Australasian Journal of Philosophy
Vol. 78, No. 1, pp. 1–15; March 2000

Should Consequentialists Make Parfit's Second Mistake? A Refutation of Jackson[1]

Ben Eggleston

I conclude that consequentialists should make the Second Mistake.
—Frank Jackson, 'Which Effects?'[2] (p. 52)

I. Introduction

The 'mistake' that Jackson has in mind is one of the five 'mistakes in moral mathematics' that Derek Parfit discusses in chapter 3 of *Reasons and Persons*.[3] The five 'mistakes' are, according to Parfit, misunderstandings that a person can have regarding what should be considered the consequences of an act, or regarding how the moral assessment of an act depends on its consequences. In identifying and correcting these five 'mistakes', then, Parfit aims to improve our moral assessments of acts by improving our accounting of the consequences of acts.

In his paper, Jackson discusses the principle that Parfit identifies as the second 'mistake'. Rejecting Parfit's indictment of it as a mistake, Jackson advocates it as a principle that consequentialists should embrace. In this paper, after briefly reviewing Parfit's account of the second 'mistake', I examine and refute Jackson's critique of it, showing that each of the three arguments that Jackson offers is unsound. I conclude by abstracting from the particular arguments Jackson offers and focusing on the conclusion he aims to establish: that consequentialists should not regard the second 'mistake' as a mistake. I argue that this conclusion cannot be maintained except on an overly narrow, and hence distorted, understanding of what consequentialism is.

II. The Second 'Mistake'

The second 'mistake' Parfit discusses is to consider an act in isolation from other acts with which it is connected by assuming, as Parfit puts it, the following:

[1] I want to thank David Gauthier and Dale Miller for their helpful comments on previous versions of this paper. Though each commented extensively on passages throughout the paper, Gauthier's influence is especially strong in section V; Miller's, in section IV.

[2] Frank Jackson, 'Which Effects?', in Jonathan Dancy (ed.), *Reading Parfit* (Oxford: Blackwell Publishers, 1997), pp. 42–53. In most instances, simple parenthetical references are to pages (and sometimes notes) of this work.

[3] Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984). In most instances, references to this work are in the following form: (*R&P*, p. 70).

> (The Second Mistake) If some act is right or wrong *because of its effects*, the only relevant effects are the effects of this particular act. (*R&P*, p. 70)

Unfortunately, the way in which Parfit expresses this 'mistake' does not seem to be what's needed in order for it to be construed as a mistake. For what could the phrase '*its effects*' refer to, if not the effects of this particular act? So it seems logically unavoidable that when some act is right or wrong *because of its effects*, the only relevant effects are the effects of this particular act. But if this is the case—that the second 'mistake' is actually a tautology—then one can hardly be mistaken in subscribing to it.

Now in claiming that the second 'mistake' is a mistake, Parfit surely does not mean to say that there's a certain tautology that it's a mistake to embrace. On the contrary, he must have in mind, as the second 'mistake', something other than what he says in the passage discussed above.[4] In order to understand what Parfit conceives of as the second 'mistake', we need to understand the second 'mistake' to be a principle to which it's possible for Parfit to launch an objection that is not logically false, even if we do not later join Parfit in finding the objection to be convincing. To meet this need, let us follow Jackson (p. 52, n. 3) in inferring that the 'its' is there erroneously, so that the second 'mistake' is the following:

> (The Second Mistake) If some act is right or wrong *because of effects*, the only relevant effects are the effects of this particular act.

Clearly, this understanding of the second 'mistake' yields a principle to which one can object without necessarily making a logical error, since it is logically consistent to deny the inference from 'effects' to 'effects of this particular act'. (For example, and to jump ahead a bit, one may claim that the effects of a *set* of acts of which this act is a member are relevant to the rightness or wrongness of *this* act.) But while we need to understand the second 'mistake' as something that it's reasonable to object to (as we have just done), we also need to understand it as something that it's reasonable—or, at least, tempting—to subscribe to. Otherwise, the second 'mistake' will be such an obvious mistake that it's superfluous for Parfit (or anyone else) to bother to expose it as a mistake. But Jackson's interpretation meets this need, too, since it *is* tempting (for consequentialists, at least) to think that only this act's effects matter to the moral assessment of this act. Indeed, it is precisely this principle that we should understand Parfit to be objecting to.

To see the upshot of making the second 'mistake', consider an example in which the effects of an act differ from the effects of a set of acts of which it is a member:

> *Case One. X* and *Y* simultaneously shoot and kill me. Either shot, by itself, would have killed. (*R&P*, p. 70)

Parfit observes that it is true that 'neither X nor Y harms me', since neither makes me worse off than I would be if he were to act differently. (Of course, if *both* were to act

[4] And other than what he says on p. 443, where he repeats *verbatim* the statement of the second 'mistake' given above.

differently, I would be better off because I would not die; but given the other's conduct, neither makes me worse off just by putting an extra bullet in me.) X doesn't harm me, and Y doesn't, either. If we make the second 'mistake', we conclude that neither X nor Y acts wrongly.

But Parfit calls this conclusion 'absurd'. He claims that in order to avoid this conclusion, we should accept the following principle, which he calls (C7):

> Even if an act harms no one, this act may be wrong because it is one of a *set* of acts that *together* harm other people. Similarly, even if some act benefits no one, it can be what someone ought to do, because it is one of a set of acts that together benefit other people. (*R&P*, p. 70)

Thus 'X and Y act wrongly because *they together* harm me. ... On any plausible theory, even if each of us harms no one, we can be acting wrongly if we together harm other people' (*R&P*, p. 70). Not to realise this is to make the second 'mistake'.

III. Jackson's First Argument

According to Jackson, however, it is Parfit who is mistaken. Jackson argues that from the strictly consequentialist perspective which Parfit purports to be elucidating, the most sensible assessment of Case One is that 'neither X nor Y act [*sic*] wrongly' (p. 42)—the assessment which, as we have seen, Parfit calls 'absurd'. In order to argue that consequentialists should *not* 'embrac[e Parfit's] view that the Second Mistake is a mistake', Jackson presents 'Three Reasons for Consequentialists to Deny the Intuition that X and Y Both Act Wrongly' (p. 45).

Jackson's first reason is premised on his tripartite classification of acts into 'right, wrong and neither (neutral)' (p. 47). Having introduced this, Jackson claims the following, which I call Jackson's tripartite principle: 'If benefiting makes an act right and harming makes it wrong, then surely doing nothing makes it neutral' (p. 47). Now Parfit and Jackson agree that neither X nor Y benefits or harms in Case One, meaning—according to Jackson—that 'neither X nor Y acts wrongly (and neither acts rightly) in Case One' (p. 47).

III.1

But Jackson's tripartite principle is flawed in a couple of ways. First, it is perfectly consistent with consequentialism to maintain that every act is either right or wrong, with none being morally neutral. As an example, consider the view that in any given situation, an act is right if its consequences are at least as good as the consequences of any other act available in that situation, and wrong otherwise.[5] This view, while obviously consequentialist, also obviously rejects the assessment of some acts as morally neutral. In

[5] This view is found in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1963). See p. 45, where Smart says that 'the right action for an agent in given circumstances is, we have said, that action which produces better results than any alternative action'.

his discussion, Jackson doesn't specify this or any other particular view as the target of his criticism, but given the possibility of such a view, it is hard to see how Jackson is then moved to claim that to deny his tripartite principle 'is to make zero special in an essentially arbitrary way' (p. 47). The 'zero' Jackson has in mind is that associated with acts that neither benefit nor harm. But clearly, the view just sketched does not assess such acts in a special way. Rather, it deems such acts right if neither benefiting nor harming is the best consequence available in the situation, and wrong otherwise. So, denying Jackson's tripartite principle doesn't necessarily involve 'mak[ing] zero special' at all, whether arbitrarily or otherwise. It follows that a consequentialist is free to reject Jackson's tripartite principle outright.

III.2

A second, independent, reason equally allows consequentialists to reject Jackson's principle. Recall that it says that 'If benefiting makes an act right and harming makes it wrong, then surely doing nothing makes it neutral'. Now since an act that literally does *nothing* is hardly an act, properly so-called, it seems fair to assume that what Jackson means by 'doing nothing' is 'not producing any benefits or harms'. On this reading, Jackson's principle means that an act that neither benefits nor harms is neither right nor wrong. But in saying this, Jackson's principle presupposes that the only effects that matter to the moral assessment of an act are the effects of that particular act—and this presupposition is precisely what Parfit identifies as the second 'mistake'. Indeed, in calling the second 'mistake' a mistake, Parfit means to say that an act can be made right or wrong by things other than its own effects; and in (C7) he gives an example of one of these other things: membership in a *set* of acts that *together* produce benefits or harms. (He would obviously add, were the issue raised, that such membership prevents an act from 'doing nothing' insofar as 'doing nothing' makes an act morally neutral.) And since Jackson's principle not only implicitly presupposes that the second 'mistake' is not a mistake but also flouts (C7), which is meant by Parfit as a corrective to the second 'mistake', it is no stretch to conclude that Jackson's principle is, itself, an *instance* of what Parfit means by the second 'mistake'. Clearly Jackson's inference relies, as a matter of logic, on the presupposition that the second 'mistake' is *not* really a mistake—thereby rendering his first argument fatally question-begging.

III.3

One final component of Jackson's first argument that should be considered is his claim that his tripartite principle is all but endorsed by Parfit himself in the text of *Reasons and Persons*. Given (as we have just seen, in section III.2) that Jackson's principle rests on what Parfit regards as a *mistake,* it may seem unlikely that Parfit's text would support it; but we should consider Jackson's claim on its own merits. In context, his claim appears as follows:

> If benefiting makes an act right and harming makes it wrong, then surely doing nothing makes it neutral. To suppose otherwise is to make zero special in an essentially arbitrary way. Indeed, there is textual evidence that Parfit himself might be sympathetic to this line of thought. (p. 47)

The textual evidence Jackson adduces involves two more of Parfit's examples: Case Two and Case Three. In Case Two,

> X tricks me into drinking poison, of a kind that causes a painful death within a few minutes. Before this poison has any effect, Y kills me painlessly. (*R&P*, p. 70)

Recall that in Case One, X and Y kill me simultaneously, though neither X nor Y is necessary for this outcome. So Case Two is like Case One, except that in Case Two, X and Y don't kill me *simultaneously*. In Case Three,

> As before, X tricks me into drinking poison of a kind that causes a painful death within a few minutes. *Y* knows that he can save *your* life if he acts in a way whose inevitable side-effect is my immediate and painless death. Because Y also knows that I am about to die painfully, Y acts in this way. (*R&P*, p. 71)

Jackson points out that he and Parfit agree that Y acts rightly in Case Three, 'because Y benefits someone else and does not harm me' (p. 47). Jackson next claims that 'there is no relevant difference between Case Two and Case Three', meaning that 'the right thing to say about Case Two is that Y acts rightly' (p. 47). And from this it follows (according to Jackson) that Y should not be faulted for his behaviour in Case One, either. The crux of Jackson's claim that the three cases are relevantly similar is found in his characterisation of Parfit's treatment of Case Three:

> [Parfit] says that 'since Y's act is not worse for me it is morally *irrelevant* that Y kills me' [p. 71 of Parfit, Jackson's emphasis]. It seems that it is making worse and making better that are the morally relevant considerations for being wrong and being right, respectively, in which case neither X nor Y acts wrongly (and neither acts rightly) in Case One. (p. 47)

So Jackson's treatment of Case One hinges on its similarity to Case Two, and on Case Two's similarity to Case Three. But Case Two is *not* similar to Case Three in the way in which Jackson needs for it to be similar. Certainly one way in which they *are* similar is that in neither case does Y harm me. But there is an important dissimilarity: in Case Two, Y's conduct is independent of X's, but in Case Three, it is not—Y would not kill me if my death were not already imminent. (This is implied in the description of the cases, and made explicit in Parfit's discussion of them.) As a result, Y's act in Case Two is blameworthy in a way that Y's act in Case Three is not. It's this difference between the two cases that leads Parfit to conclude in regard to Case Two that '(C7) implies correctly that X and Y act wrongly because they together harm me' but in regard to Case Three that 'Y is doing what he ought to do' (*R&P*, p. 71). Jackson's misinterpretation of Parfit, then, lies in thinking that just because Parfit would agree that (in Jackson's words, just quoted) 'it is making worse and making better that are the morally relevant considerations' in Case Three, he would also agree that these are the *only* morally relevant considerations that can ever present themselves. And this leads Jackson to overlook what Parfit considers to be another morally relevant consideration (namely, the way in which Y's conduct is independent of X's) that differentiates the first two cases from the third.

Admittedly, the discussion in the previous paragraph of Parfit's treatment of the three cases does not suffice to show that Parfit's judgments in regard to them are *correct*; it only indicates *what* Parfit's judgments *are*. But this is precisely at issue in this component of Jackson's first argument, since he claims (on p. 47) that 'there is textual evidence that Parfit himself might be sympathetic to this line of thought'—i.e., the line of thought discussed earlier in this section, particularly Jackson's tripartite principle that 'If benefiting makes an act right and harming makes it wrong, then surely doing nothing makes it neutral'. But just as Jackson flouts (C7) in his discussion of his tripartite principle, so does he ignore it in assuming that Parfit would be sympathetic to treating the three cases alike. As a result, Jackson's first argument fails to undermine Parfit's indictment of the second 'mistake'.

IV. Jackson's Second Argument

Jackson's second objection to the judgment that both X and Y act wrongly is that such a judgment 'runs counter to the whole thrust of consequentialist thinking about morality' (p. 47). The thrust to which Jackson refers is 'the denial of agent-relative values' (p. 48)—the insistence, roughly, that what is important from the moral point of view is what consequences are brought about, not who brings them about. This thrust is illustrated in the characteristically consequentialist judgment that if I am in a situation in which my only options are to kill one person myself or to let both that person and another person be killed by someone else, then I ought to kill the one myself, since one death is better than two. Jackson goes on to say,

> In and of itself, who does the killing is irrelevant in the consequentialist picture. But it is exactly this which a consequentialist who thinks that X and Y act wrongly in the overdetermination case [such as Parfit] must deny. For such a consequentialist holds that X ought not to shoot, but the difference between shooting and not shooting in the overdetermination case is precisely a difference in who brings my death about. (p. 48)

But this argument is unsound, as we can see in a couple of ways.

IV.1

First, let us pause to review what makes theories agent-relative and agent-neutral. Consider Parfit's account of this distinction, which Jackson explicitly endorses (p. 53, n. 6):

> C [which stands for Consequentialism] might claim that it would be worse if there was more deception or coercion. C would then give to all of us two common aims. We should try to cause it to be true that there is less deception or coercion. Since C gives to all agents common moral aims, I shall call C *agent-neutral*.
>
> Many moral theories do not take this form. These theories are *agent-relative*, giving to different agents different aims. It can be claimed, for example, that each of us should have the aim that he does not coerce other people. On this view, it would be wrong for me to coerce other people, even if by doing so I could cause it to be true that there would be less coercion. Similar claims might be made about deceiving or betraying others. On these claims, each person's aim should be, not that there be less deception

> or betrayal, but that he himself does not betray others. These claims are not Consequentialist. (*R&P*, p. 27)

Now as Jackson maintains, the denial of agent-relative values precludes us from holding that in the case Jackson sketches, my aim should be that I myself do not kill; this is because this aim is relative to me, the agent. Rather, we should hold that my aim should be that there be less killing. This aim is, obviously, agent-neutral. Similarly, the denial of agent-relative values precludes us from holding that in Case One, X's aim should be that he himself does not kill. But we do not have to hold that X's aim should be *this*, or any other agent-relative aim, in order to maintain that in Case One, X ought not to shoot. Rather, we can specify an agent-neutral aim for X—such as that there be less killing, that there be less shooting of other people, that there be less violence in general against other people, or any of an indefinite number of alternatives.

Admittedly, the issue is complicated by the fact that Case One is a case of overdetermination. Can we specify an agent-neutral aim that matters that X fails to promote when he shoots me? Deaths presumably matter, but when he shoots me, X does not cause there to be any more deaths than there would otherwise be. And although in shooting me X does cause there to be more shooting of other people and more violence in general against other people, the stipulation that X doesn't harm me might seem to imply that these don't matter. As a result, it might seem plausible to claim, as Jackson does, that there is no way for X's act to be criticised from the perspective of agent-neutral values.

But these last two inferences, tempting though they may be, are groundless. The fact that X doesn't harm me does not mean that the agent-neutral aims that X fails to promote do not matter. Which aims matter and which ones don't is a question to be addressed by a substantive theory of the good; and the best theory of the good may tell us that some of the agent-neutral aims that X fails to promote do matter. It may tell us, for example, that violence matters, even when it's not harmful. This, then, is one way in which a consequentialist may hold that X ought not to shoot, without compromising her denial of agent-relative values.

Moreover, there is another way. Even granting (as Jackson would apparently insist) that there is *no* agent-neutral aim that X's act actually lessens the realisation of—that X's act actually has *no* bad consequences—consequentialists may still maintain that X's act is wrong. Consequentialists certainly *may* hold the view that an act is wrong only if it actually has bad consequences; indeed, due to its simplicity, this view is the most obvious and probably the most well-known version of consequentialism, and I suspect that this causes Jackson's argument to look more defensible than it really is. But in fact, consequentialists are also free to hold, instead, a more complicated view of the connection between the good and the right. They are free to say, for example, that an act is wrong if it is an act of a sort that *tends* to have bad consequences—Why, then, are they not free to say, as Parfit does, that an act is wrong if it is one of a *set* of acts that together have bad consequences? To forbid this last possibility is arbitrary and, as before, question-begging.

IV.2

Still, when we return to the passage from Jackson quoted above, it may seem as compelling as before. It may be helpful, then, for us to examine it more closely, for in doing so we may divest it of its appeal. Let us begin by representing Jackson's argument in the following way:

P1 No true consequentialists think that who does the killing is relevant.

P2 Anyone who thinks that X and Y act wrongly in Case One thinks that who does the killing *is* relevant.

C Therefore, no true consequentialists think that X and Y act wrongly in Case One.

Focus on the two premises: each involves the idea of relevance. Because relevance is a relational concept (we can always ask, 'Relevant to what?'), it has many senses. Let us distinguish two. One is *relevance to the question of which outcome is better*. We may call this *outcome-relevance*. Another is *relevance to the moral assessment of agents' conduct*. We may call this *conduct-relevance*.

The first premise, then, can have either of two senses. If it means that all true consequentialists think that who does the killing is not outcome-relevant, then it is surely correct: this is what consequentialism's denial of agent-relative values means. But if it means that all true consequentialists think that who does the killing is not conduct-relevant, then it is incorrect, as Jackson's own case makes clear. In that case, who does the killing is relevant to the moral assessment of my conduct: if I do the killing—i.e., if I pull the trigger—then my conduct is (paradoxically) better than if I do not do the killing. Similarly, in Case One, when X shoots, a consequentialist may find that X's conduct is worse than if he were not to shoot, on the grounds that his shooting thwarts the pursuit of some desirable agent-neutral outcome (of the kind specified in section IV.1). It follows that the first premise is true only if relevance means outcome-relevance.

Now if Jackson's argument is to be valid, then relevance must mean outcome-relevance in the second premise, too. But this makes the second premise false: to think that X and Y act wrongly in the overdetermination case does *not* require one to think that who does the killing is relevant to the question of which outcome is better. All it requires one to think is that who does the killing is relevant to the moral assessment of agents' conduct: that who does the killing is conduct-relevant. So, the second premise is true only if relevance means conduct-relevance.

Jackson's argument, then, trades on an ambiguity in one of its central terms. As stated, it is invalid. If it is made valid through the removal of the ambiguity, then one of its premises is made false. In any case, it is unsound.

V. Jackson's Third Argument

Jackson's third objection to holding that X and Y each act wrongly begins with the distinction between 'overdetermination proper' and 'causal pre-emption' (p. 48). According to Jackson, if Case One were a case of preemption instead of overdetermination—if, for example, X's bullet were to kill me just a moment before Y's bullet were to arrive instead of the bullets' killing me simultaneously—then 'it [would not be] plausible that Y acts wrongly' (p. 49). From this Jackson infers that Parfit's judgment that X and Y each act wrongly depends crucially on the fact that Case One is a case of overdetermination, not a case of preemption. But, as Jackson points out, what makes Case One a case of overdetermination and not a case of preemption is the mere fact that the bullets kill me simultaneously. Otherwise—if 'one bullet does the deadly work a moment before the arrival of the other' (p. 49)—then it's a case of preemption. As Jackson puts it, the fact that Case One is a case of overdetermination and not a case of preemption 'depends on the fine detail of what happens inside my body' (p. 49). By extension, Parfit's judgment that X and Y each act wrongly depends on this 'fine detail'. And according to Jackson, this is what's objectionable about Parfit's understanding of Case One: Parfit's 'answer depends on what precisely happens inside me, in a way which could be accommodated within the deontologist's framework, but which is hard to make plausible from the consequentialist perspective' (p. 49).

V.1

Jackson does not explain *how* 'what precisely happens inside me' is a consideration more alien to the consequentialist perspective than to the deontological one, so one can only speculate as to his reasons for holding this view. Surely one of his reasons is *not* that consequentialist judgments of particular cases characteristically do *not* require much precision in matters of contingent fact. For it is commonly urged as a criticism against consequentialism that it requires impossibly close attention to details of contingent fact, whereas the deontological perspective grounds the moral assessment of an act not in such details, but in whether it is an act of 'this or that kind'.[6] Admittedly, ascertaining an act's 'kind' may involve as much attention to details of contingent fact as does any consequentialist inquiry, but typically it does not. After all, would a deontologist assessing the acts in Case One be concerned with whose bullet, if either, arrived first? Probably not: as long as the two bullets' moments of arrival are sufficiently close for each of X and Y to understand himself to be shooting and killing me, a deontologist is likely to judge that each acts wrongly, even if one is preempted by the other from causing my death. Clearly, considerations along these lines do not help to furnish a credible backing for Jackson's view.

Perhaps a better reason for holding Jackson's view is that such an inquiry (as to 'what precisely happens inside me', as to whether Case One is a case of overdetermination or a case of preemption) is *superfluous* from a consequentialist perspective. One might find such an inquiry superfluous from a consequentialist perspective because whatever the outcome of the inquiry is, no new light is thereby shed on the consequence of X and Y's conduct: either way, I'm dead. But insofar as the consequentialist perspective is taken up for the moral assessment not only of consequences but also of *conduct*, such an inquiry is eminently relevant. For a consequentialist assessment of conduct involves ascertaining not only what consequences are brought about, but also *how* they are brought about: by *what acts* they are brought about. In Case One, it may well matter, for a consequentialist assessment of Y's conduct, whether Y's shot contributes to my death as much as X's shot does (as in the overdetermination scenario) or harmlessly adds a bullet to my already-expiring body (as in the preemption scenario). For Parfit's judgment of Case One to turn on such 'fine detail', then, hardly threatens his consequentialist credentials.

[6] This phrase is found in David Lyons, *Forms and Limits of Utilitarianism* (Oxford: Clarendon Press, 1965), p. vii.

V.2

But a deeper problem with Jackson's third argument is his assumption that Parfit's judgment *does* turn on the fact that Case One is a case of overdetermination and not a case of preemption. As noted above, Jackson says that if Case One is a case of preemption instead of overdetermination, then 'it is not plausible that Y acts wrongly' (p. 49). But would Parfit agree?

Suppose that, for the sake of argument, we go along with Jackson and impute to Parfit the judgment that in the preemption scenario, Y does not act wrongly. Presumably, Parfit's basis for such a judgment would be the observation that Y doesn't harm me, since X single-handedly kills me first. In other words, the reason that Y does not act wrongly would be that (given X's conduct) Y doesn't harm me. But a person who reasons in this way would take note of the observation that X doesn't harm me, since Y is about to act in such a way that would single-handedly kill me if X doesn't kill me first, and the amount of time by which X shortens my life is negligible. And such a person would be forced to conclude, on pain of contradiction, that X does not act wrongly, either, since (given Y's conduct) X doesn't harm me.

Parfit would surely reject this last judgment (that X doesn't harm me in the preemption scenario), and so he must also refuse to assent to the initial judgment (that Y doesn't harm me in the preemption scenario). Indeed, it seems reasonable to expect that in the preemption scenario, he would blame both X and Y, just as he does in the overdetermination scenario. For in each scenario, all that X or Y can claim by way of justification is that his conduct causes no harm, given what the other is doing; but if this doesn't absolve them in Parfit's eyes in the overdetermination scenario, it's hard to see why it would absolve them in his eyes in the preemption scenario.

V.3

Finally, there is direct textual support for the attribution to Parfit of similar judgments in the overdetermination and preemption scenarios of Case One. Recall Case Two, in which

> X tricks me into drinking poison, of a kind that causes a painful death within a few minutes. Before this poison has any effect, Y kills me painlessly. (*R&P*, p. 70)

This is like Case One, except that it's a case not of overdetermination, but of preemption: X's act of killing me gets preempted by Y's. Parfit acknowledges that in this case, neither X nor Y harms me. But he goes on to say (as quoted above, in section III.3) that '(C7) implies correctly that X and Y act wrongly because they together harm me'. Since Parfit condemns X and Y in this case of preemption, it's far-fetched to think that Parfit's condemnation of X and Y in Case One depends on its *not* being a case of preemption.

How, then, could Jackson have thought otherwise? Well, recall the textual support that Jackson claims to find for his *first* argument. In particular, recall (from section III.3) Jackson's claim that when one considers Parfit's treatment of Case Three, 'It seems that it is making worse and making better that are the morally relevant considerations for being wrong and being right'. Applying this principle to a case of preemption, it's natural to infer that a preempted agent does not act wrongly, since a preempted agent is, by definition, kept from making worse or better. But as we noted in section III.2, Parfit countenances at least one morally relevant consideration aside from making worse or better—namely, membership in a set of acts that together make worse or better. Here, as in his first argument, Jackson overlooks this feature of Parfit's view, imputing to him a view that, in fact, he rejects.

Jackson's third argument is that Parfit's judgment in regard to Case One depends on its being a case of overdetermination instead of preemption, and this is a reason for consequentialists to reject Parfit's judgment that both X and Y act wrongly. But (as we have just seen) Parfit's view of Case One does *not* depend on its being a case of overdetermination instead of preemption; and (as argued earlier in this section) even if it did, that should not trouble consequentialists.

VI. Conclusion: Consequentialism's Big Tent

So Jackson's three arguments fail to show that consequentialists should reject Parfit's claim that the second 'mistake' is a really a mistake. The failure of Jackson's arguments does not, of course, imply that what Jackson claims is false; it could still be true that consequentialists ought to embrace the second 'mistake', just for reasons better than those provided by Jackson. But Jackson's discussion of the second 'mistake' is thoughtful and careful, evincing a close and resourceful reading of Parfit's text. As a result, I believe we should take its failure seriously and, rather than assuming that Jackson's arguments can be replaced with better ones, entertain the hypothesis that no arguments can do what Jackson intends for his to do.

VI.1

This hypothesis gains plausibility in the light of several features of consequentialist thought. For one thing, there are many different versions of consequentialism, and many of them are incompatible with other, equally consequentialist, versions of consequentialism.[7] Furthermore, there are many ways in which versions of consequentialism can come into conflict: rather than there being just a single dimension of variation, with every disagreement between versions of consequentialism being traceable to a difference of position along this one dimension of variation, there are many dimensions of variation, a difference of position along *any* of which can give rise to a disagreement between

[7] To find examples, one needn't go farther than some leading historical sources. Three views, each unique, are found in Jeremy Bentham's 1789 *An Introduction to the Principles of Morals and Legislation*, ed. by J. H. Burns and H. L. A. Hart (Oxford: Clarendon Press, 1996); John Stuart Mill's 1861 *Utilitarianism*, in Mill's *Essays on Ethics, Religion and Society*, ed. by J. M. Robson (Toronto: University of Toronto Press, 1969; volume X of *Collected Works of John Stuart Mill*); and Henry Sidgwick's 1907 *The Methods of Ethics*, seventh ed. (Indianapolis: Hackett Publishing Company, 1981). For some of Mill's reflections on Bentham specifically, including some pointed jabs (regarding 'the great fault I have to find with Mr. Bentham as a moral philosopher', how 'Mr. Bentham's writings...have done...very serious evil', and 'his first disqualification as a philosopher'), see his 1833 'Remarks on Bentham's Philosophy' and his 1838 'Bentham', both in Mill's *Essays* (quotations from p. 7, p. 15, and p. 91, respectively). For Moore's views of Bentham and Mill, see not only his *Methods*, but also his 1902 *Outlines of the History of Ethics for English Readers* (Indianapolis: Hackett Publishing Company, 1988), especially pp. 239–50 ('Bentham and His School' and 'J. S. Mill'). In twentieth-century philosophy, versions of consequentialism (utilitarian and otherwise) have only proliferated (see note 8)..

versions of consequentialism.[8] And as if all this isn't enough, the *dimensions* of disagreement themselves seem to defy any sort of non-arbitrary, comprehensive classification.[9] Such disorder within consequentialism is possible only because the notion of consequentialism itself is, from a logical point of view, very weak. And while its weakness makes it flexible enough to accommodate the diversity of consequentialist theories in circulation, its weakness also prevents it from providing the resources—the substantive propositional commitments—that are needed in order for arguments like Jackson's to succeed. There may be good reasons for rejecting a claim such as Parfit's indictment of the second 'mistake', but such reasons should not be sought in the consequentialist perspective *per se*.

What can we find in the consequentialist perspective *per se*? I take consequentialism to be the view that acts are right and wrong in virtue of the consequences they produce.[10] In Parfit's words, it's the view whose 'central claim' is that 'There is one ultimate moral aim: that outcomes be as good as possible' (*R&P*, p. 24). (As Parfit explains on p. 3, he calls 'aims' the things that theories 'tell us to try to achieve'.) Parfit adds that

[8] An obvious source of conflict among consequentialists is the question of what makes consequences good. One answer is the modern-day Benthamic one offered by T. L. S. Sprigge in *The Rational Foundations of Ethics* (London: Routledge & Kegan, 1987), p. 223: 'experience which is felt to be good in the actual living of it'. Another answer is found in R. M. Hare, *Moral Thinking: Its Levels, Method, and Point* (Oxford: Clarendon Press, 1981), p. 104: the maximal satisfaction of 'our present preferences'. A third is found in G. E. Moore, *Principia Ethica*, revised ed., ed. by Thomas Baldwin (Cambridge: Cambridge University Press, 1993), p. 237: 'certain states of consciousness, which may be roughly described as the pleasures of human intercourse and the enjoyment of beautiful objects'.

A second question is whether the good is to be *maximised*. Some suggest not; see Michael Slote, part I of Michael Slote and Philip Pettit, 'Satisficing Consequentialism', *The Aristotelian Society* supplementary volume 58 (1984), pp. 139–63. Even those who favour maximisation then offer conflicting answers to a third question: What is to be chosen in a maximising way? By default, acts are often considered the object of maximising choice, but other possibilities include practices (see John Rawls, 'Two Concepts of Rules', *Philosophical Review* 64 (1955), pp. 3–32), rules (see R. B. Brandt, 'Some Merits of One Form of Rule-Utilitarianism', *University of Colorado Studies, Series in Philosophy No. 3: The Concept of Morality* (Boulder, Colorado: University of Colorado Press, January 1967), pp. 39–65), motives (see R. M. Adams, 'Motive Utilitarianism', *The Journal of Philosophy* 73 (1976), pp. 467–81), and courses of action (see Eric B. Dayton, 'Course of Action Utilitarianism', *Canadian Journal of Philosophy* 9 (1979), pp. 671–84).

Fourth, although we normally think of *universalistic* consequentialism, *egoistic* consequentialism is an option, too. Sidgwick regarded the choice between these as 'the profoundest problem of Ethics' (*Methods*, p. 384, n. 4); also see Jesse Kalin, 'Two Kinds of Moral Reasoning: Ethical Egoism as A Moral Theory', *Canadian Journal of Philosophy* 5 (1975), pp. 323–56. A few pages later in the same journal, R. I. Sikora's 'Utilitarianism: the Classical Principle and the Average Principle' (pp. 409–19) reminds us of a fifth dilemma for consequentialists (non-egoistic ones, at least). I omit discussion of further questions, on matters ranging from moral psychology to meta-ethics (and beyond).

[9] In this vein it is worth comparing the questions I list in note 8 to the nine questions listed by Philip Pettit in 'Introduction', Philip Pettit (ed.), *Consequentialism* (Aldershot, England: Dartmouth, 1993), pp. xiii–xix. Although we cover much the same ground, we map it rather differently.

[10] Essentially this characterisation of consequentialism is given in the unsigned entry for consequentialism in Simon Blackburn (ed.), *The Oxford Dictionary of Philosophy* (Oxford University Press, 1994), p. 77; in James P. Griffin's entry for consequentialism in Ted Honderich (ed.), *The Oxford Companion to Philosophy* (Oxford University Press, 1995), pp. 154–56; and in Dan W. Brock's entry for utilitarianism in Robert Audi (ed.), *The Cambridge Dictionary of Philosophy* (Cambridge: Cambridge University Press, 1995), pp. 824–25.

Some consequentialists also think that *each* act is right or wrong in virtue of the consequences *it* produces. But this view is a stronger view than that of consequentialism *per se* and can be inferred consequentialism also says that 'What each of us ought to do is whatever would make the outcome best', and that 'If someone does what he believes will make the outcome worse, he is acting wrongly'[11] (*R&P*, p. 24). In regard to what makes outcomes good and bad, Parfit is studiously agnostic—as we must be, in order to avoid prematurely discarding viable versions of consequentialism. We must join Parfit in allowing, for example, that 'Consequentialists...may...believe that, in some cases, the best outcome is not the one in which people are benefited most' (*R&P*, p. 77).

I give the foregoing account (of how the odds are stacked against Jackson) not only to give Jackson's discussion its due and to provide a general explanation of its failure that complements the piecemeal analysis offered in sections III, IV, and V, but also to suggest the futility of arguing for the conclusion that Jackson's discussion seeks to establish: that the rejection of the second 'mistake' is incompatible with consequentialism. The considerations just adduced show that a great many things are compatible with consequentialism, and it should be no surprise that proscribing the second 'mistake' is among them.

VI.2

Still, that something should be no surprise does not mean that it is the case. Can consequentialism, in fact, accommodate Parfit's indictment of the second 'mistake'? To see that it can, begin by considering Jackson's claim that 'An act is objectively right if it *in fact* makes things better – that is, benefits' (p. 49). Now this is undoubtedly a perfectly natural thing for a consequentialist to think, and it is the kind of thing that gives Jackson's defence of the second 'mistake' some of its intuitive appeal. But it isn't mandatory for a consequentialist to think this: we must heed Parfit's warning (just quoted, in section VI.1) that 'Consequentialists...may...believe that, in some cases, the best outcome is not the one in which people are benefited most'. This means that a consequentialist could well think that it is obligatory to abstain from some acts that, in fact, benefit.[12] Mill, for example, claims that it can be obligatory to abstain from an act whose 'consequences in the particular case might be beneficial', with 'the ground of the obligation to abstain from it' being that 'the action is of a class which, if practised generally, would be generally

[10] *continued . . .*
from the latter only by way of a fallacy: the 'fallacy of decomposition', perhaps. Coincidentally, this fallacious inference is almost identical to Parfit's second 'mistake', which is that 'If some act is right or wrong because of effects, the only relevant effects are the effects of this particular act'. Note that in calling the inference fallacious, I do not mean to be joining Parfit in saying that the inferred view is false; I only mean to say that it is not implied by consequentialism *per se*. It could still, *pace* Parfit, be true.

[11] Parfit uses the phrase 'what *he believes* will make the outcome worse' (my emphasis) because, as he says on p. 372, 'in assigning blame, we must consider not only actual but predictable effects'.

[12] What does Parfit think? He prefaces (C6) by saying that 'I should act in the way whose consequence is that most lives are saved' (p. 69; I assume that he means '*the* most lives'); but in the very next paragraph, he says 'On any plausible moral theory, we should *sometimes* try to do what would benefit people most' (p. 69; my emphasis). Presumably, this latter claim is meant to allow that on some plausible moral theories, it's sometimes permissible to forgo the attainment of the best possible consequences. But it is unclear whether Parfit also thinks that on *every* plausible moral theory, it's sometimes permissible to forgo the attainment of the best possible consequences; and it is also unclear whether Parfit also thinks that on any or all plausible moral theories, it's sometimes *obligatory* to forgo the attainment of the best possible consequences.

injurious'.[13] Another consequentialist, taking her cue from Parfit's indictment of the second 'mistake', may claim something analogous: that it's sometimes obligatory to abstain from an act whose consequences in the particular case might be beneficial (or, of course, harmless), with the ground of the obligation to abstain from it being that the act is one of a set of acts that together do some harm. Such a claim is not only consequentialist but also compatible with Parfit's indictment of the second 'mistake'.

Now it may be objected that these versions of consequentialism—Mill's and the parallel one just devised to exemplify the avoidance of the second 'mistake'—are unsatisfactory. It may be objected, for example, that it's incoherent to regard a maximally beneficial act as wrong, on the grounds that it belongs to a class of generally harmful acts, or to a set of collectively harmful acts. (The objection allows that it's not incoherent to deem a maximally beneficial act wrong if things aside from benefits and harms—such as the keeping of promises and the creation of beautiful things—are included in one's account of the good. What the objection regards as incoherent is to cite a benefit-or-harm-related consideration as a reason, in some cases, to abstain from maximising benefits.) Objections like this one appear frequently in contemporary discussions of consequentialism.[14] But this very fact—the fact that such objections do figure prominently in the contemporary consequentialist dialogue (and, indeed, have so figured for some time)—shows that the issue is far from settled. On the contrary, versions of consequentialism against which the incoherence objection is urged represent serious and formidable proposals about how consequentialism should be understood and elaborated. So although attempts have been made to evict such views from the house of consequentialism, it would misrepresent the state of the literature to say that these attempts have succeeded. In this respect, consequentialism is a house divided. So, the claim that it's sometimes obligatory to forgo the attainment of the most beneficial outcome may be unsatisfactory, but not from the point of view of consequentialism *per se*. And it follows that consequentialism *per se* does not show Parfit's indictment of the second 'mistake' to be objectionable.

VI.3

To sum up: In sections III, IV, and V, we found some factors *internal* to Jackson's three arguments that keep them from succeeding—we found logical flaws because of which his arguments happen to fail. In this section, we have identified *external* factors because of

[13] John Stuart Mill, *Utilitarianism*, p. 220. I should acknowledge, however, that not even Mill's status as a consequentialist is undisputed. For an exceptionally thorough discussion, see Christopher Miles Coope, 'Was Mill a Utilitarian?', *Utilitas* 10 (1998), pp. 33–67, part of the abstract of which reads, 'It is even doubtful whether he [i.e., Mill] was a consequentialist in any sense'. But even if scrutiny of Mill's work reveals his thought to be an infelicific example of consequentialism, I assume that the position I ascribe to him could well be held by a consequentialist.

[14] Recent examples include Brad Hooker, 'Rule-Consequentialism, Incoherence, Fairness', *Proceedings of the Aristotelian Society* n.s. 95 (1995), pp. 19–35, and works referenced therein. Earlier examples include J. Harrison, 'Utilitarianism, Universalisation, and Our Duty to Be Just', *Proceedings of the Aristotelian Society* n.s. 53 (1953), pp. 105–34; J. J. C. Smart, 'Extreme and Restricted Utilitarianism', *The Philosophical Quarterly* 5 (1956); R. David Broiles, 'Is Rule Utilitarianism Too Restricted?', *Southern Journal of Philosophy* 2 (1964), pp. 180–87; and George C. Kerner, 'The Immorality of Utilitarianism and the Escapism of Rule-Utilitarianism', *The Philosophical Quarterly* 21 (1971), pp. 36–50.

which, in a sense, his arguments have to fail, since he intends for them to show something that cannot be shown. He intends for them to show that consequentialists, *qua* consequentialists, should not join Parfit in regarding the second 'mistake' as a mistake; and we have seen that the logical weakness of consequentialism *per se* makes this impossible.

None of this is to say, of course, that consequentialists *must* endorse Parfit's indictment of the second 'mistake'; a consequentialist may take it or leave it as she sees fit. And a consequentialist might commit herself to other claims that imply, whether she realises it or not, that the second 'mistake' cannot be a mistake. (Indeed I believe that Parfit so commits himself in his discussion of the first 'mistake', creating an inconsistency.[15]) All that the present paper shows is that it's not the case that consequentialism *per se* requires one to embrace what Parfit calls the second 'mistake', and that to make such a claim involves imputing to consequentialism more than can legitimately be inferred from it. To avoid *this* mistake, we have to avoid perceiving consequentialism to be narrower and more determinate than it really is, and must instead remember that consequentialism is an open and flexible doctrine that can be instantiated in a variety of ways. Remembering this is the key to seeing how consequentialists may, if they so choose, consistently join Parfit in regarding the second 'mistake' as a mistake.

University of Pittsburgh

Received: March 1999
Revised: July 1999

[15] See my 'Does Participation Matter? An Inconsistency in Parfit's Moral Mathematics' (unpublished).