of this information, which among the options in S it would be reasonable or permissible for us to choose? A natural suggestion is this:

Avoid Defeated Options (ADO): In choosing among any set S of available outcomes, the S-given reasons favor avoiding any outcome that is defeated by any of the alternatives.

But there is a problem with ADO, at least when it is conjoined with the Person-Affecting Principle of Defeat (PAPD). For PAPD implies that there can be choice situations in which there are no undefeated options. This is true, for example, in a choice among A, B, and C, where PAPD implies that C > B > C > A. And so it follows that there are choice situations in which it is impossible to follow ADO.

One response to this problem would be to maintain that, in choice situations in which there are no undefeated options, rational choice breaks down, and so we shouldn't expect to find principles that apply in such cases. But I believe this view is mistaken. To see why this is so, consider the choice among the following four alternatives (the number in parentheses after a given person's name indicates this person's welfare level in the outcome in question):

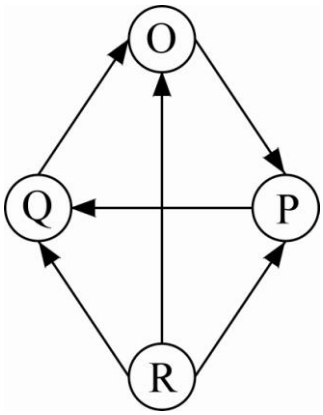
O: Tom (10), Dick (5), Paul (15)

P: Dick (10), Harry (5), Paul (15)

Q: Harry (10), Tom (5), Paul (15)

R: Peter (10), Paul (5), Mary (15)

Note that O, P, and Q are just like A, B, and C, except that they include an extra person, Paul, whose welfare level is 15. Thus, Q defeats P, which defeats O, which defeats Q. And just as it seems, for reasons of symmetry, that we should be indifferent among A, B, and C when those are the available options, it likewise seems, for reasons of symmetry, that we should be indifferent among O, P, and Q in the present choice situation. It seems, however, that if we take person-affecting considerations seriously, then we should choose *any* of these first three outcomes over R. For there is someone, namely Paul, who fares much worse in R than in any of the other alternatives. And there is no one who fares better in R. And so it follows from PAPD that each of the first three alternatives defeats R with respect to the person-affecting dimension of evaluation. And since, apart from this dimension, the four alternatives are equivalent, it follows that each of first three outcomes defeat R all things considered. We can represent the defeat relations among these outcomes by the following diagram, where an arrow proceeding from a circle representing a first outcome to a circle representing a second outcome indicates that the second outcome defeats the first:



Since, because of person-affecting considerations, R is defeated by all the other alternatives, it seems clear that anyone who takes person-affecting considerations seriously should avoid alternative R and choose one of the other alternatives instead. And so the mere fact that there are no undefeated options does not imply that anything goes, or that we should reject any attempt at rational guidance. What principle, then, should we appeal to in such contexts? One natural explanation as to why we should avoid R, in the choice situation just considered, is that it is defeated by all the available alternatives. This suggests the following principle:

> **Avoid Universally Defeated Options (AUDO)**: In choosing among any set S of available outcomes, the S-given reasons favor avoiding any outcome that is defeated by every one of the alternatives.

This principle appears to be correct, so far as it goes. But it also seems too week. To see why this is so, suppose we were to add the following alternative to S:

> **T**: Groucho (10), Chico (5), Harpo (15)

Note that, while T has exactly the same welfare distribution as each of the other four alternatives, it contains none of the same people. Hence, person-affecting considerations are irrelevant when comparing it to the other alternatives. It doesn't seem, therefore, that the availability of this fifth alternative could affect the person-affecting considerations bearing on the other alternatives. In particular, it doesn't seem that availability of this alternative could change the fact that person-affecting considerations strongly favor each of O, P, and Q over R, and hence that it would be a mistake to choose R when O, P and Q are available. Hence, we want a principle that implies that, even when T is added to the set of available options, it would still be a mistake to choose R. However, AUDO fails to imply this. For, when T is added to the set of available alternatives, it

ceases to be the case that R is defeated by each of the other alternatives. Hence, we need a stronger principle.

One possibility would be to supplement Avoid Universally Defeated Options with a restricted version of Avoid Defeated Options, namely the following:

> **Avoid Defeated Options Whenever Possible (ADOWP):** In choosing among any set S of available outcomes, if S contains at least one undefeated outcome, the S-given reasons favor avoiding any outcome that is defeated by any of the alternatives.

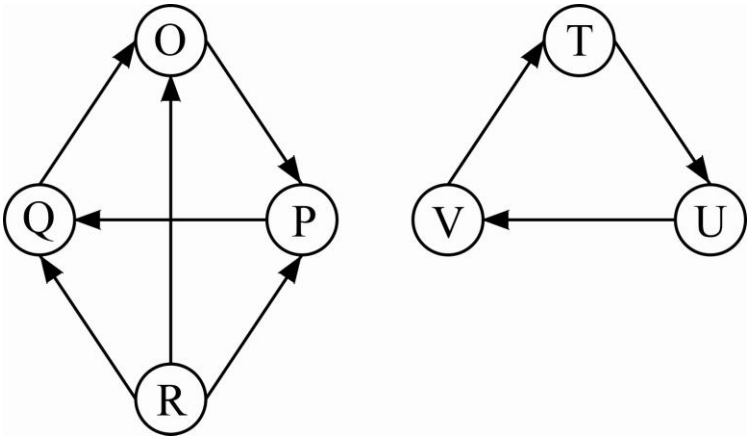
This principle, it seems, could explain why, when T is added to the set of available options, it remains the case that one should not choose R. For T is undefeated. And so ADOWP implies that, when T is added, we are required to avoid any defeated option, including R.

Unfortunately, this explanation is insufficiently general. To see why this is so, let us add two more outcomes to the set of available alternatives, namely the following:

U: Chico (10), Zeppo (5), Harpo (15)

V: Zeppo (10), Groucho (5), Harpo (15)

Note that T, U and V form a cycle of defeat that is isomorphic to the cycle formed by O, P, and Q. Thus, when these outcomes are added to the set of available alternatives, we can represent the defeat relations among these alternatives as follows:



Like T, U and V contain none of the same people as any of the first four other alternatives (O, P, Q and R). Hence, just as adding T alone shouldn't affect the way we rank these four alternatives, so adding T along with U and V should likewise not affect the way we rank these four alternatives. Hence, even when T, U and V are all present, it seems it would still be a mistake to

choose R. So we want a principle that still tells us to avoid R when all these alternatives are available. But the conjunction of AUDO and ADOWP fails to do this. AUDO fails to rule out R, since R is no longer universally defeated, and ADOWP fails to rule out R, because now there are no undefeated options, and so ADOWP does not apply.

What principle, then, will explain why we should avoid R, when outcomes O through V are all available? To answer this question, let's consider the diagram above. One can think of the arrows in this diagram as representing the possible trajectories one might follow in deliberating about the available alternatives, if one were always to compare alternatives two at a time, and if one were to move only from a defeated alternatives to the alternative by which it is defeated. Thought of in this way, we can see, from the diagram, that there is an important respect in which R differs from all the other alternatives. For there are paths of possible deliberation that lead *irreversibly away* from R. Thus, if one begins from R, and if, in one's deliberation, one moves from a given outcome only to another outcome that defeats it, then one can move from R to any of O, P, or Q, but one can never return. By contrast, in the case of all the other alternatives, while such deliberation can lead one away from them, continuing such deliberation further can always lead one back to one's original starting point.

In order to express this difference, it will be useful to introduce some terminology. For any two outcomes, O1 and O2, if it is possible, by way of the kind of deliberation just described, to move from O1 to O2, then let us say that O2 *directly or indirectly defeats* O1. That is, O2 directly or indirectly defeats O1 just in case there is a sequence of outcomes beginning with O1 and ending with O2 such that each outcome in this sequence is defeated by its successor. And let us say that an outcome O1 is *subordinate* to an outcome O2 just in case O2 directly or indirectly defeats O1, and not vice versa. Thus, O1 will be subordinate to O2 just in case, proceeding by the kind of deliberation described earlier, it is possible to move from O1 to O2, but impossible to move back again. Plausibly, it is the fact that R is subordinate to other available alternatives that explains why this option should be avoided. This suggests the following principle:

> **Avoid Subordinate Options (ASO)**: In choosing among any set S of available outcomes, the S-given reasons favor avoiding any outcome that is subordinate to any of the alternatives.

I will now argue that this principle, combined with the Person-Affecting Principle of Defeat (PAPD) avoids each of the four problems that we have considered for the alternative versions of the person-affecting principle.

First, ASO + PAPD avoids the Non-Identity Problem. It does so in a manner that is now familiar, namely, by treating person-affecting considerations as only one dimension of evaluation.

Second, ASO + PAPD avoids the problem of Symmetric Intransitivity. For this problem arises in cases like the choice among A, B, and C, where there is a cycle of defeat relations. And whenever the defeat relations among the available alternatives form a cycle, none of these alternatives will be subordinate to any other, in the sense defined above. And so ASO + PAPD will not imply that we should avoid any of these alternatives. Thus, ASO + PAPD makes sense of our intuitive judgment that, in such a choice situation, it would be permissible to choose any one of the available options.

Third, ASO + PAPD avoids the Mere Addition Problem. Recall that the Mere Addition Problem, which faces both Temkin's Revised Principle and the Weakened Revised Principle, is that both these principles imply that adding someone (or a group of people) who fare(s) *better* than any of the original people can make an outcome *worse*. But ASO + PAPD can never have this implication. To see why not, let ORIGINAL be any arbitrary outcome, and let EXTRA be an outcome that can be derived from ORIGINAL by adding one or more people, each of whom fares better than anyone in ORIGINAL. Since no one fares better in ORIGINAL than in EXTRA, it follows from PAPD that, with respect to the person-affecting dimension, ORIGINAL does not defeat EXTRA. And, with respect to non-person-affecting considerations taken together, EXTRA clearly defeats ORIGINAL, since EXTRA is superior with respect to total, maximum and average utility. Hence, all things considered, EXTRA will defeat ORIGINAL.

But if EXTRA defeats ORIGINAL, then ASO can never favor ORIGINAL over EXTRA. That is, no matter what set of alternatives may be available, ASO will never imply that we should avoid EXTRA, without also implying that we should avoid ORIGINAL. For ASO will imply that we should avoid EXTRA only if EXTRA is subordinate to some other option. But since EXTRA defeats ORIGINAL, it follows that if EXTRA is subordinate to some other option, then original must likewise be subordinate to this other option. And, in this case, ASO will imply that we should avoid both EXTRA and ORIGINAL. Hence, it is impossible for ASO to

imply that we should avoid EXTRA without also implying that we should avoid ORIGINAL. Thus, in contrast with Temkin's revised principle, it is impossible for ASO to favor ORIGINAL over EXTRA.

Lastly, ASO + PAPD avoids the Problem of Improvable Life Avoidance. For this problem arises in cases with the following structure:

(i) There is a possible group of people, G, such that we have an option (call it PREVENTION) that involves not creating group G, and we also have a finite number of options (call them the *Creation Options*) that involve creating group G.

(ii) None of the Creations Options is best for everyone in G. That is, for every Creation Option, there is at least one alternative Creation Option in which someone in G fares better.

(iii) All of the available options are neutral with respect to their effects on independently existing people, and they are likewise neutral with respect to non-person-affecting considerations.

As we have seen, Temkin's Revised Principle, as well as the weakened versions thereof, all imply that, in such a choice situation, we should always choose PREVENTION, so as to minimize comparative harm. By contrast, ASO + PAPD never has this implication. To see why not, consider the Creation Options. It follows from condition (i) that there are only finitely many such options. And if there are finitely many Creation Options, then it follows that there must be at least one Creation Option that is not subordinate to any other Creation Option. For the only way there could be a finite set of Creation Options such that every outcome in this set is subordinate to some other outcome in this set would be if the subordination relations formed a cycle. But this is impossible. To see why, let us suppose, for the sake of reductio, that there were such a cycle. That is, suppose there were a sequence of outcomes O1 through On such that O1 is subordinate to O2, … , which is subordinate to On, and On is subordinate to O1. It follows from the definition of the subordination relation that O2 must be directly or indirectly defeated by On, which in turn is defeated by O1. And so it follows that O2 must be indirectly defeated by O1. And this contradicts our supposition that O1 is subordinate to O2.

Thus, since there can be no cycle of subordination relations, it follows that at least one of the Creation Options must not be subordinate to any of the other Creation Options. Call this Creation Option X. It's easy to show that Option X cannot be subordinate to PREVENTION. For Option X could be subordinate to PREVENTION only if it were directly or indirectly

defeated by PREVENTION. And this would require that there be at least one Creation Option that is directly defeated by PREVENTION. However, given condition (iii), above, PREVENTION could defeat a creation option only if PREVENTION were better for someone in G. And this is ruled out by the fact that no one in G exists in PREVENTION. Therefore, PREVENTION does not defeat any of the Creation Options, and so Option X cannot be subordinate to PREVENTION. And so it follows that Option X is not subordinate to any other option. And so ASU does not imply that we should avoid Option X. Consequently, ASU does not imply that we should avoid all the Creation Options. Hence, unlike Temkin's Revised Principle, ASU does not imply that we should choose PREVENTION.

Thus, there is much to be said for the conjunction of ASO and PAPD, as a way of taking person-affecting considerations into account. For, as we have seen, this combination of views has significant explanatory power and avoids a number of difficulties facing other person-affecting views. Of course, further exploration may well reveal that ASO + PAPD itself faces difficulties, and requires further revision. Nonetheless, I believe that the strengths of this combination of views suggests that we can make progress on a very important issue that Temkin's book raises but does not answer, namely the question of how we are to make rational choices in contexts in which the defeat relations among the available options are intransitive.

6. Rethinking Intransitivity

What do person-affecting principles tell us about the *better-than* relation? In particular, is Temkin right that such principles give us reason to question the transitivity of the *better-than* relation?

The answer to this question will depend on *which* of these principles we accept. Temkin's first such principle, PAP, has very strong implications concerning the *better-than* relation. Consider, once more, the following three outcomes:

A: Only Tom and Dick exist. Tom has welfare level 10 and Dick has welfare level 5.

B: Only Dick and Harry. Dick has welfare level 10 and Harry has welfare level 5.

C: Only Harry and Tom exist. Harry has welfare level 10 and Tom has welfare level 5.

As we have seen, PAP entails that C is better than B, that B is better than A, and that A is better than C. Hence, PAP entails that the *better-than* relation is intransitive. And yet these

implications concerning A, B, and C are precisely what led both Parfit and Temkin to abandon the original formulation of the Person-Affecting Principle.

What about the other formulations of the Person-Affecting Principle that solve the problem of Symmetric Intransitivity? None of these principles are stated in terms of the better-than relation. Consider, for example, Temkin's Revised Principle. This formulation is explicitly about what we should be "aiming at" in comparing options. Hence, it appears to be about *how we should choose* among alternatives. And this principle never implies any intransitivity in how we should rank outcomes within a single choice situation. To the contrary, it implies that, within any given choice situation, we can always give a fully transitive ranking among the available outcomes, in terms of the degree of comparative harm they contain, relative to the set of available outcomes. The only kind of intransitivity that is directly implied by the Revised Principle is a kind that extends across different choice situations, namely the following:

> **Cross-Context Choice-Theoretic Intransitivity:** There are trios of outcomes, O1, O2 and O3 such that one should choose O2 when the only alternative is O1, O3 when the only alternative is O2, and O1 when the only alternative is O3.

Indeed, when it comes to the various versions of the person-affecting principle we have considered that avoid the problem of Symmetric Intransitivity (namely RPAP, WRPAP, and ASO + PAPD), Cross-Context Choice-Theoretic Intransitivity is the only kind of intransitivity they directly imply.

And the existence of Cross-Context Choice-Theoretic Intransitivity doesn't imply anything about the intransitivity of the better-than relation. For such intransitivity can arise even in contexts in which it is clear that the better-than relations among the outcomes in question are fully transitive. As an illustration, let's consider a modified version of a trolley case from Frances Kamm.[5] Suppose that Trisha, the benevolent train switch operator, knows that, unless she does something to stop him, then Condorcet, the evil trolley conductor, will drive his trolley down track A, killing two innocent people. There are two ways in which she can prevent him from doing so. First, she could wait until he has begun his homicidal journey, and then redirect the train onto track B, where it will kill only one innocent person, Innis. Alternatively, before Condorcet begins his journey, she could shoot Innis in the knee caps with a shot gun. If she does

[5] This example is inspired by Kamm. See esp. p. 26ff.

so, then Condorcet will be so appalled by the gory site that he will lose his desire to kill, and refrain from driving the trolley down either track. Thus, there are three possible outcomes:

W: Trisha does nothing, and the two innocent people on track A are killed.

X: Trisha redirects the train onto track B, and Innis is killed.

Y: Trisha shoots off Innis's knee caps, and no one is killed.

It's very plausible that, if Trisha knows with certainty that her only options are W and X, then she should choose X, since she should prefer for one innocent person to be killed rather than two. And if she knows with certainty that her only options are X and Y, then she should choose Y, since she should prefer severely injuring Innis to killing him. And yet it's also very plausible that, if she knows with certainty that her only options are W and Y, then she should choose W, since it is not permissible for her to permanently maim Innis, and cause him excruciating agony, as a means to saving two lives. And so these three outcomes instantiate Cross-Context Choice Theoretic Intransitivity.

If this intransitivity is to be explained by the intransitivity of the better-than relation, then it must be that, while Y is better than X and X is better than W, Y is not better than W. But this is false, for Y *is* better than W. After all, Y involves less harm than W (an injury rather than two deaths), and Y also involves less wrongdoing than W (a maiming carried out by Trisha rather than a double murder carried out by Condorcet). We therefore need another explanation of the Cross-Context Choice-Theoretic Intransitivity. Fortunately, such an explanation is available. For we can maintain that, while outcome Y is better than outcome W, it would nonetheless be wrong for Trisha to choose Y given a choice between W and Y, since doing so would violate a *claim* of Innis's not to be treated in such a way.

It is natural to suppose that a similar explanation might be available for the instances of Cross-Context Choice-Theoretic Intransitivity that are implied by the person-affecting principles, such as the intransitivity involving outcomes A, B, and C. In particular, one might propose the following:

> **Deontological Hypothesis**: When, purely in virtue of person-affecting considerations, one should choose an outcome O1 over an outcome O2 when these are the only available alternatives, this is to be explained not in terms of O1 being better than O2, but rather in terms of an *agent-centered restriction* prohibiting one from choosing O2 over O1. And this agent-centered restriction is itself to be explained in terms of *claims* of those who would fare worse given the choice of O2.

Consider, for example, a case in which Chelsea is choosing between A and B. If one adopts the Deontological Hypothesis, then one will maintain that what explains why she should choose B is not that B is *better than* A, but rather that the only person whose welfare is differentially affected, namely Dick, has a special claim on Chelsea in this choice situation. Just as one might maintain that the harms *Trisha* commits should weigh more heavily in Trisha's deliberations than the harms *Condorcet* commits, despite the fact that Trisha's harms don't contribute any more the overall values of outcomes, so one might maintain that, since Dick is the only person whose welfare is differentially affected by Chelsea's choice, *his* welfare should weigh more heavily in Chelsea's deliberation than the welfare of Tom or Harry, despite that fact that Dick's welfare contributes no more to the overall values of the outcomes. This could easily explain why Chelsea should choose B over A. And similar explanations could be given for why one should choose C over B, and A over C. Hence, the Deontological Hypothesis can explain Cross-Context Choice-Theoretic Intransitivity without appealing to, or implying, the claim that the better-than relation is transitive.

Thus, the kind of intransitivity implied by the person-affecting principles will support the claim that the better-than relation is intransitive only if we reject the Deontological Hypothesis, and adopt in its place a view on which the relevant choice-theoretic relations correspond to better-than relations. The simplest such view is the following:

> **Simple Axiological Hypothesis**: In cases where, purely in virtue of person-affecting considerations, one should choose an outcome O1 over an outcome O2 when these are the only available alternatives, O1 is *better simpliciter* than O2.

The problem with this view, however, is that it gives rise to the problem of Symmetric Intransitivity. For it implies that C is better simpliciter than B, that B is better simpliciter than A, and that A is better simpliciter than C.

Fortunately, Temkin does not appear to endorse the Simple Axiological Hypothesis. Rather than maintaining that B is better than A simpliciter, he maintains that B is better than A *relative to a choice situation in which A and B are the only two available alternatives* (2012, p. 431). Thus, he appears to endorse the following:

> **Relativized Axiological Hypothesis**: **Simple Axiological Hypothesis**: In cases where, purely in virtue of person-affecting considerations, one should choose an outcome O1 over an outcome O2 when these are the only available alternatives, O1 is better than O2 relative to the set of alternatives {O1, O2}

While this Relativized Axiological Hypothesis does not imply that the better-than relation is intransitive in the ordinary sense, it does imply the following, weaker claim:

> **Cross-Context Intransitivity of Better Than**: There are trios of outcomes, O1, O2 and O3 such that O1 is better than O2 relative to the set of alternatives {O1, O2}, O2 is better than O3 relative to {O2, O3}, and O3 is better than O1 relative to {O1, O3}.

And this would itself be an interesting and important result.

How, then, are we to adjudicate between the Deontological Hypothesis and the Relativized Axiological Hypothesis? In order to do so, we would need an account of the distinction between axiological and deontological reasons for choice. And Temkin himself suggests such an account. He says that the claim that an outcome O1 is better than an outcome O2 "reflects the judgment that from an impartial perspective there would be most reason to prefer the former outcome to obtain rather than the latter and, analogously, greater reason, from that perspective, to regret an outcome where [O2] obtained rather than O1, other things equal" (p. 11). And the "impartial perspective," he tells us, can be thought of as corresponding to the perspective of an ideal observer "who does not himself have any deontological or agent-relative reasons to favor one outcome over another" (2012, p. 10).

On this view, when O1 is better than O2, this means that there is *agent-neutral* reason to prefer O1 to O2. And so this suggests that what distinguishes between axiological reason deontic reasons is that the former are agent-neutral. If one should choose O1 over O2 because O1 is better than O2, then one's reason for preferring O1 over O2 is a reason that everyone shares. By contrast, if one should choose O1 over O2 because of an agent-centered restriction, then one's reason for preferring O1 over O2 will not apply to other agents who are not faced with such a choice.

On the face of it, this account of axiological reasons would appear to favor the Deontological Hypothesis over the Relativized Axiological Hypothesis. For the reasons to which person-affecting considerations give rise do not appear to be agent-neutral. For we have seen that, if we are to take person-affecting considerations seriously while avoiding the Problem of Symmetric Intransitivity, then we must maintain that, while B is preferable to A in a context where only A and B are available, A is *not* preferable to B in a context where A, B, and C are all available. Thus, if one person is in the first kind of choice situation, and another person is in the second kind of choice situation, then the first person will have reason to prefer B to A, while the

second person will have no such reason. And so it seems the first person's reason for preferring B to A must be agent-relative. But if the reasons to which person-affecting considerations give rise are agent-relative, whereas better-than relations correspond to agent-neutral reasons, then it seems that person-affecting considerations can't give rise to better-than relations.

There is, however, a reply to this argument, namely the following:

> You are conflating *agent-relativity* and *alternative-relativity*. While the reasons to which person-affecting considerations give rise are indeed alternative relative, they are nonetheless agent-neutral. Thus, if person-affecting considerations favor the choice of outcome O1 over outcome O2, then, in any situation in which O1 and O2 are the only available alternatives, *all agents* will have reason to prefer that O1 comes about, not just an agent who is choosing between O1 and O2. And because *all agents* have reason to prefer that O1 come about in such a situation, it makes sense to say that O1 is *genuinely better* than O2 relative to {O1, O2}.

This response turns on the following claim:

> **Agent-Neutrality Thesis**: If there is person-affecting reason to choose an outcome O1 over an outcome O2 then, in any situation in which O1 and O2 are the only two possible outcomes, everyone will have reason to prefer O1 to O2.

If this claim is correct, then we will have strong reason to accept something like the Relativized Axiological Hypothesis. And, conversely, if this claim is false, then we will have strong reason to reject any such axiological view, and adopt something like the Deontological Hypothesis. And so the relevance of person-affecting consideration to the *better-than* relation seems to turn on the plausibility of this claim.

While I don't have a knock-down argument against this claim, I think it can be shown that this claim has counterintuitive implications. To see why this is so, it will be useful to focus on cases where, whether outcome O1 or O2 comes about is determined by a *mindless mechanism*, rather than by a free choice. For it is in such cases that the Deontological Hypothesis and the Relativized Axiological Hypothesis most clearly come apart. (In cases where *an agent is choosing* between O1 and O2, the proponent of the Deontological Hypothesis might maintain that there is agent-neutral reason to prefer that the agent choose the outcome favored by person-affecting reasons, on the ground that the opposite choice would involve wrongdoing, and there is agent-neutral reason to prefer that wrong-doing be avoided).

Let's begin by considering a fairly simple case. Suppose that one or another of A, B, and C is guaranteed to obtain, and there is some mindless mechanism that determines which of these outcomes obtains, with the result that each outcome has an objective chance of 1/3. In this case, it seems clear that an ideal observer should be indifferent as to which of these outcomes is brought about, and so she should be indifferent as to how the mechanism operates.

But now let's add some more details to the story. Let's suppose the mechanism works as follows: a six-sided die is cast, and then three coins are tossed. And suppose that which of A, B, and C is brought about is determined as follows:

(1) If the die comes up 1 or 2, and the 1st coin comes up Heads, then A obtains.
(2) If the die comes up 1 or 2, and the 1st coin comes up Tails, then B obtains.
(3) If the die comes up 3 or 4, and the 2nd coin comes up Heads, then B obtains.
(4) If the die comes up 3 or 4, and the 2nd coin comes up Tails, then C obtains.
(5) If the die comes up 5 or 6, and the 3rd coin comes up Heads, then C obtains.
(6) If the die comes up 5 or 6, and the 3rd coin comes up Tails, then A obtains.

It seems clear that learning all these details about the functioning of the mindless mechanism should not change the fact that an observer should not care how this mechanism behaves. The ideal observer should be just as indifferent concerning the operations of this mechanism after learning how it works as she was beforehand. Thus, she shouldn't care in the slightest about the outcomes of the die roll or of the coin tosses.

Note that, in the mechanism described above, the first coin makes a difference only if the die comes up 1 or 2—if the die comes up any other number, then the first coin has no effect. And so it follows that an observer's *unconditional* preferences concerning the first coin toss should be the same as her preferences concerning this coin toss *conditional on the die coming up 1 or 2*. Hence, since the observer should be indifferent concerning this coin toss unconditionally, she must likewise be indifferent concerning this coin toss conditional on the die coming up 1 or 2. And if she should initially be indifferent concerning this coin toss conditional on the die coming up 1 or 2 then, *upon learning* that the coin has come up 1 or 2, she should remain indifferent concerning this coin toss.

But this result conflicts with the Agent-Neutrality Thesis. For, as all the person-affecting principles agree, there is person-affecting reason to choose B over A. And so the Agent-Neutrality Thesis implies that, in any situation where A and B are the only two possible

outcomes, there is agent-neutral reason to prefer B over A. But once the die has come up 1 or 2, outcome C is eliminated, so and A and B are the only possible outcomes. Ergo, the Agent-Neutrality thesis implies that, once that die has come up 1 or 2, there will be agent-neutral reason to prefer B to A. And since there is no competing reason to prefer A to B, it will follow that an ideal observer should prefer B to A all things considered. But in the context in question, where the die has come up 1 or 2, A will obtain just in case the first coin comes up Heads, and B will obtain just in case the first coin comes up Tails. Hence, since the Agent Neutrality thesis implies that an ideal observer should prefer B to A, this thesis likewise implies that an ideal observer should prefer for the first coin to come up Tails rather than Heads.

Thus, the defender of the Agent-Neutrality Thesis must reject the claim I defended earlier, namely that, upon learning that the die comes up 1 or 2, the ideal observer should be indifferent between the first coin's coming up Heads and the first coin's coming up Tails. And in order to reject this claim, she must maintain one of the following:

(i) At the outset, even though an ideal observer should be indifferent among outcomes A, B, and C, she should not be indifferent concerning the outcomes of the coin tosses that determine which of A, B and C obtain, or

(ii) While an ideal observer should initially be indifferent concerning the outcomes of the coin tosses, she should form such preferences as soon as she learns the outcome of the die roll.

Since each of these claims is very hard to believe, I conclude that there is strong reason to be skeptical of the Agent-Neutrality Thesis. And if we are skeptical of the Agent-Neutrality Thesis, then we should likewise be skeptical of the axiological hypotheses, and so we should be skeptical of Temkin's view that person-affecting considerations bear on the better-than relation.

If this is right then, at least insofar as Temkin's arguments turn on person-affecting considerations, these arguments may not require us to rethink *the good* per se. Nonetheless, these arguments will require us to rethink the *moral landscape* in very radical ways. For they indicate the need to develop new moral theories that give person-affecting considerations their due, as well as new theories of practical reasoning that allow us to address the complexities that such considerations introduce into our decision making.

References

Kamm, F. M. 2007. *Intricate Ethics*. New York: Oxford University Press.

Norcross, Alistair. 1999. "Intransitivity and the Person-Affecting Principle." *Philosophy and Phenomenological Research* 59 (3): 769-776.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press.

Temkin, Larry. 1987. "Intransitivity and the Mere Addition Paradox." *Philosophy and Public Affairs* 16 (2): 138-187.

Temkin, Larry. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. New York: Oxford University Press.