FRANK ARNTZENIUS and DAVID McCARTHY

SELF TORTURE AND GROUP BENEFICENCE*

ABSTRACT. Moral puzzles about actions which bring about very small or what are said to be imperceptible harms or benefits for each of a large number of people are well known. Less well known is an argument by Warren Quinn that standard theories of rationality can lead an agent to end up torturing himself or herself in a completely foreseeable way, and that this shows that standard theories of rationality need to be revised. We show where Quinn's argument goes wrong, and apply this to the moral puzzles.

A few years ago Warren Quinn described a situation in which it would seem that people by their own rational decisions are forced to torture themselves in an obviously irrational way.[1] The puzzle he posed was whether there was a rational way of avoiding this self-torture. We offer a solution to this puzzle. We then apply what we learn from this solution to a much more widely discussed moral problem involving acts which are said to produce imperceptible benefits or harms.

1.

Suppose there is a device which can be used to apply electric current to a person's thumb. This causes a pain in the thumb the intensity of which depends on the current applied. The device has 1000 settings, from 0 (off) to 1000 (high but non-lethal current). Some sadistic psychologists make use of this device on Harry, a reputedly highly rational individual. They implant the device in his thumb and tell him how it works. They ask him whether he can feel anything when the device is at setting 0. He replies that he can't. They then turn it to level 1000 and ask him what it feels like. "Extremely painful," says he. Would any amount of money compensate him for suffering that level of pain for the rest of his life? "Certainly not." Not even a million dollars? "No, not even ten million." They turn the device to various other settings to give Harry an idea of what the pain is like at those settings. They then turn the device to various adjacent settings, like 232 and 233, or 765 and 766, and ask him whether he can tell the difference in how the adjacent settings feel. He thinks carefully, and admits that he can't. He eventually becomes convinced that he can't ever

Erkenntnis **47:** 129–144, 1997.

tell the difference between adjacent settings. This is what the psychologists want; if he had been able to sometimes tell the difference, they would have made the increases more fine grained, as they would be with 10 000 steps between no current and the maximum. Surely, they think, there must be some sufficiently fine grained series of increases such that Harry could not tell the difference between adjacent settings. So it does not matter that our story assumes that this is so in the set up we have described.

The psychologists then turn the setting back to 0 and offer Harry the following deal. Each morning he can increase whatever setting the device is at by 1, and if he does, he will receive $1000. But once he increases the setting, he can never turn it back. For example, if he increases the device from 45 to 46, he will feel at least that much current, at least that amount of pain, for the rest of his life. Call this *Self Torture 1*. (This is slightly different from Quinn's original case, which we will discuss later.)

The problem is this. Each morning Harry reasons as follows. "I am reliably informed that the current will have no adverse effects on me apart from the pain, and the only thing that matters to me about pain is the way it feels. I can't tell the difference between the present setting and the next one up, so, money aside, I am indifferent between the two settings. But I'd like to have an extra $1000, so I should up the setting by one." Since Harry reasons like this each morning, in one thousand days he drives the setting up to the maximum and thereby earn a million dollars. But at the start he told the psychologists that not even ten million dollars would compensate him for that pain. So he started off in a painless state. He then made a sequence of apparently rational decisions and ended up in a clearly less desirable state than the one he started in. Being a highly rational individual, it was easy for him to foresee this from the start. Something's gone wrong. No good theory of rationality can have this result. The puzzle Quinn poses is to say where Harry went wrong.

Quinn thinks it obvious that Harry should not keep going to the maximum, but he thinks there was nothing wrong with Harry's series of decisions according to standard theories of rationality. So, Quinn says, those theories must be revised.

We say that it does not follow from the description of the case that it is rational for Harry to increase the setting by one at any level. We say that there is some setting below 1000 at which it's rational for Harry to stop. And we say this follows from standard theories of rationality. Far from having to be revised, those theories help show us what's wrong with Harry's sequence of decisions.

By 'standard theories of rationality' we mean variations on the idea that an agent has certain subjective preferences which determine a subjective

value function over states of affairs such that at any time it is rational for the agent to perform the act which of all acts available to him or her has greatest *expected* value as determined by his or her subjective probabilities. Most of these theories provide a formal, mathematical framework for describing and assessing questions of rational choice, but for most of this paper we will try to give an intuitive, informal explanation of where Harry's reasoning goes wrong, sketching at the end how this can easily be explained in these more formal models.

2.

One indication that there is something wrong with Harry's reasoning is this: he can be turned into a money pump in a way he could easily foresee. For suppose that on day 1001 the psychologists offer to reset the setting to 0 for a million and one dollars. Given his preferences he should accept this deal. So he would end up with one dollar less than he started with (and would have suffered a lot of pain along the way). Even if he knew in advance that this would be offered to him on day 1001, he would still have acted as he did on the previous 1000 days, if we assume the same reasoning and preferences as before. This makes it even clearer that there is a real problem about his sequence of apparently rational decisions. It suggests that the problem is that his preferences are intransitive. Given the extra thousand dollars he gets for each increase in the setting, and assuming it costs him a million and one dollars to go from 1000 to 0, he sees himself as getting strictly better off at each step as he goes from 0 to 1, 1 to 2, . . . , 999 to 1000, and 1000 to 0. Intransitivity of preferences is something most standard theories of rationality cannot cope with. But it's not enough just to rule them out, as they arise in such a reasonable way. The only thing for Harry that is bad about pain is the way it feels, and since he cannot seem to tell the difference between the way adjacent settings feel, and since he can tell the difference between having and not having an extra thousand dollars, each of his stepwise preferences seems entirely reasonable. If the fault lies with some of his preferences, we need to explain *why* they are unreasonable.

It will help if we describe our strategy and conclusions. We will argue that Harry should take the apparent indiscernibility of the pain levels at adjacent settings to indicate that it's hard to tell by direct introspection whether the pain levels are different, but not that they're the same. We will argue that Harry has evidence that at least some, and plausibly all, of the pain levels at adjacent settings are different, and that he should care about such differences. This is all that is needed to block the reasoning that led

Harry to go to the maximum. It also explains where standard theories of rationality tell Harry he should stop, and why that's before the maximum.

Let's consider first a slightly different case, *Self Torture 2*. Before Harry is allowed to start increasing the settings from 0, the psychologists do the following experiment. They administer a very long sequence of electrical shocks on him with the device. They randomize the setting in between each shock, but they don't tell Harry what the setting is. They tell Harry to describe how painful each shock feels to him, in whatever vocabulary he prefers to use. He may use numerical and precise vocabulary, he may use qualitative vocabulary, he may use vague predicates, it does not matter, as long as he tries to describe the level of pain he feels on each occasion as accurately as he can. If a shock at some time appears exactly as painful to him as some other shock at some other time he must describe that pain using the same words, and if a shock appears either more painful or less painful than some other one, he must describe them differently. After doing a whole sequence of experiments the psychologists tabulate the results. For each level of the device, they group together all the reports Harry made of the intensity of pain at that level. They then show Harry these results. Only then is Harry allowed to start increasing the setting by one each morning.

There seems very little difference between Self Torture 2 and Self Torture 1. And it may seem to Harry that the reasoning in Self Torture 1 is just as sound in this new case, with the result that it seems rational for him to gradually increase the setting to 1000. In part we agree: the reasoning in Self Torture 2 is just as sound as in Self Torture 1. The evidence Harry has in Self Torture 2 should make it clear to Harry why that reasoning is unsound in Self Torture 2, and that should make it clear to him why the reasoning is unsound in Self Torture 1. (More precisely: at least one *instance* of Harry's reasoning each morning before he reaches 1000 is unsound in each of Self Torture 1 and 2.)

We shall establish this by considering the two possibilities for the groupings of reports. The first is that whenever the settings are the same, Harry gives the same description of the pain he feels. The second is that Harry does not always describe his pain in the same way when the settings are the same.

Let us begin with the first possibility. Although it is somewhat unlikely, it would immediately show that Harry's reasoning was unsound in this case. For it would immediately follow that some, and possibly all, adjacent settings of the device in fact are discernible by Harry in terms of the pain he feels at those settings. For if his descriptions are always the same if the settings are the same, one can attach to each setting a unique description. Since setting 0 is not at all painful to Harry, and setting 1000 is very painful,

the descriptions must change, and the pain felt be described as worse, at least once as the setting goes from 0 to 1,1 to 2, . . . , 999 to 1000. Thus for some N, setting $N + 1$ is noticeably more painful for Harry than setting N. One of the premises in Harry's reasoning each morning was: "I can't tell the difference between the present setting and the next one up." So on at least one morning, that premise was false, so Harry's argument that he should up the setting by one on that morning was unsound.

Some people will think that even if Harry can notice the increase in pain between some or all adjacent settings, it will still be rational for Harry to go to the maximum setting: at each stage the one step increase in pain is at best hardly noticeable whereas the extra thousand dollars is a clearly noticeable benefit, swamping the slight increase in pain and creating a similar version of the argument that on standard conceptions of rationality it is rational for Harry to go to the maximum setting. We do not believe that is correct, and we will address it later. But for now, notice that regardless of the merits of such an argument, it is not a form of the one Quinn gave to show that standard theories of rationality would lead Harry to 1000. That argument is based on the premise that Harry can't tell the difference between the way adjacent settings feel. If the first possibility obtains, that premise is false, at least in Self Torture 2.

We now turn to the second possibility, the more complicated case where Harry does not always give the same descriptions to the same settings. In this case we cannot simply classify settings according to the descriptions of pain that Harry gives at those settings, for he does not always give the same description at the same setting. But we can classify them according to the (long run) frequencies of the reports he gives. For instance, setting 235 might be such that 83% of the time he describes the accompanying pain as very painful, and 17% of the time he describes it as somewhat painful. And since the frequencies of reports must differ for the settings 0 and 1000, it immediately follows that at least some adjacent settings must give rise to a different frequency distribution of reports. In fact, given that currents between any adjacent settings always differ by the same small amount it would seem plausible that all settings would give rise to differing frequencies of reports, although the frequencies will differ very little if the settings are adjacent. Thus at least some, and most plausibly all, adjacent settings, are discernibly different on the basis of direct reports of pain felt, in the sense that the frequencies of reports differ for those adjacent settings. We will now argue that Harry should care about such differences. To see this, consider two interpretations of such differences in frequencies.

The first interpretation is that Harry's reports are always accurate and veridical descriptions of the pain he feels. If so, the pain that Harry feels

is not always the same at the same setting. Moreover, at least some of the time, and plausibly very often, Harry feels exactly the same pain at adjacent settings. This would also explain the sense in which adjacent settings are indiscernible: most of the time they give rise to exactly the same pain. But adjacent settings are discernible in that they give rise to different frequencies of such pains. And Harry should care about such differences. For example, if he describes setting 235 as very painful 83% of the time and somewhat painful 17% of the time, and setting 236 as very painful 84% of the time and somewhat painful 16% of the time, then other things being equal, that is, the extra money aside, he should regard setting 236 as worse than setting 235: in the long run, he will feel slightly more pain at setting 236. More precisely, the expected (disvalue of the) pain he will feel is greater at setting 236 than 235, where the expected pain is given by the sum of an appropriate measure of each possible pain multiplied by its probability. This is just what standard theories of rationality tell Harry he should care about.

The second interpretation of the difference in frequencies of Harry's reports for the same setting is that they are not always accurate and veridical descriptions of the pain he feels. Since he is trying to report the pain he feels at different settings as accurately as he can, the obvious explanation of why his reports may be inaccurate is that he makes various kinds of errors in memory or judgment. But what would follow if such were the case?

Suppose Harry experienced a certain setting last week and now reports that the setting he is experiencing now is slightly more painful. And suppose that psychologists point out to him that they have evidence that people tend to underestimate the painfulness of past experiences in proportion to how far back in the past they were. So Harry now comes to believe that his report is based in part on an error, and that the setting he is experiencing now is actually less painful than the one he experienced last week. Suppose he is told that he must experience one of the two settings for the rest of his life; which should he choose? It is clearly rational for him to choose the setting which he has most evidence to believe is less painful, and not the setting which he mistakenly reported was less painful. So in the present case, he should try to infer the pain he *actually* feels at each setting from his *reports* of the pain he feels at that setting, taking into account his estimate of the ways his reports might be based in part on errors of memory or judgment. Quite plausibly Harry could infer that the long run frequency of the pain reports that he gave at a given setting reliably indicates the (expected) actual pain caused by that setting. But even if Harry thinks that the relation between pain reports and actual pain is more complicated,

this will not make all of the arguments he gave himself each morning valid unless he assumes that his error is so great that the (actual or expected) pain associated with each setting is exactly the same as the settings adjacent to it. But that would mean that the pain associated with 0 is the same as the pain associated with 1000, and he has overwhelming evidence to believe that is not true. Thus even if some or all of the variation in the reports he makes of the pain at each setting is based in part on errors, he has no reason to believe that the expected pain associated with each setting is the same. Hence by the arguments of the preceding paragraphs at least some instances Harry makes each morning of the argument that led him to 1000 are unsound, so the argument does not show that it is rational for him to up the setting on those occasions.

Let us summarize our argument. Since Harry can clearly distinguish settings that are reasonably far apart, it follows that the frequencies of descriptions of the pain he feels at a given setting must differ for at least some adjacent settings, and quite plausibly for all adjacent settings. This shows that for at least some, probably all, adjacent settings either the pain is always different, though in such a way that he usually fails to report any difference, or that the pain for two adjacent settings is usually the same, but the frequencies with which he feels certain levels of pain differs slightly even for adjacent settings, or a mixture of these two cases. In either case, for at least some settings, increasing the setting by one makes Harry worse off, so, money aside, he has a reason to prefer not to increase the setting by one. Hence in Self Torture 2, Harry was mistaken to think that, money aside, he was indifferent between adjacent settings.

One might object that although the experiment we have described with 1000 different settings would reveal to Harry that at least some of the adjacent settings differed in terms of what matters to him, it would not reveal such a difference if it took 10 000, or 100 000 steps to get from no current to the maximum. Surely for sufficiently small increases, a one step increase makes no difference to what matters to Harry. But in response, the fact that it took 1000 steps in the variation we considered played no role in our argument. The argument goes through if 1000 is replaced by 10 000 or 100 000, or what have you.

So much for Self Torture 2; what about Self Torture 1? In that variation, the psychologists did not do the experiment we have been describing on Harry. They merely implanted the device in Harry's thumb with a full explanation of the set up and evidence which suggested to Harry that it was at least very difficult for him to tell the difference between each member of pairs of adjacent settings. But Harry could easily *imagine* the experiment we have described, just as we have been doing. So even in

the absence of the experimental evidence we have described, Harry has perfectly good reason to believe that money aside, at least some, quite plausibly all, of the one step increases are worse for him in terms of what he cares about. Hence at least some in the sequence of arguments that led Harry up to 1000 are unsound, and, contrary to Quinn, no argument has been given that standard theories of rationality lead Harry in a completely foreseeable way to end up torturing himself.

3.

It may be objected that at least in Self Torture 2 we have not said *which* of the instances of the argument Harry rehearsed each morning is unsound. It may also be objected that in our discussion of Self Torture 1 and 2 nowhere have we said whether it is rational for Harry to go beyond setting 0, and if so, where it's rational for him to stop. But this is a good thing, since which instances are unsound and what it's rational for Harry to do depends on further facts we haven't specified.

First, all our rebuttal of Quinn's original argument against standard theories of rationality relied upon was the claim that there are at least two adjacent settings such that the higher setting is, money aside, worse for Harry in terms of what he cares about – the expected pain – than the lower setting. But that claim is clearly consistent with a vast range of hypotheses about how the expected pain varies with the setting. For example, there may or may not be a threshold level of current below which there is no expected pain. And the increase in expected pain may be fairly linear with the increase in current, or alternatively, the increase might be relatively steep over some settings, and relatively flat over others. (It might even decrease over some settings.) Which instances of the argument Harry rehearsed each morning are unsound depends on his estimate of the relative likelihoods of these different hypotheses. And which setting it is rational for Harry to stop at depends at least in part on his estimate of those likelihoods. But it also depends on the next two factors.

Second, different individuals can reasonably attach different values to the same pains. A boxer might care a lot less about a given pain than most of the rest of us. Where it is rational for Harry to stop therefore depends at least in part on his personal view of pain.

Third, different individuals can rationally attach different values to the same amount of money. Harry might care greatly about getting $10 000 for the hip replacement his aunt needs, but be otherwise quite content with having little money. Or he might care greatly about having $100 000 so he can put his daughter through law school. Where it is rational for Harry to

stop therefore depends at least in part on the value he attaches to having different amounts of money.

In short: where it is rational for Harry to stop depends on the probabilities he attaches to different hypotheses about the relation between expected pain and setting level, on his subjective valuation of different levels of expected pain, and his subjective valuation of different levels of wealth. Until these are specified, we cannot say whether it is rational for Harry to go beyond 0 at all, and if it is, where he should stop. But given all these facts about Harry, standard theories of rationality determine exactly where he should stop, and our discussion shows why that's before the maximum.

In Self Torture 2 Harry has more evidence than in Self Torture 1, so it is likely that the probabilities he attaches to different hypotheses about the relationship between the different settings and the pain he will feel will differ, so the point at which it is rational for him to stop may not be the same in each case. We have delayed discussing Quinn's original case because it is more complicated than the variations we have been discussing because Harry has more opportunity to gather evidence. This can make a difference to what it is rational for him to do. Quinn's original case is Self Torture 3.

Here Harry has the option at the beginning of experimenting with different settings to find out what those settings are like, and at any stage he can try out settings higher than the one he is currently on before deciding whether to up the setting by one. Thus the real question for Harry is first, what evidence is it rational for him to gather, and second, where it is rational for him to stop given the results of his evidence collection. The argument that it is not rational for Harry to go up to the maximum is just the same in this case, so we won't repeat it. Instead, we describe how standard theories of rationality tell Harry to assess what evidence he should gather.

Like us, at the start of the affair Harry only has prior estimates of the likelihood of various hypotheses about what the expected pain is like at various settings. If he had the results of extensive experimentation, like the experiment we imagined in Self Torture 2, he would have a much more accurate estimate of the likelihood of those hypotheses, and a decision based on that has no lower, and almost certainly higher, expected value than the decision about where to stop it would be rational for him to make in the absence of having such evidence. It's possible to prove a result like this using standard theories of rationality: having more evidence always has non-negative, and usually positive expected value.[2] So it is rational for Harry to get as much evidence as possible first, if the expected costs of getting such evidence were zero. But the costs of obtaining evidence

are not zero: it takes time, and it hurts! But given the high stakes (lots of money and lots of potential pain), we believe it would be reasonable for Harry to obtain a fair bit of evidence first. But exactly what evidence it's rational for Harry to obtain depends on his prior estimates of the value of obtaining that evidence, which will depend, roughly, on his prior estimates of the value of learning about particular kinds of evidence and his prior estimates of how likely he is to get those kinds of evidence. The picture is even more complicated by the fact that if at any point Harry increases the setting by one, that will itself give him new evidence to factor into the decision of whether to go to the next step, whether he should obtain still more evidence first, and so on. (That was, of course, true of Self Torture 1 and 2.) Of course all this is horribly complicated, but the goal of standard theories of rationality is to provide a *criterion* for which act it is rational to perform based on the agent's subjective preferences and probabilities. It does not try to provide a simple method for an agent whose computational abilities are limited to *determine* which act is rational to perform. This may show that there are important and interesting questions about rationality which standard theories of rationality do not address. We have no quarrel with that. But the argument Quinn gave was an attempt to show that standard theories of rationality fail to do what they try to do. If we are right, Quinn has given no good argument that they lead Harry to self-torture, and no good argument that they fail to do what they try to do.

4.

Even if we have refuted one argument that said that standard theories of rationality tell Harry to go to the maximum, it may appear that a similar argument can be run. When Harry realizes what was wrong with his reasoning, the psychologists kindly set the device back to zero and allow him to start again. The problem is that Harry now reasons like this each morning. "Even if all adjacent settings are different in terms of what I now realize I should care about, the expected pain, the increase in expected pain is at most only very slight. But having an extra $1000 is a great benefit. So on balance, the good of the extra $1000 outweighs the slight increase in expected pain. So I should increase the setting by one." Again, this leads poor Harry to end up torturing himself. Where's his mistake this time?

Just as before, if Harry accepted this line of reasoning, he would have intransitive preferences and could be turned into a money pump. But again, it still has to be shown what's wrong with them since they arise in an apparently reasonable way.

One fallacy Harry might be committing is this. He might say: "Having an extra $1000 is clearly discernible, but while the pain associated with a one step increase is in some statistical sense discernible after lengthy scrutiny, it is still very hard to discern. Since clearly discernible differences must be greater than hard to discern differences the value of having the extra $1000 always outweighs the disvalue of a one step increase."

But while having an extra $1000 is clearly discernible – as a big pile of dollar bills, say – it's the *value* of the money to Harry that matters to him, and that might be much less easily discernible. Suppose, for example, that the only thing Harry can buy with $1000 is a painkiller which will fractionally reduce the pain he suffers over the rest of his life, and suppose that the reduction in pain due to this painkiller is hard to discern, indeed is only discernible in terms of long run frequencies of pain reports. That would make it very hard for Harry to discern the value of the $1000, and it would not be obvious at all that it outweighed the disvalue of the increase in expected pain.

Let us now see what further mistakes Harry might make when he tries to balance the value of an extra $1000 against the disvalue of the increase in expected pain.

One thing Harry might miss is the diminishing marginal value of money. Unless he has very unusual preferences, or is in very unusual circumstances, which we shall ignore, then at some point, the next $1000 will be worth less to him than the previous. Crudely, $1000 means more to a poor person than a rich person. Thus the value of the extra $1000 Harry would get for going from setting 600 to 601 might be very little, small enough to be outweighed by the increase in expected pain, even if, as Harry thought, that increase is "slight."

There is one further point Harry might overlook. When he compares the pain he feels at adjacent settings, it is hard for him to tell the difference. And even if the position we have been arguing for is correct, the difference understood in terms of statistical discernibility is only slight, at least when understood as a sort of instantaneous comparison. But does that mean that the increase in expected pain is, as Harry thought, only slight? No. Harry is going to suffer the pain caused by whatever setting he goes to for every minute of every hour of every day for the rest of his life. Remember Bentham's admittedly crude Hedonic Calculus: pleasure equals intensity times duration. When we factor in the duration of the "slight" increase, the net result may be that the expected increase in pain is actually quite significant, and not slight at all; something small times something large can be significant. Moreover, from the point of view of pleasure spread out over an entire life, the value of an extra $1000 begins to look fairly small:

if Harry lives for another fifty years, it works out at about five cents a day. And it's not hard to believe that a little extra pain for a whole day is not worth five cents.

In summary, once we notice the distinction between the discernibility of a given amount of money and the discernibility of the value of that amount of money, the diminishing marginal value of money, and the duration of the pain, there is no difficulty in seeing that at some setting the value to Harry of getting an extra $1000 will be outweighed by the increase in expected pain, so there is no argument that standard theories of rationality tell Harry to end up torturing himself. In fact, Harry's problems nicely illustrate much of what is attractive about standard theories of rationality: the notion of expected value, their ability to distinguish between money and value, to explain puzzles about the diminishing marginal utility of money, and to provide a framework to assess questions about rational evidence gathering.

5.

Suppose that there are a thousand intensely thirsty wounded soldiers in the desert. A water-cart is being taken through Sally's town to the soldiers. When it gets there, the water will be divided evenly among the thousand. There were one hundred pints of water in the cart to start with, and when it gets to Sally, she can add her pint at little cost to herself. Here's how Sally reasons.

"On any plausible moral theory the consequences of an action are relevant to its permissibility. On some theories, things other than its consequences can be relevant to its permissibility. But here, only the consequences are relevant. As it happens, I know the amount of water the soldiers receive will have no adverse effects on their health, so the only thing that is bad about their suffering is the way it feels. If I add my pint, I will only give each soldier an extra thousandth of a pint, an amount which would result in no perceptible difference to any soldier, so no difference in the way any soldier feels. So I have no moral reason to add my pint, so no plausible moral theory can say that I ought to add my pint."

Since it is a noticeable inconvenience for Sally to add her pint – crossing the street in the hot sun – she doesn't. As it happens, Sally was one of nine hundred potential pint of water donors, each reasoning in the same way. Each does not add his or her pint, and the net result is that the soldiers only get a tenth of a pint. Had each person added his or her pint, each of the soldiers would have ended up with a full pint, a marked improvement given how intense their thirst is. So at little cost to himself or herself, each of the nine hundred could have helped to greatly relieve the soldiers' suffering.

Something seems to have gone wrong with the reasoning that led Sally and the others to conclude that it wasn't true that they ought to add their pints.

This problem has been discussed under different guises by many writers.[3] Two kinds of solutions have been offered. For the sake of argument, we shall accept that although it is not clear whether, as consequentialism says, the rightness of actions is always a function of their consequences, it seems plausible that in Sally's case only the consequences are relevant.[4] Then the Group Beneficence Solution says: "By itself, Sally's adding her pint would produce no good consequences. But Sally ought to add her pint because that act is a member of a *set* of acts such that if each member of that set were performed, the consequences would be best."[5]

The Group Beneficence Solution is subject to a number of severe difficulties brought out by Derek Parfit and Michael Otsuka which we need not repeat,[6] except to say that they all stem from the claim that, by itself, Sally's adding her pint would produce no good consequences.[7] Furthermore, no other solution looks possible if we accept that, by itself, Sally's adding her pint would produce no good consequences. Since there is surely some solution to this puzzle, these writers recommend that we should accept that, by itself, Sally's adding her pint produces better consequences than her not adding her pint. Hence Sally's problem is solved by the Restricted Consequentialist Solution: "Sally ought to add her pint because, by itself, the consequences of her adding it are better than the consequences of her not adding it."[8] And, along with Frances Kamm,[9] these writers add that, since the benefit to each soldier of getting an extra 1000th of a pint is imperceptible, we have to accept that there can be imperceptible benefits. This claim sounds counterintuitive, but its denial is worse; it leaves us with no solution to Sally's problem.

Where does this leave us? We have to believe in imperceptible benefits because that is necessary to adequately solve a moral problem. But there is something very odd about there being imperceptible benefits. If the only thing that is bad about the soldiers' suffering is the way it feels, and they could perceive no difference in having an extra 1000th of a pint, how can having the extra 1000th be a benefit? To be comfortable with the Restricted Consequentialist Solution, or something similar, we need to be told much more about how Sally's adding her pint is a benefit to the soldiers.

The discussion of Harry's problem establishes that for at least some n between 0 and 999, (and quite plausibly for all n) some (and plausibly all) of the soldiers are better off for having $n + 1$ 1000ths of a pint of water rather than n 1000ths of a pint. This leads to two questions. Are these benefits imperceptible? What does this show about what Sally ought to do?

We find the claim that the benefits are imperceptible somewhat misleading, and this may be responsible for the air of paradox this kind of case has continued to have. Harry, for at least some, and plausibly all, of the settings could perceive the difference in the pain that they caused, in the sense that his long run frequency of pain reports differs for different settings. This, despite the fact that most of the time he would give the same pain report for adjacent settings. Now one can argue whether a difference in long run frequencies of pain reports should be called a perceptible difference. But, as we have argued, it is clear that it is a difference that Harry should care about. Now consider Sally's case. Since large additions of water always make a clearly perceptible difference to each soldier, it follows that, at least at some water levels, and most plausibly at all water levels, small additions of water make a perceptible difference, in the sense of the relative frequencies of perceptual reports of the soldiers. And as in Harry's case, the soldiers should care about these differences. So there is a sense in which such differences are perceptible, and the soldiers, and Sally, should care about perceptible differences in that sense of "perceptible". This is true even if there is also a sense in which these differences are imperceptible, there being more than one reasonable way of making precise what in such a case counts as a perceptual ability.

Ought Sally to add her pint? This question is complicated by the fact that it is a noticeable inconvenience for Sally to add her pint – crossing the road in the hot son. In fact, whether Sally ought to add her pint is underdetermined by the description of the case in just the same way in which the description of Harry's case underdetermines the number of times, if at all, he ought to increase the setting. For example, if for each soldier the marginal value of an extra 1000th of a pint rapidly diminishes with each extra 1000th of a pint, the consequences may be on balance best if, say, exactly 700 people add their pints. If the marginal value is constant, the consequences would be best if 1000 people add their pints. But there is one further complication for Sally and her fellow potential donors which Harry does not face: if the consequences are best if some number less than 1000 people add their pints, Sally and company face what is, at least in theory, a horrendous coordination problem. What each ought to do may in principle be solved only by a far more complicated theory than the simple (restricted) consequentialism which underlies the Restricted Consequentialist Solution. These issues have been very well discussed by others.[10] However, our main point stands. The argument that Sally ought not to add her pint since it would confer only an imperceptible benefit to the soldiers is not a valid argument. In the relevant sense of "perceptible", it is a perceptible benefit.

NOTES

* Derek Parfit gave very detailed and generous comments, in part correcting some mistakes we had made in interpretation of some of his views, and an anonymous reviewer for *Erkenntnis* gave us a very useful guide to some literature of which we were not aware. Judith Thomson was also very generous with her comments. This article was written while the second author was receiving funding from the Greenwall Foundation, which is gratefully acknowledged.

[1] Quinn (1990).

[2] See Skyrms (1990).

[3] Recent discussion of this problem derives from Parfit (1984), who in turn attributes the example to Glover (1975). A very similar problem is discussed in Regan (1980). This variation of the problem seems to come from Harrison (1952–53), who in turn bases it on Harrod (1936). Uttara Bharath tells us that a variation of Sally's problem has been part of a South Indian folk-tale for hundreds of years.

[4] We are sympathetic to the claim of Foot (1985) and Thomson (1992) that there is an important mistake in consequentialism's freewheeling talk of better and worse consequences, but in this context it would be distracting to avoid such talk. Our substantive claims can easily be restated without this locution.

[5] Variations on this solution are endorsed in Harrison (1952–53) and Harrod (1936). It is offered as a possible solution in Parfit (1984), although it is not his preferred solution.

[6] See Parfit (1984); Parfit (1987); and Otsuka (1991).

[7] So these criticisms of the Group Beneficence Solution are not, by themselves, criticisms of attempts to solve general coordination problems by appeals to sets of acts, such as forms of rule consequentialism and the coordination utilitarianism defended in Regan (1980).

[8] To a first approximation, this is also Regan (1980)'s solution. Regan's full account lies within the context of a theory too complex to describe here, although we will later indicate some of the power of Regan's account.

[9] Kamm (1991).

[10] In particular, see Regan (1980) and the references therein.

REFERENCES

Foot, P.: 1985, 'Utilitarianism and the Virtues', *Mind* 94, reprinted in Samuel Scheffler (ed.), *Consequentialism and Its Critics*, Oxford University Press, Oxford 1988.

Glover, J.: 1975, 'It Makes No Difference Whether or Not I Do It', *Proceedings of the Aristotelian Society*, Suppl. Vol. 49, 71–90.

Harrison, J.: 1952–53, 'Utilitarianism, Universalisation, and Our Duty to Be Just', *Proceedings of the Aristotelian Society* **53**, 105–34.

Harrod, R.: 1936, 'Utilitarianism Revised', *Mind* **45**, 137–156.

Kamm, F.: 1991, *Morality*, *Mortality*, Vol.1, Oxford University Press, Oxford.

Otsuka, M.: 1991, 'The Paradox of Group Beneficence', *Philosophy and Public Affairs* **20**, 132–149.

Parfit, D.: 1984, *Reasons and Persons*, Oxford University Press, Oxford.

Parfit, D.: 1987, *Reasons and Persons*, Oxford University Press, Oxford (a revised edition of Parfit 1984).

Quinn, W.: 1990, 'The Puzzle of the Self Torturer', *Philosophical Studies* **59**, 79–90.

Regan, D.: 1980, *Utilitarianism and Cooperation*, Oxford University Press, Oxford.
Skyrms, B.: 1990, *The Dynamics of Rational Deliberation*, Harvard University Press, Cambridge.
Thomson, J.: 1992, 'On Some Ways in Which a Thing Can Be Good', *Social Philosophy and Policy* **9**, 96–117.

Manuscript submitted March 25, 1996
Final version received October 7, 1996

University of Southern California
Department of Philosophy
3709 Trousdale Parkway
Los Angeles CA 90089-0451
U.S.A.