Philosophical Studies (2005) 123: 149–173 © Springer 2005
DOI 10.1007/s11098-004-5220-3

ALASTAIR NORCROSS

HARMING IN CONTEXT

The Standard consequentialist approach to harm is illustrated by the following principle, defended by Derek Parfit:

(C6) An act benefits someone if its consequence is that someone is benefited more.
An act harms someone if its consequence is that someone is harmed more. (69)

How should we understand what it is for someone to be "harmed more"? Intuitively, it is for someone to be made worse off by the action. But worse off than what? With what do we compare the result of the action? One suggestion that we can quickly dispense with is that we should compare the welfare of the victim before and after the action. Consider a case involving a doctor, named 'Doctor' and a patient, named 'Patient'. Call this case *Doctor*: Patient is terminally ill. His condition is declining, and his suffering is increasing. Doctor cannot delay Patient's death. The only thing she can do is to slow the rate of increase of Patient's suffering by administering various drugs. The best available drugs completely remove the pain that Patient would have suffered as a result of his illness. However, they also produce, as a side-effect, a level of suffering that is dramatically lower than he would have experienced without them, but significantly higher than he is now experiencing. So the result of administering the drugs is that Patient's suffering continues to increase, but at a slower rate than he would have experienced without them. The very best thing she can do has the consequence that Patient's suffering increases. That is, after Doctor's action Patient is suffering N amount of suffering as a direct result of Doctor's action, and N is more than Patient was suffering before the action. Has Doctor *harmed* Patient if she slows the rate of increase of Patient's suffering as much as she can? This hardly seems plausible. It is a far more plausible description of this case to claim that Doctor has in fact greatly benefited Patient.

Clearly, we can't simply compare Patient's welfare before and after a particular action. Doctor *has* made Patient better off: not better off than he was, but better off than he would have been. We compare levels of welfare, not across times, but across worlds. Doctor has benefited Patient, because she has made Patient better off than he would have been had she done something else. Even though Patient is now suffering more than he was, he would have been suffering even more, if Doctor had done anything else instead. In thinking about harm, then, what most consequentialists, including Parfit, have in mind is the idea of someone being worse off than they would otherwise have been. Thus, we get the following:

HARM An act A harms a person P iff P is worse off, as a consequence of A, than she would have been if A hadn't been performed. An act A benefits a person P iff P is better off, as a consequence of A, than she would have been if A hadn't been performed.

At first glance, this seems perfectly straightforward. I shoot and kill you. As a consequence of my act, you are worse off than you would otherwise have been, that is, than you would have been if I hadn't shot you. However, on closer investigation, things turn out not to be so straightforward. I will investigate the standard consequentialist approach to harm, by focusing on cases in which it appears that a group of people can together harm someone, even though none of the members of the group harms anyone. I will examine Derek Parfit's approach to group harms, and argue that it is unsuccessful. I will argue for an alternative account of harm that applies both to individual acts and to group acts.

First, though, I will address a possible source of confusion. The notion of harm with which I am concerned, and which is of most interest to consequentialists, is the notion of harm *all things considered*, as opposed to harm in some respect or other. Let me illustrate this distinction with an example that might appear to challenge the consequentialist account of harm I have suggested.[1] Tonya and Nancy are rivals in the cut-throat world of competitive ice-skating. Tonya, seeking to ensure victory in an upcoming contest, attacks one of Nancy's knees with a baseball bat. Nancy's injuries keep her out of competition, and require extensive medical treatment. In the course of the treatment, a tumor, unrelated to the injury, is discovered and successfully treated. If Nancy hadn't been injured and undergone extensive tests, the tumor wouldn't have been discovered in time, and she would have suffered a painful and career-ending illness (her knee, though, would have been fine). It appears, then, that, as a consequence of Tonya's savage attack, Nancy is better off than she would otherwise have been, and thus that Tonya's attack didn't harm Nancy, but rather benefited her. But surely, we might object, Tonya's attack harmed Nancy. It severely mangled her knee. Doesn't this type of case show that HARM cannot be the correct account of harm and benefit? Not at all. Tonya's attack benefited Nancy in the long-term, while it certainly harmed her in the short-term. It also harmed Nancy's knee, while benefiting Nancy overall. We may be reluctant to judge that Tonya benefited Nancy, but that is because we believe, erroneously, that such a judgment must somehow redound to Tonya's credit. The reason why we correctly believe that Tonya's action doesn't ground a positive judgment of her character is, of course, that the benefits to Nancy were entirely unforeseen by Tonya. Suppose it were otherwise. Tonya is deeply concerned for Nancy's welfare, and knows that Nancy has a nascent tumor, although she is unable to persuade Nancy of this or to persuade her to undergo tests (perhaps Nancy, being somewhat snooty, suspects that Tonya is trying to undermine her confidence in order to gain a competitive advantage). Tonya realizes that only a severe knee injury will result in the discovery and treatment of Nancy's tumor. She therefore attacks Nancy, knowing that she herself will be caught, banned from competitive ice-skating, and even sent to jail. To this version of the story, we have no difficulty responding that not only did Tonya benefit Nancy, but that her action was an heroic piece of self-sacrifice.

While it is clear that HARM may be adapted to render restricted judgments about harm or benefit in certain respects (harm to the knee versus benefit to the person overall, harm in the short-term versus benefit in the long-term), it is not clear why these should be of interest to a consequentialist. Unless we take the implausible (at least from a consequentialist perspective) position that we have stronger reasons not to harm than we do to produce comparable benefits,[2] judgments of overall harm or benefit will provide the same reasons for action as the sum of all the relevant restricted judgments. Henceforth, therefore, I should be understood to be talking about harm all things considered, unless I explicitly say otherwise.

Derek Parfit claims, in *Reasons and Persons*, that an act that makes no difference to anyone's welfare can nonetheless be wrong "because it is one of a *set* of acts that *together* harm other people."[3] This claim (part of what Parfit calls '(C7)') is illustrated by the following case of overdetermination:

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

Even though neither X nor Y harms me, since I would have been killed by the other, even if one had not shot me, Parfit claims that it is absurd to conclude "that X and Y do not act wrongly". (70) He continues:

> X and Y act wrongly because *they together* harm me. They together kill me... On any plausible moral theory, it is a mistake in this kind of case to consider only the effects of single acts. On any plausible theory, even if each of us harms no one, we can be acting wrongly if we together harm other people. (70)

This is also supposed to apply in cases of preemption, where the acts are not simultaneous.[4] Parfit presents the following illustration:

> *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. (70)

Here, as with Case One, Parfit claims that neither X nor Y harms me, but also that they both act wrongly "because they together harm me. They together harm me because, if *both* had acted differently, I would not have died." (71) Against the objection that Y does harm me, since he kills me, Parfit presents:

> *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. (71)

In this case, Y not only doesn't harm me, but he acts as he ought to. Y doesn't harm me because, "if Y had acted differently, this would have made no difference to whether I died." (71) X, on the other hand, both harms me and acts wrongly, "because it is true that, if X had not poisoned me, Y would not have killed me." (71) Since Y affects me in the same way in Case Two as in Case Three, Y doesn't harm me in Case Two either. What makes Y's act wrong in Case Two is that "he is a member of a group who together harm me." (71) What prevents this being true of Y in Case Three? After all, it is true in Case Three that, if both X and Y had acted differently, I would not have been harmed. In response to this, Parfit points out that it is also true that, if X, Y, and Fred Astaire had all acted differently, I would not have been harmed. It doesn't, of course, follow that Fred Astaire is a member of a group who together harm me. We need a clearer account of what it is to be a member of a group who together harm, or benefit, others. Parfit attempts to provide such an account with:

> (C8) When some group together harm or benefit other people, this group is the smallest group of whom it is true that, if they had all acted differently, the other people would not have been harmed, or benefited. (71-2)

This group consists of X and Y in Case Two, but only of X in Case Three.

Parfit's (C7) and his treatment of Cases One, Two, and Three might appear to be in tension with his rejection, in the preceding section of *Reasons and Persons*, of what he calls the "Share-of-the-Total view", as it applies to the following case:[5]

> *The First Rescue Mission:* I know all of the following. A hundred miners are trapped. . . .If I and three other people go to stand on some platform, this

> will ... save the lives of these hundred men. If I do not join this rescue mission, I can go elsewhere and save, single-handedly, the lives of ten other people. There is a fifth potential rescuer. If I go elsewhere, this person will join the other three, and these four will save the hundred miners. (67–68)

The Share-of-the-Total view would have me help save the 100 miners, since my share of the credit would thus be 25 lives, as opposed to a mere 10, if I save the other 10 single-handedly. This, as Parfit points out, is clearly mistaken. What I should do is save the 10, since this will make a difference of 10 lives saved over helping to save the 100. If an act can be wrong because, although it doesn't harm, it is a member of a group that together harms, can't an act be right because, although it doesn't benefit, it is a member of a group that together benefits? Parfit has something to say about this. He presents:

> *The Third Rescue Mission.* As before, if four people stand on a platform, this will save the lives of a hundred miners. Five people stand on this platform. (72)

This case, Parfit says, demonstrates the need to add a further claim to (C8), because "there is not *one* smallest group who together save the hundred lives." (72) Parfit returns to this kind of case in a later section (section 30), where he claims that the crucial question in such cases concerns what an agent knows or has reason to believe. He offers the following additional principle:

> (C13) Suppose that there is some group who, by acting in a certain way, will together benefit other people. If someone believes that this group either is, or would be if he joined, too large, he has no moral reason to join this group. A group is too large if it is true that, if one or more of its members had not acted, this would not have reduced the benefit that this group gives to other people. (83)

Parfit is here talking of moral reasons in the subjective sense. To have a moral reason in this sense is to have a reason that is epistemically available in a robust sense. Such a reason may also be an objective reason. If, in (C13), the agent's belief about the size of the group is true, then he also has no objective moral reason to join the group. One can have objective moral reasons that are not epistemically available, in which case one doesn't have the corresponding subjective reasons. One can also have subjective moral reasons that don't correspond to objective reasons. If the agent's belief about the size of the group is false, she may have an objective reason to join the group, but no subjective reason. It is easy to see now that Parfit's notion of group harms and benefits, as articulated in (C7), (C8), and (C13), does not imply that I have a moral reason (either objective or subjective) to join the others in The First Rescue Mission, since it is part of the description of the case that I know all the relevant facts. In both Case One and Case Two, however, nothing is said about what X and Y know or believe (in contrast to Case Three). On the most intuitive readings of the cases, neither knows about the other's actions. In which case, both have subjective moral reason not to kill me. Even if they did know of the other's actions, their actions wouldn't be rescued from wrongness by (C13), which only concerns benefits and not harms. This asymmetry, though, appears *ad hoc*, so a charitable reading would have (C13) apply to harms and benefits.

Parfit's rejection of the Share-of-the-Total view is not inconsistent with (C7). A worry remains, however, about the motivation for his treatment of group harms and benefits. If the agents' belief states are as important in the ascription of moral reasons as (C13) would have them be, why not appeal directly to the fact that both X and Y believe themselves to be harming me to justify the judgment that what they both do is wrong? In fact, in section 10 of *Reasons and Persons*, Parfit, speaking of how he will use moral terms, declares that "*wrong* will usually mean *subjectively* wrong, or *blameworthy*." (25) Why does Parfit bother with (C7) and (C8), if he can do the job by a simple appeal to the distinction between subjective and objective wrongness?

I suspect that the actions of both X and Y in Case One and, perhaps, in Case Two, would be intuitively judged wrong, even if it were specified that each knew about the other's action. It is probably this intuitive judgment that Parfit is trying to capture in his notion of group harms. Furthermore, it seems clear that I am harmed in Case One. After all, I am killed. What could be a clearer case of harm than that? Since neither X nor Y individually harm me, according to (C6), we need an account of who does the harming.

Recall the difference between Y's action in Case Two and in Case Three. (C8) tells us that Y is a member of a group that harms me in Case Two, but not in Case Three, because, if X hadn't poisoned me, Y would still have shot me in Case Two, but not in Case Three. So it turns out that whether Y acts wrongly in the actual world depends on Y's behavior in certain possible worlds – most likely the closest worlds in which X doesn't poison me. This is a somewhat strange result. To see just how strange it is, let's consider some variations on Case Two.

Case Two-and-a-Quarter. As before, X tricks me into drinking the poison that will shortly result in my painful death. Before any of the real unpleasantness can set in, however, the slightly bitter taste of the poison prompts me to visit the nearest soda machine to purchase a Coke to take the taste away. Y is lurking by the soda machine waiting for a victim on whom to practice his assassination skills. Y shoots and kills me instantly and painlessly. If I hadn't drunk the poison, I wouldn't have visited the soda machine, and wouldn't have been shot. In fact, if I hadn't visited the soda machine, Y would have grown tired of waiting for a victim, and would have decided to become an accountant instead of an assassin.

If we apply (C8) to Case Two-and-a-Quarter, we get the result that the group who harms me consists just of X. If X had not poisoned me, I wouldn't have been harmed. And yet, the intuition that Y acts wrongly in Case Two applies equally to Case Two-and-a-Quarter, which is quite consistent with Case Two. We don't, in general, excuse behavior that appears to be wrong, if we discover that the agent wouldn't have had the opportunity to perform the wrong act, were it not for the seemingly unrelated behavior of someone else. Consider an even more challenging case.

Case Two-and-a-Half. X tells me a particularly funny joke. I laugh so much that I become hoarse, and visit the soda machine to purchase a Coke. Y, once again lurking to practice his assassination skills, shoots and kills me

> painlessly. The Coke that I had purchased, and was about to drink, had been infected with a deadly poison in a freak undetected soda canning accident.

Y doesn't harm me in this case, since I would have died of Coke poisoning, if he hadn't shot me. The smallest group who harms me consists of just X, since, if X hadn't told me the joke, I wouldn't have needed to visit the soda machine. In this case X harms me and Y doesn't. What is more, Y isn't even a member of a group who harm me (at least according to (C8)). It seems to follow that X acts wrongly and Y doesn't. Do we have to accept these rather counterintuitive results, if we follow Parfit's approach?

My cases Two-and-a-Quarter and Two-and-a-Half, and Parfit's Cases One and Two are underdescribed with respect to a crucial question. Does Y (and X in Case One) believe that he is harming me? That is, does Y know that I am about to die anyway? The most natural reading of the cases, as I said earlier, assumes that Y believes he is harming me. On this reading Parfit can say that Y acts wrongly in the *subjective* sense. However, as we have seen, Parfit's appeal to group harms can be at least partially motivated by the desire to judge both X's and Y's actions to be wrong in Case One and Case Two, on the assumption that they both know about the other's action. Let's consider, then, what to say about these cases, on the assumption that Y knows that I will die anyway. Parfit's appeal to (C7) and (C8) implies that the crucial difference between Y in Case Two and in Case Three is that his killing me in the latter case depends on X having already made my death inevitable, but in the former case it doesn't. In Case Three, if X hadn't already poisoned me, Y wouldn't have been willing to bring about my death in the course of saving you.

My examples, though, should make us suspicious of this claim. Y's killing me in both my examples depends on X having already made my death inevitable, and yet Y's behavior in my cases is intuitively judged on a par with case Two, not Case Three. Isn't the intuitively crucial difference between Case Two and Case Three that Y's killing me in the latter case depends on his *belief* that my death is already inevitable, whereas it doesn't in the former case? This raises an interesting question concerning Y's status in Case Three. According to Parfit, Y isn't a member of the group that harms me, because, according to (C8), that group consists of only X. If X hadn't poisoned me, Y wouldn't have killed me. However, suppose that Y's knowledge of X's poisoning me comes from overhearing X plotting to poison me. Suppose further that the closest possible world in which X doesn't poison me is one in which he changes his mind at the last minute, after Y has overheard him plotting. In that world, Y kills me anyway, because he still believes that I am going to die from poisoning. These details are quite consistent with Parfit's description of the case.[6] Is Y, in the actual world, a member of the group that harms me? It seems that he is, since it is not true of X that if he had acted differently, I would not have been harmed. What is worse, one plausible reading of the relevant counterfactual yields the result that Y alone is a member of the group that harms me. Suppose that the closest possible world in which Y doesn't kill me is one in which he doesn't overhear X plotting to poison me. This is because X *doesn't* plot to poison me (and doesn't poison me) in this world. So it is true of Y that, if he had acted differently, I would not have been harmed. We could attempt to block this move by banning the use of backtracking counterfactuals in applying (C8). When we consider what would have happened if someone had acted differently, we must suppose the world changed only at and after the time of the action. While this move may prevent Y from being the *only* member of the group that harms me, it doesn't address the previous point, that Y may be *a* member of that group, because his killing me depends not on X's already having made my death certain, but on Y's belief that this is so. In fact, banning backtracking counterfactuals blocks one line of defense against this result. When we consider the closest world in which X doesn't poison me, we can't consider a world in which he doesn't plot to poison me, since the plotting precedes the actual poisoning.

If, as I say, the intuitively crucial difference between Case Two and Case Three is that Y's killing me in the latter case depends on his *belief* that my death is already inevitable, whereas it doesn't in the former, we can see why Y's behavior in my two cases, Two-and-a-Quarter and Two-and-a-Half, is intuitively on a par with his behavior in Case Two. In my cases, even though Y would not have killed me if X hadn't made my death inevitable, he would still have killed me, if he hadn't *believed* that my death was already inevitable. That is, given that I was at the soda machine, Y's belief that I was about to die anyway (In Case Two-and-a-Half suppose that I had already drunk the tainted Coke by the time Y spotted me) played no part in his decision to shoot me. Suppose that it did. Suppose that, in Case Two-and-a-Quarter, Y is a sniper in a crack troop of marines who are only used against vicious terrorists who are holding innocent hostages. About to be dispatched on a vital mission, Y discovers that his aim is off. He needs practice on human targets, but he would never willingly harm an innocent person. Along I come, about to die painfully anyway. Y would never have considered shooting me, if he hadn't known this. How are we now to judge Y's behavior in shooting me? Far from being wrong, it appears now to be quite admirable. Not only doesn't it involve harm (it actually saves me from a painful death), but it is motivated by a desire to save the lives of innocent people. Furthermore, it considerably improves the chances of saving such lives.

The appeal to group harms seems to be motivated by a desire to accommodate the intuition that both X and Y act wrongly in Case One, even if each knows that the other will shoot. It is not clear, though, that utilitarianism can't accommodate such intuitions as are worth accommodating without an appeal to group harms.[7] So, how *should* a utilitarian judge Case One, on the assumption that both X and Y believe that the other will shoot? At first glance, it would appear that their actions are both objectively and subjectively right (assuming that they couldn't have been doing something better, if they hadn't been shooting me). Both these judgments can be challenged, though. X may believe that Y is going to shoot me anyway, but it is unlikely that he believes that there is *no chance* that Y will fail to kill me. Perhaps Y's gun will jam, or I will make an unexpected movement at the moment of shooting, or Y will simply change his mind. No matter how small these chances may be, X is not entitled to ignore them completely. X is doing what he expects will make my death even more likely than it already is. This is enough to make his action subjectively wrong. (The same applies to Y.) Similar things can be said about one type of objective wrongness. If the world is not completely deterministic, there are objective probabilities other than 1 and 0. If the right act is the act with the highest objective expected utility, it is unlikely that X's and Y's actions will be right. It seems that the utilitarian can judge both X and Y to have acted wrongly, without appealing to the problematic notion of group harms. However, the original spirit of the problem can be resurrected, if we modify Case One.

Consider Case One modified to Case One-and-a-Half as follows. Suppose that X's shooting me has the side-effect of curing the paralysis in the left leg of an innocent child, Suzie. Suppose further that Y's shooting me has the side-effect of curing the paralysis in Suzie's right leg. If neither shoots me, Suzie will be permanently unable to walk. If both shoot me, Suzie will live a physically normal life. If only one shoots me, Suzie will walk with the aid of a crutch. Now, suppose that both X and Y know of the side-effect of their own act of shooting, but neither knows that the other even exists. If they both shoot me, we can say that they both act wrongly in the subjective sense. They do what they believe will make things worse, assuming that curing the paralysis in one leg does not justify killing an innocent person. However, each act is right in the objective sense, since it didn't harm me, given that the other also shot me, and it made things better for Suzie. One important difference between Case One and Case One-and-a-Half is that X's and Y's actions in the latter are right in the objective expected utility sense, given certain reasonable probability estimates. Suppose that there is only a 1% chance that Y will not shoot. Suppose further that the chance of X's shot curing Suzie's paralysis in her left leg is at least 99.99%. It is plausible now that the objective expected utility of X's shooting me is higher than that of any alternative not involving shooting me.

The tiny chance that I would not have died if he hadn't shot me is outweighed by the near certainty that Suzie's leg will be cured if he does shoot me. Perhaps someone will object that a one in a hundred chance of death is not outweighed by the near certainty of a cure for a paralyzed leg. If so, we can adjust the probabilities accordingly. Someone may object that *any* chance of death outweighs a cure for a paralyzed leg. Such a view is not only highly implausible,[8] it is also not likely to be held by a utilitarian.

Now, suppose that X and Y both know that the other will shoot me, and what the effects of both actions will be. How do we intuitively judge them? That depends on what role their knowledge of the other's action played in their decision. Suppose first that neither is influenced in his decision by the knowledge of what the other will do. Even though he knows that he is making things better than they would have been, that is not why he acts as he does. He would have shot me anyway, even if he had not known of the other's existence. The normal intuitive judgment, in this case, may well be that they both act wrongly. Even though a utilitarian would have to say that their actions are right both objectively and subjectively, she can still render a negative judgment on their characters. Since they are both quite willing to do what they believe will make things much worse, the fact that they don't believe their current actions to make things worse doesn't excuse them. The utilitarian thus explains the common intuitive judgment (if there is one) that X and Y both *act* wrongly as the result of the all too common confusion of judgments of actions with judgments of character. If we know all the facts about the agents' motivations, we can see that their actions, in some sense, spring from bad characters, even though the very same actions could have come from good characters. A consequentialist account of character will ultimately connect evaluations of character with evaluations of actions. For example, we might say that a character trait, (C1) is better than another, (C2), just in case the possession of (C1) makes one likely, *ceteris paribus*, to perform better actions than does the possession of (C2).[9] Given the limited plasticity of human nature, it is clear that a good character trait may sometimes lead to an action that is significantly suboptimal, and that a bad character trait may sometimes lead to performing the best, or close to the best, action available.

Suppose now that both X and Y know that the other will shoot me, and that is why they are also prepared to shoot me. If X believed there was much of a chance of Y not shooting me, he wouldn't shoot me. So X's shooting me depends on his belief that Y will also shoot me, and the same goes for Y. Each shoots me because he believes (correctly) that he is not thereby harming me (or that the risk of harming me is very small), and that he is greatly benefiting Suzie. Each does what is subjectively right and objectively right, both as regards actual results and objective expected utility. Furthermore, neither appears to display a bad character, at least not what a utilitarian should call a bad character. The willingness to shoot someone who is going to be shot at the same time anyway, in order to achieve a great benefit – curing a paralyzed leg – seems to be an admirable character trait from a utilitarian perspective. Perhaps this last judgment could be challenged, on the grounds that the possession by both X and Y of this character trait on this occasion makes the world worse. If neither had this character trait, I wouldn't have been shot. However, this is a rather peculiar perspective from which to assess character traits. It is tied both to this occasion, and, even more peculiarly, to regarding X and Y as a group. If only X hadn't had this character trait, the world would have been worse. I would still be dead, and Suzie would have been paralyzed in one leg. The same goes for Y. Given the limited plasticity of human nature, the appropriate perspective for assessing character traits will not be one that is tied to highly specific situations.[10] It appears, then, that a utilitarian should judge both X's and Y's behavior positively, unless she appeals to (C7) with its problematic notion of group harms.

Why *not* judge X's and Y's behavior positively? Given all the assumptions of the preceding paragraph about their beliefs and motivations, it's not clear that there's even a common-sense moral intuition that they act wrongly, or that they display bad character. The problem, of course, is that X's and Y's supposedly admirable behavior leads to a worse state of affairs than would have resulted if they had both acted badly, according to utilitarianism.

Recall that Parfit's explanation of Case One, which would also apply to Case One-and-a-Half, is that, although neither X nor Y individually harms me, they harm me as a group. Although we have seen that Parfit's account of group harms is fraught with problems, the intuition remains that the group consisting of X and Y does harm me. How else do I end up being harmed?

I have argued that Parfit's account of group harms runs into some serious problems. Part of the reason for this is that he builds his account from the standard consequentialist approach, as given by HARM:

HARM An act A harms a person P iff P is worse off, as a consequence of A, than she would have been if A hadn't been performed. An act A benefits a person P iff P is better off, as a consequence of A, than she would have been if A hadn't been performed.

There are, however, some obvious problems with HARM, considered just as an account of individual harms. HARM approaches the question of whether an act harms by comparing the world in which it occurs with a world in which it doesn't occur. But which world in which it doesn't occur is the relevant one?

Perhaps the intuitive reading of HARM involves a comparison with the world in which the agent is inactive. When we ask whether P would have been worse off if the act hadn't been performed, we are considering a world in which the agent simply doesn't exercise her agency. So, what is it not to exercise one's agency? One obvious possibility is to remain completely immobile. But this clearly won't do. Consider the following case, that I will call *Button pusher*: An agent, named Agent stumbles onto an experiment conducted by a twisted scientist, named Scientist. He is seated at a desk with one hundred buttons, numbered '0' through '99', in front of him. He tells her that the buttons control the amount of pain to be inflicted on a victim, named Victim. If no button is pressed within the next 30 s, Victim will suffer excruciating agony. If the button marked '99' is pressed, Victim will suffer slightly less; if '98' is pressed, Victim will suffer slightly less, and so on down to '0', which will inflict no suffering on Victim. He was, he explains, about to sit and watch as Victim suffered the maximum amount. However, to honor her arrival, he turns control of the buttons over to Agent. She is free to press any button she wishes, or to press none at all. Agent pushes '99', inflicting almost maximal suffering on Victim. If she had remained immobile, Victim would have suffered even more. According to HARM, then, Agent's act doesn't harm Victim, and even benefits him, since he would have suffered even more, if she hadn't pushed any button. But surely her act doesn't benefit Victim. It led to excruciating agony for him, when he needn't have suffered at all. She could have pressed '0' instead. If any act harms, it seems clear that this one does.

Perhaps we should compare the results of Agent's act with what would have happened if she hadn't even been on the scene. There seem to be two ways to interpret this suggestion: (i) We imagine a world identical to the actual world before t, in which the agent miraculously vanishes from the scene at t; (ii) We imagine a world as similar as possible to the actual world before t, in which the agent is non-miraculously absent from the scene at t. That is, we imagine what would have had to have been different before t in order for the agent to have been absent at t. (i) Runs foul of *Button pusher*. If Agent had miraculously vanished, instead of pushing button '99', Victim would have suffered even more. But this consideration clearly doesn't incline us to judge that Agent's act benefited Victim. (ii) Seems more promising. How do I know whether I have harmed or benefited someone? I ask myself whether they are better or worse off than they would have been if I hadn't even been here in the first place. But this won't do, either. Once again, it gives the wrong results in *Button pusher*. If Agent hadn't even shown up in the first place, Scientist would have let Victim suffer the maximum amount, but we don't on that count judge Agent's act to benefit Victim.

In *Button pusher* Agent inflicts excruciating agony on Victim, but he would have suffered even more had Agent been inactive, either through immobility or absence from the scene. The problem is not just that inactivity gives unacceptable results in particular cases, but rather that the comparisons it invites do not seem relevant to whether an act harms or benefits. If I do something that seems to be very harmful, such as inflicting excruciating agony on someone with the press of a button, why should it matter that they would have suffered even more if I had been immobile or absent from the scene? Whether it is harmful to inflict pain on someone doesn't seem to depend on whether they would have suffered even more if I had been inactive, unless, perhaps, my inflicting such pain on them is the only alternative to more suffering. In *Button pusher*, however, Agent could easily have prevented Victim from suffering altogether. These counterfactuals, then, don't seem relevant to the question of whether an act harms or benefits.

There are other ways to read the counterfactuals in HARM, that will give different accounts of harm and benefit. The most obvious alternative reading involves a judgment about which other possible world is closest to the world in which the action occurs.[11] Instead of comparing the world in which the act occurs with a world in which the agent is either immobile or absent from the scene, we compare it with a world that is as much like it as possible, consistent with the act not occurring. Sometimes that will be the world in which the agent is immobile, but often it will be a world in which the agent does something else instead.

So, how does a standard possible worlds analysis of the counterfactuals in HARM hold up? Consider another example: suppose you witness the following scene at Texas Tech University: A member of the Philosophy department, passing Bobby Knight on campus, waves cheerily and says "Hey, Knight." Bobby Knight, turning as red as his sweater, seizes the hapless philosopher around the neck and chokes her violently, while screaming obscenities. By the time Bobby Knight has been dragged away, the philosopher has suffered a partially crushed windpipe and sustained permanent damage to her voicebox, as a result of which she will forever sound like Harvey Fierstein. Has Bobby Knight's act harmed the philosopher? The intuitive answer is obvious, and HARM seems to agree. The philosopher is much worse off than she would have been had Bobby Knight not choked her (unless, perhaps, she has always wanted to sound like Harvey Fierstein). But suppose we discover that Bobby Knight has recently been attending anger management classes. Furthermore, they have been highly successful in getting him to control his behavior. When he becomes enraged, he holds himself relatively in check. On this particular occasion (only the third violent outburst of the day), he tried, successfully, to tone down his behavior. In fact, if he hadn't been applying his anger management techniques, he wouldn't have choked the philosopher, but would rather have torn both her arms from her body and beaten her over the head with them. Since it took great effort on Bobby Knight's part to restrain himself as much as he did, it seems that the closest possible world in which he doesn't choke the philosopher is one in which she is even worse off. HARM, in this case, seems to give us the highly counterintuitive result that, not only does Bobby Knight's act of choking not harm the philosopher, it actually benefits her.

Now apply this reading of HARM to *Button pusher*. Let me add a couple of details to my previous description of the case. Agent delights in the suffering of others. She is initially inclined to press no button, so that Victim will suffer maximally, but she's dissatisfied that this will involve, as she sees it, merely *letting* Victim suffer, rather than actually *making* him suffer. She wants Victim to suffer as much as possible, but she also wants to make him suffer. At the last second she changes her mind, and pushes '99'. If she hadn't pushed '99', she wouldn't have pushed any button. She didn't even consider the possibility of pushing a different button. The only question she considered was whether she should *make* Victim suffer excruciating agony or *let* him suffer even more. Clearly, the closest world in which Agent doesn't push '99' is one in which she doesn't push any button, and Victim suffers even more. Once again, HARM judges Agent's act to benefit Victim. But we are no more inclined to believe that her act is beneficial than we were before we knew about her character defects. The fact that Agent's character made it highly probable that she would have done even worse than she did doesn't alter our intuitive judgment that her act of inflicting excruciating agony harms Victim. What is particularly disturbing for a consequentialist about this latest reading of HARM is that it makes the character of the agent relevant to whether the act harms or benefits.[12] The better the agent, the harder it is for her to benefit someone, and the worse she is, the easier it is.

It seems that none of the interpretations of HARM can provide the consequentialist (or anyone else) with a satisfactory account of what it is for an act to harm or benefit. The intuition on which they were based is that a harmful act makes someone worse off. The difficulty lies in producing a general formula to identify the particular possible world (or worlds), in comparison with which the victim is worse off, as a result of a harmful act. Any unified theory requires a way of fixing the contrast point, but the contrast point varies from situation to situation. Part of the problem is that our intuitions about whether particular acts harm or benefit are often influenced by features of the context that it would be difficult to incorporate into a general account.

The key to solving this puzzle is the realization that the fundamental consequentialist account of harm is an essentially comparative one. An act harms someone if it results in their being worse off than they would have been if some alternative act had been performed. Thus harm is always relative to some alternative. A particular act may harm me relative to one alternative and benefit me relative to another. There is no fundamental non-comparative moral fact of the form 'act A harms person X'. The fundamental moral facts are of the form 'act A results in a better world than alternative act B. The A-world is better than the B-world by a certain amount n, because person X is better/worse off in A than in B by amount n1, person Y is better/worse off in A than in B by amount n2,...' Our intuitive judgments about what acts *really* harm what people are explained by appeal to conversational context.

If I say that Booth's shot harmed Lincoln, the context selects, as an appropriate alternative act of Booth, pretty much anything else except shooting Lincoln. It may be true that Booth could have shot Lincoln in such a way as to lead to a much more agonizing death than the one he in fact suffered. This alternative, however, is normally not salient (and may never be). Likewise, in discussing Bobby Knight's behavior, very few (if any) contexts make the arm-tearing alternative the appropriate comparison. Even though it is, in some obvious psychological sense, more likely that Bobby Knight will tear the arms off the source of a perceived slight than that he will not assault her at all, when we discuss his behavior, we probably have in mind an alternative that would have been more likely for most other people. Sometimes different, equally normal, contexts can render one act a harming or a benefiting. For example, my father writes a will, in which I receive half his estate. This is the first will he has written. Had he died intestate, I would have received all of his estate. Two among his many other options were to leave me none of his estate or all of it. Does my father's act of will-writing harm me or benefit me? Imagine a conversation focused on my previous plans to invest the whole estate, based on my expectation that I would receive the whole estate. It might be natural in such a context to describe my father's act as harming me. I end up worse off than if he had left me all his estate, which I had expected him to do, either by not making a will at all, or by making one in which he left me the whole shebang. Imagine, though, a different, but equally natural, conversation focusing on my lack of filial piety and the fact that I clearly deserve none of the estate. In this context it may be natural to describe my father's act as benefiting me. After all, he *should* have left me nothing, such a sorry excuse for a human being I was.

At this point an objection may arise. Introducing the previous example, I said that different contexts can render one act a harming or a benefiting. Given that I am talking about harm *all things considered*, how can I claim that one act can be both a harming and a benefiting? Wouldn't this be contradictory? No. In order to see why not, we need to be precise about what I am committed to. I say that one act can be correctly described *in one conversational context* as a harming, and can be correctly described *in a different conversational context* as a benefiting. The reason why no contradiction is involved is that the claim 'act A harmed person P' can express different propositions in different contexts. On my suggested account of harm, to claim that act A harmed person P is to claim that A resulted in P being worse off than s/he would have been if *the appropriate alternative* to A had been performed. Given the context-relativity of *the appropriate alternative*, claims about harm (and benefit) have an indexical element. Just as 'today is a good day to die' can express different propositions in different contexts of utterance, so can 'Smith's act of will-writing harmed his son'.

How, then, can we apply this approach to group harms? Just as there are facts of the form 'act A results in person P being worse off than he would have been if alternative act B had been performed', there are facts of the form 'the combination of X's doing A and Y's doing B results in person P being worse off than if X had done C and Y had done D'. In some contexts it will be appropriate to take such a fact as grounding a claim of the form 'the group consisting of A and B harms P'. It may well be that Case One provides such a context. Since I end up dead, and death is considered a paradigm case of a harm, we need to assign the harm to something. X's shooting me is naturally contrasted with his not shooting me (and doing pretty much any other non-lethal thing instead), as is Y's shooting me. Since it's not a fact that if X hadn't shot me, I wouldn't have died, nor is it a fact that if Y hadn't shot me, I wouldn't have died, we need to find another fact in the vicinity that hooks up with my being dead rather than alive. Such a fact is that if neither X nor Y had shot me, I wouldn't have died. There may be many other facts that also hook up with my death. For example, if my neighbor had been more persistent in getting me to fix her leaking faucet, I would have been delayed and would not have been in a position to be shot by X and Y. This fact is unlikely to be salient. Neither is the fact that, if X and Y had not shot me and Fred Astaire had done something different too, I would not have died. The salient fact is that, if nether X nor Y had shot me, I wouldn't have died. There is, however, no fundamental moral fact of the form 'the group consisting of X and Y *really* harmed me'.

The contextualist approach can also deal with those cases in which, as Parfit says, there is not *one* smallest group who together bring about the harm. Consider a harming version of the Third Rescue Mission:

Overkill. If four people stand on a platform, this will kill a hundred miners. Five people stand on this platform.

There is not one smallest group of whom it is true that, if they had all acted differently, the other people would not have been harmed. Nevertheless, it seems natural to claim that the five people together harm the 100 miners. The salient fact that would make such a claim appropriate is of the form 'if any two of that group had acted differently (done almost anything except stood on the platform), the miners would not have been killed'.

At this point I should clarify the role of salience[13] in my contextualist account of harm. Salience often plays a role in determining which alternative the context selects as the appropriate one, but salience may not be the only determining factor. To see this, consider an example that might be thought to pose a problem for my account, if salience is solely responsible for selecting the appropriate alternative.[14] Imagine a group of comic-book enthusiasts talking about how great it would be if their leader, Ben, had the abilities of Spiderman. After an hour or three of satisfying fantasizing, they are joined by Ben himself, who apologizes for being late. He explains that he was on his way when his grandmother called him on his cellphone. She had fallen, and she couldn't get up without his help. It took him more than an hour to get to her, because of traffic congestion, during which time she had been lying uncomfortably on the floor. Once he helped her up, though, she was fine. He is sorry that he is late, but the rest of the group, who are also devoted grandsons, must agree that benefiting his grandmother is a good excuse. ''Au contraire'', reply his riends, that is the ''worst excuse ever''. He didn't benefit his grandmother at all, but rather harmed her, since he would have reached her a lot sooner, and prevented much suffering, if he had simply used his super spider powers to swing from building to building, instead of inching his way in traffic. Furthermore, he would have reached the meeting on time. Clearly, something is amiss here. Even though the alternative in which Ben swings through the air on spidery filaments is, in some sense, salient, it is not thereby the appropriate alternative with which to compare his actual behavior. We can't make an alternative appropriate simply by talking about it, although we may be able to make it salient that way. Perhaps we should add to salience, among other things, a commitment to something like 'ought implies can'. Since Ben cannot swing through the air on spidery filaments, this is ruled out as an appropriate alternative.[15] I don't here have the time (or the inclination) to give a detailed account of how conversational context determines the appropriate alternative. I suspect that the correct account will be similar to the approach of contextualists in epistemology, such as David Lewis, Mark Heller, and Keith deRose.

Even when it is contextually appropriate to take a fact about what would have happened if the members of a particular group had acted in certain different ways as grounding a claim about group harms, there may be no simple recipe for assigning credit or blame to the members of the group. In the version of Case One-and-a-Half in which both X's and Y's actions depend on their knowledge that the other will shoot me, it is probably inappropriate to blame either one, although it may be appropriate to assign the harm to the group consisting of the two of them. In Case One, it is probably appropriate to blame both X and Y, and to assign the harm to the group consisting of the two of them.

So, where does this discussion leave the place of harm in consequentialist theory? I am not proposing that we do away with all talk of harm or benefit in our ordinary moral discourse. What I am claiming is that, for the purposes of ethical theorizing, harm and benefit do not have the kind of metaphysical grounding required to play fundamental roles in ethical theory, nor do judgments of harm or benefit make any distinctive contributions to reasons for action. If, in considering whether to do A, I correctly judge that A would harm P, I am judging that A would result in P being worse off than s/he would have been if I had performed the appropriate alternative action. This is certainly a relevant consideration from a consequentialist perspective, but it is also one that I would already have taken into account, if I had considered all my available alternatives.

Although my argument has been conducted in a consequentialist framework, I suspect that plausible accounts of harm in non-consequentialist theories will be subject to the same considerations. I do not here have the time to consider whether this presents any serious problems for non-consequentialist theories.[16]

NOTES

[1] I owe the example to Frances Howard-Snyder.

[2] An asymmetry between reasons not to harm and reasons to benefit is also implausible from a non-consequentialist perspective that treats harms in certain respects as more basic than overall harms. Consider an approach to harm that classifies causing pain (among other things) as harming, and preventing pain as benefiting, and that claims that our reasons not to harm are stronger than our reasons to benefit. Such an approach may well, depending on the size of the supposed asymmetry between reasons, judge that Doctor has stronger reason not to administer the pain-relieving drugs than to administer them. This should clearly be unacceptable to non-consequentialists as well as consequentialists.

[3] Derek Parfit, *Reasons and Persons*, 70.

[4] Parfit doesn't use the term 'preemption'.

[5] This was the thesis of Ben Eggleston's paper, "Does Participation Matter? An Inconsistency in Parfit's Moral Mathematics", presented at the APA Eastern Division, December 1999, to which I delivered the response. The current paper has grown out of my reply to Eggleston's paper.

[6] It might be objected that my elaboration of the example prevents Y's true belief in the actual world that X has poisoned me from being knowledge. If the closest world in which X doesn't poison me is one in which Y believes that X has poisoned me, it seems that Y's actual belief doesn't track the truth in the right way to be knowledge. This objection relies on a controversial theory of knowledge. It's not even clear that it succeeds in the context of that theory. Given that the contexts in which we consider whether Y has knowledge and in which we consider whether Y is a member of the group that

harms me are different, different possible worlds may be relevant to each. Furthermore, if we simply changed Case Three to specify either that Y simply believes that I am about to die painfully, or that Y knows that it is almost certain that I am about to die painfully, our intuitive judgements of Y's behavior would remain unchanged.

[7] Frank Jackson makes a similar point in his "Which Effects?", in *Reading Parfit*, ed. Jonathan Dancy, Blackwell 1997.

[8] See my "Comparing Harms: Headaches and Human Lives", *Philosophy and Public Affairs*, Spring 1997, for arguments against this view.

[9] This approach can be subject to many variations. For example, do we compare (C1) and (C2) with respect to a particular person, a particular type of person, the "average" person, etc.? Do we compare propensities with respect to the circumstances a particular individual is likely to encounter, given what we know about her, given her social position, given "normal" circumstances, etc.?

[10] For discussion of this point see my "Consequentialism and Commitment", *Pacific Philosophical Quarterly*, December 1997.

[11] See, for example, the accounts of counterfactuals developed by David Lewis and Robert Stalnaker.

[12] The problem here is both that the proposal makes character relevant to whether actions harm or benefit, and that it does so in a particularly counterintuitive way. For a consequentialist, the former problem is more significant.

[13] I mean by salience, roughly, the degree to which the participants in a conversational context consciously focus on an alternative. There may be more sophisticated accounts of salience, but this is certainly a common one.

[14] I owe at least the general idea of this example, though not the details, to Ben Bradley. He suggested something like this in discussion as a problem for my account.

[15] I owe thes suggestion to Julia Driver.

[16] I am grateful for comments on various versions of this paper by Jonathan Bennett, Ben Bradley, Julia Driver, Doug Ehring, Mark Heller, Frances Howard-Snyder, Elinor Mason, Stuart Rachels, George Sher, Steve Sverdlik, and an anonymous referee for this journal.

Department of Philosophy
Rice University
Houston, TX 77005
E-mail: norcross@rice.edu