GRAHAM ODDIE

KILLING AND LETTING-DIE: BARE DIFFERENCES AND CLEAR DIFFERENCES

(Received in revised form 15 April 1996)

1. FROM BARE TO CLEAR DIFFERENCES

Is killing in itself worse than letting-die? In medical practice, in law, and in folk morality the answer is pretty clearly *yes* – and the fact that it is worse is typically held to make a difference to what it is morally permissible to do. Some interesting thought experiments, however, suggest that in itself killing is no worse than letting die. These thought experiments typically involve the method of "bare-difference".[1]

Take one of the most famous of these thought experiments – that of Smith and Jones, each of whom stands to gain enormously through an inheritance from the death of his six year old nephew. Jones undertakes to kill the nephew while the boy is in the bath – by pushing him over, knocking his head on the edge of the bath thereby rendering him unconscious, standing by to hold him under the water if necessary, and then making the whole thing look like an accident. As it happens Jones doesn't have to hold the nephew down – that knock is sufficient to keep him unconscious. In the barely different scenario, Smith also plans to kill the nephew in the very same way. Smith comes into the bathroom with the same intent but finds that the boy has just fallen over accidentally, knocking his head on the edge of the bath, and is now lying unconscious, face down in the water. As in the Jones scenario, Smith stands by ready to keep his head down if the boy looks like emerging, but this turns out to be unnecessary. After a bit of thrashing about the boy drowns. Jones kills his nephew. Smith lets his nephew die.

If you judge these two scenarios to be equally bad then it seems you are thereby committed to denying that killing is in itself worse than letting die. For if killing were in itself worse than letting die

Philosophical Studies **88:** 267–287, 1997.

then the mere shift from letting-die to killing would have to make the situation as a whole worse. So one way to refute a claim of intrinsic worseness and perhaps establish the value-equivalence of two features is to construct a pair of situations such that the only (or bare) difference between them involves the two features at issue, but which have exactly the same value.

Bare difference arguments have come in for a range of important criticisms,[2] but there is one kind of response to such arguments which has not been widely discussed and which is not easily rebutted. When presenting the above thought experiment I often meet with intuitive resistance to the judgement that the scenarios are equally bad. Some responders insist that Jones's killing is worse than Smith's letting-die, and their judgement remains even when I try to locate, and remove, what I suspect are various extrinsic differences they are tacitly smuggling in to make their response seem reasonable. It strikes me that having ascertained that the only difference between Jones and Smith involves killing and letting-die, the responder's judgement is then driven by a prior commitment to the principle that killing is, in itself, worse than letting die. But can one say anything in response to the stubborn resistance of intuition in such cases or does one simply have to stick, equally stubbornly, to a judgement of identity of value in the barely-different pair?

Fortunately the method of bare-difference does not exhaust our repertoire of argumentative strategies. In this paper I develop a kind of argument which does not appeal to bare-difference scenarios and their accompanying judgements of identical value. Instead the argument involves pairs of clearly different scenarios, scenarios which differ in obvious and detectable ways, and which elicit more modest judgements of value. From a class of clear differences we can then establish that killing is not worse, in itself, than letting die.

2. THE ARGUMENT

We will be comparing situations in which people are either killed or allowed to die. In order to distill out the particular contribution, if any, made to differences in value by the difference between killing and letting-die the cases will have to be otherwise as similar as is possible (as in standard bare-difference arguments). So I assume that

the deaths of the people involved are otherwise of the same disvalue, that they involve the same amount of loss, of pain, anxiety, fear and so on. This, of course, makes the thought experiments somewhat artificial, but that in itself is not an objection to them.

The argument draws from the catalogue of trolley stories.[3] A runaway railway trolley, this one a sleeper with one sleeping passenger in it, is careering out of control down a track, as usual, towards a fork. You are at a fork in the track within easy reach of the switch. Down the left-hand fork there is a stationary trolley and in it an unsuspecting passenger. If things are left to themselves the runaway trolley will head down the left-hand fork and collide with the stationary trolley causing the death of both passengers. This right hand fork leads to a siding with a brick wall at the end of it. If you were to activate the switch the runaway trolley would be diverted down the right fork hitting the brick wall and killing the sleeping passenger. So in this case you could either do nothing, thereby letting two people die, or you could send the trolley into the brick wall, thereby causing the death of one (that is, killing him). So, you could either let two people die or you could kill one person.[4]

There is a clear difference between your two options, and it also seems clear that it would be better to kill one rather than let both die. However, all I need for the argument is a more modest claim – that it would be *no worse* (or at least as good) to kill one than to let both die. Let L_2 be the (dis)value of letting 2 people die and K_1 the (dis)value of killing one. Then we can record the information that it would be no worse to kill the one than to let both die in the following inequality:[5]

$$K_1 \geq L_2.$$

So much for the $K_1 - L_2$ scenario. Now suppose that the situation is as before except that there are now two sleeping passengers in the out-of-control trolley rather than one, and still just the one in the stationary trolley. So you could let three die, or else kill two. Again, it seems no worse (at least as good) to divert the trolley into the siding, thereby killing two, than allowing the moving trolley to collide with the stationary trolley, thereby letting all three die. So the $K_2 - L_3$ scenario tells us that:

$$K_2 \geq L_3.$$

Continuing on in this way, it seems we can add as many passengers to the moving trolley as we like without destroying the fact that diverting the moving trolley to kill its passengers is not worse than letting them all die along with the additional passenger in the stationary trolley. That is:

(1) for any n, $K_n \geq L_{n+1}$.

I now want to employ, provisionally, a fairly substantive although plausible assumption: that letting two people die is twice as bad as letting one person die, that killing two is twice as bad as killing one, and so on. That is to say:

(2) for any n, $K_n = nK_1$ and $L_n = nL_1$.

Some people might find (2) plausible, others might not. However, at this stage I want to use it simply to illustrate the form of the argument. It turns out, somewhat surprisingly, that however plausible (2) is, the argument doesn't rely on anything nearly as strong.

One further assumption – the position against which I am arguing certainly endorses the principle that letting a person die is no worse, in itself, than killing. And in any case, this is independently plausible. That is to say:

(3) $L_1 \geq K_1$.

Now, for any n, we have:

$K_n \geq L_{n+1}$	from (1),
$K_n = nK_1$	from (2),
$L_{n+1} = (n + 1)L_1$	from (2),

And these jointly entail:

$$nK_1 \geq (n + 1)L_1,$$

from which, by dividing both sides by n, we immediately get:

(*) $K_1 \geq [(n + 1)/n]L_1$.

In conjunction with (3), our result (*) tells us that for any number n, the disvalue of killing one person is sandwiched between, on the one hand, L_1 (the disvalue of letting one person die), and on the other, L_1 multiplied by the factor [(n + 1)/n].

$$\text{for any n, } L_1 \geq K_1 \geq [(n + 1)/n]L_1.$$

The factor [(n + 1)/n] gets closer and closer to 1 as n gets larger and larger. We can make it as close to 1 as desired, simply by making n sufficient large. Since the formula holds for any n, (*) and (3) jointly entail the value-equivalence thesis:

$$L_1 = K_1.$$

And in conjunction with (2) we can derive the more general equivalence of killing and letting die:

$$\text{for any n, } L_n = K_n.$$

3. OBJECTIONS AND REFINEMENTS

3.1. *Diverting is Not Killing*

Go back to the original case – one passenger in the runway trolley and one in the stationary trolley. Someone might be inclined to deny that switching is, in this case, tantamount to killing. The idea seems to be that since the sleeping passenger in the runaway trolley is going to die anyway, or is going to die at more or less the same time, you cannot really kill him. This just seems false.[6] If this is not immediately obvious, vary the circumstances a little. Suppose the runaway sleeper has on board it some explosives which can be set off by remote control. The explosives were put in place by some terrorists whose plot you have discovered. You now have the controller in your hand. Unfortunately the only way to stop the trolley is to blow it up some distance before it reaches the stationary trolley. If you do so you would clearly be killing the moving passenger, even though he would have died a few seconds later had you not done so. By parity, if the trolley, when diverted, would hit the brick wall just a few seconds before it would have hit the stationary trolley

when left to itself, diverting would be tantamount to killing. Your judgement on the relative value of killing one and letting two die in the explosives case may differ from the brick wall case, but if so it must involve some difference other than that between killing and letting die.

3.2. *Failing to Divert is Not Letting-die*

A related objection attacks the judgment that by failing to intervene (in the original situation) you allow two people to die. Quite generally a necessary condition of your letting P happen is that you be able to prevent P. While it is true that in the original case (one sleeper, one stationary passenger) you can prevent the state of affairs consisting in *both* passengers dying, you cannot prevent the state of affairs consisting in at least one passenger dying – naming, the one in the moving trolley. Thus at most you can prevent one from dying and so at most, by failing to intervene, you allow one to die – the one in the stationary trolley.

To see that the judgement is puzzling, note firstly that it seems undergirded by the idea that you do not allow someone to die if, come what may, he is going to die in some way or other. But consider a doctor with a terminally ill patient who is fairly close to death. He could intervene and give him a fatal dose of something, say morphine, or he could . . . just let him die. That seems the natural way to characterise his dilemma. We might add that he could just let him die *without intervening*. But that seems rather redundant. What this shows is that the relevant necessary condition on letting P happen is somewhat weaker than that stated. You let P happen only if you could either intervene to prevent it happening or intervene to change in some significant respect the particular way P comes about or is realised.

Even if the judgement were not strange and could be granted it would not help the opponent of non-equivalence. For if in the original scenario failing to divert is tantamount to letting only one die then, giving the plausible judgement that it is no worse to divert, we are forced to conclude that killing one is, in itself, no worse than letting one die. Hole in one.

In any case, even granted the objection, the story can be amended to make the intuitive judgement even clearer. Suppose that the run-

away sleeper has one passenger, and that the stationary trolley has two. Allowing the runaway trolley to proceed down the left hand fork is tantamount to allowing (at least) two people to die while diverting is tantamount to killing one. This story does yield the options of either letting (at least) two people die or killing one, and clearly diverting is now clearly no worse (and clearly at least as good as) doing nothing. By adding one passenger to the stationary carriage we generate an even more strikingly compelling argument for assumption (1).

3.3. *Bad Motives for Intervening*

The conclusion of the argument has an obvious but possibly disturbing implication. A runaway trolley with one passenger in it is careering out of control down a track towards the left-hand fork further along which sits the stationary but this time empty trolley. A collision will cause the death of the passenger on the moving trolley. The right-hand fork still leads to the siding with the brick wall at the end of it. You could either do nothing, thereby letting one person die, or you could send the trolley into the brick wall, thereby intervening to cause his death. You could either let him die or you could kill him. According to equivalence, there is no value-difference between the two.

Some may be unwilling to agree that killing one would be no worse than letting one die in this scenario. How can this judgement be squared with the derivation? As it stands, of course, it cannot – something has to do – and those who dislike the conclusion might give up either assumption (1) or assumption (2).

The defender of equivalence can explain away the recalcitrant intuition in the $K_1 - L_1$ case by appealing to an extrinsic evil. The judgement of a value difference in the $K_1 - L_1$ scenario may be undergirded by the idea that in such a case you could have no reason for intervening, or at least no *good* reason, and that a prime candidate for so acting would be malice (perhaps the thought, "I want to be directly *involved* in the cause of his death") which would be lacking if you simply let him die. But such a thought could only constitute maliciousness if you *believed* your killing him was doing him some harm that letting him die would not do. If you thought that killing him was, in itself, doing him less harm than letting him die (a highly

implausible thesis, I admit) that very thought could be part and parcel of a merciful motivation. Perhaps something like this "*I* can't stand by and just let him die. I have to *do* something for him, even if that doesn't make any difference to his level of suffering. I want to be directly *involved* in the cause of his death, to help him on his way." Of course this sounds crazy, because a departure from equivalence in this non-standard way *is* crazy (which is just to say that assumption (3) is highly plausible.) But what it shows is this. Killing him *might* be motivated by malice, but not necessarily. If it were it would be at most an extrinsic difference, and further, one which could only exist where the participant himself believed that killing in itself is worse than letting die.

Let's suppose there is some other difference in value between the two collisions, one which is not insignificant although falling well short of the magnitude of disvalue involved in lost life. Suppose, for example, that the stationary trolley is actually a luggage van with the sleeping passenger's luggage on board – the two trolleys belonged to the same train and have become separated in a bizarre accident. As it happens the passenger was taking with him on his journey all his worldly belongings – a modest but not completely insignificant collection – as well as the only copy of his last will and testament. The collision will destroy his belongings along with his will. So letting him die involves destroying his worldly goods along with his ability to hand them on as his wished. This is definitely a harm to the person – a harm which is avoided by diverting the trolley and thereby killing him. But it is a harm which most people would judge to be much less than that of losing one's life (perhaps excepting those who are both very rich and very old). For example, normally it would not be worthwhile killing someone in order to ensure that his will is carried out accurately. But would it not be better to divert the trolley in such a case? And could not that diversion be motivated precisely by a desire to benefit the sleeping passenger?

To vary the story in a more realistic and pertinent direction, let's suppose that there is a comparatively small but not insignificant difference between the manner of the two possible deaths. Suppose that the passenger's death in a trolley-wall collision would be instant, but that because the two-trolley collision involves transfer of some of the momentum to the stationary trolley, the impact on the passenger's

body would not be as great, and that as a consequence the passenger's death in the latter case would be a lingering affair involving some pain. You, an expert on railway trolley collisions, quickly judge that an instant death would be better than a lingering and painful one. Your desire to spare the passenger suffering thus motivates you to divert the trolley, to spare the passenger unnecessary suffering.

In both these cases a comparatively small difference in value could tip the balance of value in favour of diverting, and could also form the basis of a benevolent motive for so doing.

3.4. *The Possibility of Diminishing Marginal Disvalue*

I have assumed (in (2)) that each additional killing/letting-die is just as bad as the last. Is this right? Is the hundredth killing as bad as the ninety-ninth? Indeed, is even the second as bad as the first? Some might find it plausible to suppose that the additional disvalue of killing another person, or of letting another person die, diminishes. This is not the implausible thesis that it is as bad to kill ten as to kill a hundred, but rather that the difference in disvalue between killing ninety nine people and of killing hundred is not as great as the difference in disvalue between killing none and killing one. The same would presumably apply to the disvalue of letting people die. The difference in disvalue between letting ninety-nine people die and of letting one hundred die, is not as great as the difference in disvalue between letting none die and letting one die. Let's call this the diminishing marginal disvalue of lost lives.

As we will see, this idea of diminishing marginal disvalue has less appeal than it might appear on the surface to have. But for the sake of the argument let's run with it.

The general form of assumption (2) is embodied in (2a):

(2a) for any n, $K_n = f_n K_1$ and $L_n = f_n L_1$.

This simply says that the disvalue of killing n people, or of letting n people die, is some function f of the disvalue of killing one, or of letting one die. Simply generalizing the derivation in section 2, instead of $K_1 \geq [(n + 1)/n]L_1$ we get:

$$K_1 \geq [f_{n+1}/f_n]L_1.$$

which in conjunction with (3), yields:

(**) $L_1 \geq K_1 \geq [f_{n+1}/f_n]L_1$.

Given diminishing (or even non-increasing) marginal disvalue, it can be shown that the ratio f_{n+1}/f_n behaves just like the simpler ratio $(n + 1)/n$ (For a proof of this see appendix, lemma 1). That is to say, given non-increasing marginal disvalue:

f_{n+1}/f_n converges to 1.

From (**) it follows that K_1 is sandwiched between, on the one hand, L_1 and, on the other, something larger than but as close to L_1 as you like. That is, it again follows that

$K_1 = L_1$,

and (in conjunction with 2a),

for all n, $K_n = L_n$.

However, despite its initial plausibility and despite the fact that the argument can be amended to accommodate it, it is not obvious that the idea of diminishing marginal disvalue can be sustained. Does it apply to killings by a single person or is there no restriction on the killers? Let's suppose the former – that the thesis of diminishing marginal disvalue applies to the killings of a single killer. Imagine Simpson plans to kill his ex-wife while a drug dealer plans to kill her boyfriend. In the first scenario they both carry out their plans. But in the second scenario the drug dealer gets delayed by a messy transaction, whereas Simpson, while in the middle of killing his ex-wife, is interrupted by the boyfriend and ends up having to kill the boyfriend too. If neither has killed anyone before, or they have both killed the same number of people before, the thesis of diminishing marginal disvalue would imply that, other things being equal, the second situation is not as bad as the first. If the drug dealer is an old hand at killing and Simpson a neophyte, then Simpson's double killing of ex-wife and boyfriend may well be worse than each killing one, and considerably worse than the drug dealer killing both himself. Also, his ex-wife's murder would have been worse than it is, and

that of the boyfriend's better, if only they had occurred in the reverse order. These seem strange judgements.

Suppose then, that the thesis applies to the total class of killings. It follows that the additional disvalue of each new killing has been steadily declining all over the universe ever since Cain killed Abel. Worse, in a world in which people have been around for an infinite amount of time already, and there has been, say, one killing a year the disvalue of killing will have reached the lower limit. Indeed, it will have been at the lower limit at all previous moments.

However, if diminishing marginal disvalue cannot be defended then the simpler linear assumption on which the first version of the argument is based is that much more attractive despite its logical strength.

3.5. *A Threshold Objection*

A detractor of equivalence might well object that the passage to assumption (1), suggested by a consideration of the $K_1 - L_2$ scenario and then more generally the $K_n - L_{n+1}$ scenarios, was too swift. They might be reluctant to concede that it is no worse to divert when there is just one extra life at stake, that saving one life is too trivial a consideration to outweigh the possible extra disvalue of killing. It is interesting, for instance, that in the original trolley problem and in situational variations of it, killing one is usually contrasted with letting *five* die. Thus the value of four saved lives seems to be a threshold at which most are comfortable with the idea.

Let there be some number of people (say, n + 1) in the stationary trolley and just one in the trolley out of control. Then I have assumed that even if n is 1, it is no worse to divert the runaway trolley. But perhaps one is too small. Some might think that for every n, n is too small – that it would always be worse to divert the trolley and kill 1 than to let (n + 1) people die. This embodies the absolutist idea that killing is so bad that no number of saved lives could ever compensate for the disvalue of it. There would probably be no way by means of these thought experiments to budge that opinion. However, consider someone who is not quite that extreme, someone who thinks that while saving one life may be too trivial to justify taking the step of killing, there is nevertheless some number T (T for threshold) of lives such that it would be no worse to kill 1 than to let (T + 1)

die. Perhaps T is large – say, ten, or a thousand, or even a trillion. But once we had that many people in the stationary trolley (or on a planet under threat from an accidental nuclear holocaust) it would be no worse to engage in killing, than to let the moving trolley (or the accidentally released missile) career into the stationary trolley (or initiate the holocaust) thereby letting a T + 1 die:

$$K_1 \geq L_{1+T}.$$

Now suppose we add a passenger to each trolley. Since we have reached the threshold where killing becomes no worse, wouldn't it still be no worse to divert the moving trolley? If so we have:

$$K_2 \geq L_{2+T}.$$

Continuing adding a person to each trolley, we have more generally:

(1a) $K_n \geq L_{n+T}.$

Now combining this with (2a) we get:

$$K_1 \geq [f_{n+T}/f_n]L_1.$$

which in conjunction with (3), yields:

(**) $L_1 \geq K_1 \geq [f_{n+T}/f_n]L_1.$

Given non-increasing marginal disvalue, it follows that the ratio f_{n+T}/f_n behaves just like the simpler ratio f_{n+1}/f_n. (See appendix, lemma 2.) That is to say

f_{n+T}/f_n converges to 1.

And this is enough, as before, to guarantee equivalence. Thus even if there is a threshold T, no matter how large, such that you concede that it would be no worse to kill 1 than to let T + 1 die, and such that, more generally, it would be no worse to kill n than to let (n + T) die,

you are still committed to the idea that killing one is no worse than letting one die.

3.6. *Infinitesimal Differences in Value*

Suppose that in making decisions about what to create God operates two principles of choice. One is *the more the better* and the other is *the bigger the better*. Further, she ranks these two principles in that order. Thus if facing a choice between creating three mice and two elephants she goes by number of creatures, preferring three creatures to two. If facing a choice between three mice and three elephants she goes by bulk, preferring three big things to three small ones. Where Mice_n is the value of n mice, and Elephants_n is the value of n elephants we thus thus have:

(*) $\text{Mice}_{n+1} > \text{Elephants}_n$ (by numbers) and
$\text{Elephants}_n > \text{Mice}_n$ (by bulk).

Surely this is a coherent evaluative ordering, one which could consistently guide God's creative choices? However, it is not difficult to see that God's two-tier ordering entails equivalents of assumptions 1 and 3:

$$\text{Mice}_{n+1} \geq \text{Elephants}_n \text{ and } \text{Elephants}_n \geq \text{Mice}_n.$$

If we add either linearity (that God values mice equally and values elephants equally) or diminishing marginal values of extra animals, then the argument yields:

(**) $\text{Mice}_n = \text{Elephants}_n$,

contradicting (*). *Something* is wrong. Does the fault lie in God's ordering, or in the argument?

So far I have been tacitly assuming that a coherent evaluative ordering should be representable by an assignment of real numbers. Only given this assumption, together with our other substantive assumptions, can we apply the mathematical operations to derive the evaluative equivalence of killing and letting-die. The assumption of representability in the reals has been there all along, albeit hidden in the formulation. The apparent incompatibility of the two-

tier ordering with linearity (or diminishing marginal value) can be resolved if we drop the requirement of representability within the reals. Furthermore, we can still have numerical representability provided we include infinitesimally small numbers along with the usual real numbers.

An infinitesimally small number is a number larger than 0, but smaller than any positive real number (other than 0). While the normal arithmetical operations of addition and substraction apply to infinitesimals if you multiply an infinitesimal, ι say, by a real number r, the result, ιr, is itself an infinitesimal. That is, for any r, ιr is also smaller than any positive real number (other than 0).

In God's evaluative ordering each animal is assigned a value which is within an infinitesimal of any other animal's value. Let's take the value of a mouse to be our unit – each mouse is worth *exactly* one unit of value – and assume for simplicity that all mice are equal: n mice are worth n units of value. An elephant, however, is worth ever so slightly more than a mouse, that is: $(1 + \iota)$ units where ι is an infinitesimal. All elephants are equal, so n elephants are worth $n(1 + \iota) = (n + n\iota)$ units of value. $(n + n\iota)$ is larger than n (albeit infinitesimally so) and so n elephants are strictly more valuable than n mice. However, since $n\iota$ is an infinitesimal $n\iota$ is smaller than any real number, including 1, and so $(n + n\iota)$ is smaller than $(n + 1)$. So $(n + 1)$ mice are always strictly more valuable than n elephants. The two-tier ordering is preserved.

Thus if we were to allow representability by a system of reals supplemented with infinitesimals it would be possible to preserve the thesis of the evaluative non-equivalence of killing and letting-die. Further, it is not difficult to see that our original proof shows that if killing is worse than letting die it can only be so by an infinitesimal amount. Non-equivalence can be coherently preserved and even numerically represented.[7]

If we allow infinitesimals, the extra added disvalue which killing adds to letting-die is infinitesimal compared with the intrinsic disvalue of letting someone die.

$$K_1 = L_1 + \iota L_1.$$

Now imagine three seconds of awful pain. Such an amount of pain has some disvalue, say P. Let's suppose that there is some minimal sequence of such painful episodes, of length m, say, such that it would be no better or worse to allow a person to suffer such a sequence than to allow her to die (m would obviously be very very large, but still finite). Then:

$$mP = L_1 \text{ and so } P = L_1/m.$$

It follows that:

$$L_1 + P = L_1 + L_1/m = (1 + 1/m)L_1.$$

But since 1/m is a real number $(1 + 1/m) > (1 + \iota)$ and so (since is L_1 negative):

$$(1 + \iota)L_1 > (1 + 1/m)L_1.$$

Hence:

$$K_1 > L_1 + P.$$

That is to say, given that a brief episode of awful pain has a small but finite disvalue (compared with letting-die) killing a person would be preferable to letting them die while in addition suffering that brief episode. In other words, if killing is only infinitesimally worse than letting-die then anything of finite disvalue, like enduring a brief episode of pain, will more than make up the value difference between the two.

3.7. *Modest Ubiquity*

The method of bare-differences (as well as the method of clear-differences) involves a controversial ubiquity thesis.[8] However, there are two quite different ubiquity theses which are not always clearly distinguished and disentangled. The first, which I will call *bottom-up* ubiquity is the modest assumption that if B is worse in itself than A, then merely replacing A by B (holding all else fixed) will make the overall situation worse. Slightly more formally let S(B/A) is the situation which is just like S except that A has been replaced by B:

> *Bottom-up ubiquity*: If B is worse (in itself) than A then (for any situation S) S (B/A) is worse than S.

What could it mean that B is *in itself* worse than A if merely replacing A with B did not make things worse? Another quite different principle, which I will call *top-down ubiquity*, says that if merely replacing B with A makes things worse in *any* single case, then B in itself is worse than A

> *Top-down ubiquity*: If (for some situation S) S(B/A) is worse than S, then B is worse in itself than A.

The ubiquity principle that is sometimes criticized is the conjunction of both bottom-up and top-down ubiquity. The conjunction is by no means a modest principle. The two principles are naturally useful for establishing quite different theses. Using bottom-up ubiquity we can refute the thesis that B is, in itself, worse than A, by means of a single counterexample. All we need is a single situation S, such that replacing B for A in S does not make things worse. Refutation of claims of intrinsic value are thus rather easy.

Suppose we want to establish the thesis that B is, in itself, worse than A. We cannot do so using the bottom-up principle. However, with the top-down principle we can establish the thesis in one swift blow. A single pair of barely different situations, S and S(B/A), is sufficient to order A and B. Thus top-down ubiquity is an incredibly powerful principle, even in isolation, but especially when employed in conjunction with bottom-up ubiquity.[9] Given both principles, a single barely different pair has implications for all barely-different pairs of that kind.

Judith Jarvis Thomson thinks that chopping off a person's head is, in itself, worse than punching a person on the nose – and the rest of us are naturally inclined to agree.[10] But imagine that Jones has a strange nasal condition such that punching him on the nose would kill him. This makes a barely different pair out of punching Jones on the nose and chopping off Jones's head. In these conditions it seems that punching Jones on the nose would be just as bad as chopping off Jones's head. Thomson suggests, however, that this one case should not force us to revise our judgement that chopping off a person's head is, in itself, worse than punching a person on the nose. If Thomson is right then even bottom-up ubiquity fails, and the existence of a single value-identical barely-different A/B-pair is not sufficient to refute a claim of intrinsic worseness.

Thomson's judgement on the pair is not obvious to me. There are bad aspects of a head chopping which are absent from a nose-punching, even when both kill the victim. For example, the former displays less respect for the integrity of the victim's body. (Hanging, drawing and quartering is worse than hanging.) But if we do grant Thomson's particular comparative judgement then I think we should indeed give up the idea that *in itself* head-chopping is worse than nose-punching. However, this conclusion is not nearly as counter-intuitive as it first might sound. What survives such examination by barely-different pairs is the claim that chopping off a *normal* person's head is, in itself, worse than punching a *normal* person on the nose. That is, a *normal* head-chopping is worse than a *normal* nose-punching – and that is surely what was intended all along. The unstated condition of normality builds in various important features of noses and necks. However, it would not do to try and rescue the traditional judgement on killing and letting-die in this way, because there is no comparable notion of a *normal* killing or a *normal* letting-die in the sense in which there are normal noses and normal necks.

Although there is no such notion as that of a normal killing, perhaps a related point can be made about *typical* killings. Just as a normal A may be worse than a normal B, as it happens so too a *typical* A may be worse than a *typical* B, because of typical concomitants of A and B.[11] As it happens, a typical killing, for instance, has horrible features which a typical letting-die lacks – malicious intent, unnecessary suffering, acting without the person's informed consent, perhaps violation of certain rights, and so forth. But, according to the equivalence thesis, it is those other horrible features which make killing typically worse than letting-die. And it is precisely in those situations in which the badness of killing humans is controversial – in the case of the terminally or congenitally ill, say – that these other horrible features may well be absent from a killing and present in a letting-die. The point of the bare-difference arguments is to force us to abstract from the typical concomitants of killing or letting-die and focus on the possible value-contribution of killing and letting-die in themselves.

Note that some have been tempted to cite a single barely-different, value-equivalent A/B-pair as a refutation of the claim that the A/B-distinction is evaluatively *significant* or *relevant*.[12] Such a conclusion

is indeed unwarranted but that is because evaluative significance is a relatively weak condition. Compare the situation with causal relevance or significance. Suppose a wet, oxygen-surrounded match fails to burn when struck, and so too does a wet, oxygen-deprived match. It would clearly be absurd to take this as a refutation of the causal relevance of oxygen to burning. Being oxygen-surrounded is causally relevant to burning if the bare shift from being oxygen-deprived to being oxygen surrounded *can* (rather than *must*) make a difference to burning. Similarly the moral or evaluative significance of the A/B-distinction is guaranteed if substituting B for A can make a difference to value. The existence of a single case S in which S(B/A) is worse than S makes the A/B distinction evaluatively significant, but provided top-down ubiquity is false, the evaluative significance of A/B is quite compatible with other cases in which the substitution of B for A makes no difference to value.

The upshot is that bottom-up ubiquity is a necessary constraint on intrinsic value, and given this we can refute a claim of intrinsic worseness either by means of a single barely-different, value-equivalent pair, or by means of a clear difference argument which implies the existence of such pairs.

3.8. *Do Killing and Letting-die Have an Intrinsic Value?*

Another problem emerges. Without top-down ubiquity, a bare-difference or clear-difference refutation of the claim that A is worse than B, even in conjunction with the assumption that B is not worse than A, does not establish the intrinsic value-equivalence of A and B, or the evaluative-insignificance of the A/B-distinction. Similarly, demonstrating that in a certain class of cases the shift from letting-die to killing does not makes things worse does not show that they have exactly the same intrinsic value. For all I have shown, there could be cases in which the bare shift from killing to letting-die makes an evaluative difference, through connections with other evaluatively significant features. As it stands the argument in section 2 establishes only that the difference between killing and letting-die makes no difference in a certain range of cases. More is required to show that killing and letting-die are quite generally of the same intrinsic value.

One way of filling the gap would be, of course, to employ top-down ubiquity, but this is a very strong and controversial principle.

Another would be to assume, simply, that killing and letting-die are *comparable* for intrinsic value. Then if killing is no worse than letting-die (assumption 3) either killing is evaluatively equivalent to letting-die, or killing is worse than letting die. The argument then proceeds by refuting the last disjunct. To summarize then, the argument establishes the following interesting and rather important result:

> If killing and letting-die are comparable for intrinsic value, then they are (within an infinitesimal) evaluatively equivalent.

The condition here is by no means a trivial one. However, to justify a strong presumption against active euthanasia while retaining a soft spot for passive euthanasia, one would need to invoke the thesis that killing is, in itself, worse than letting-die. If it is not worse to kill then each case will have to be assessed on other merits and demerits. But the thesis that killing is worse, in itself, than letting die implies that killing and letting-die are comparable for value.[13] And that's all we need to establish the equivalence thesis.

APPENDIX

Lemma 1 Given marginal non-increasing disvalue, the sequence f_{n+1}/f_n converges to 1.

Proof $f_1 = 1$ and $f_{n+1} = f_n + i_n$, where i_n measures the extra disvalue of the $(n + 1)^{th}$ case of killing. That is, i_nK_1 is the extra disvalue of the $(n + 1)^{th}$ case of killing, i_nL_1 is the extra disvalue of the $(n + 1)^{th}$ case of letting-die.) It is clear that (other things being equal) you can't mitigate previous killings simply by killing some more, so f has to be a non-decreasing function of n (that is, for each n, $f_{n+1} \geq f_n$), and so i_n is some number greater than or equal to 0. The assumption of non-increasing marginal disvalue tells us that for any n, $i_{n+1} \leq i_n$. From elementary analysis we know that if i_n is a non-increasing sequence with a lower bound of 0, then i_n converges to some limit ≥ 0. Consider now the sequence:

$$f_{n+1}/f_n = (f_n + i_n)/f_n = 1 + i_n/f_n.$$

Either f_n is bounded above or it is not. Suppose not. Then whatever the limit of i_n, the sequence i_n/f_n clearly converges to 0 – since i_n converges and f_n grows without bound – and so f_{n+1}/f_n $(= 1 + i_n/f_n)$ converges to 1. Now suppose f_n is bounded above. Then, for the sake of a reductio, suppose that i_n converges to some number larger than 0, say r. Since i_n is non-increasing, $i_n \geq r$, and so $f_{n+1} \geq f_n + r$ for every n, and f_n is unbounded afterall – contradiction. Hence i_n converges to 0. Since f_n is a non-decreasing bounded sequence it converges to some number s, say. Thus i_n/f_n converges to $0/s = 0$, and again f_{n+1}/f_n $(= 1 + i_n/f_n)$ converges to 1.

Lemma 2 Given marginal non-increasing disvalue, the sequence f_{n+T}/f_n converges to 1.

Proof It is sufficient to note that

$$f_{n+T}/f_n = [f_{n+T}/f_{(n+T)-1}]\,[f_{(n+T)-1}/f_{(n+T)-2}] \ldots [f_{(n+2)}/f_{n+1}]\,[f_{n+1}/f_n]$$

and that for any j, $[f_{(n+T)-j}/f_{(n+T)-j-1}]$ tends to 1 if $[f_{n+1}/f_n]$ tends to 1. The limit of a product is the product of the limits, and so f_{n+T}/f_n also tends to 1.

NOTES

[1] Rachels (1975) and (1986), pp. 111–4, and Tooley (1980).

[2] In particular, several authors (Kagan (1988), Malm (1989), Thomson (1976)) have questioned the important "ubiquity" thesis which is presupposed by the method of bare-differences. In section 3.7 ubiquity is discussed at some length.

[3] See Foot (1978) and Thomson (1976).

[4] Both judgements – that by diverting you kill one and that by leaving things as they are you let two die – can be challenged. For the sake of clarity and the unimpeded exposition of the argument I will deal with all such objections in section 3.

[5] Obviously I am allowing here for the possibility of cases where differences may make a difference to value. In the usual trolley stories the one who is killed is distinct from any of the five who are allowed to die and that might make a difference. The particular relations between the potential intervener and those under threat may also make a difference.

[6] And not only false, but certainly self-defeating as part of a strategy for defending the evaluative difference between, say, active and passive euthanasia!

[7] I am indebted here to criticisms and comments of my colleagues Luc Bovens, Steve Leeds, and Michael Tooley. Luc and Steve both urged the rationality of

lexicographic orderings, suggesting my original argument must be wrong because too strong. Steve suggested God's preferences. Finally, Michael suggested the use of infinitesimals to represent such orderings.

[8] Kagan (1988).

[9] For a more extensive discussion of the two principles and their relation to organic unity see my (forthcoming) "Axiological atomism".

[10] Thomson (1976), p. 69.

[11] It is important to note that *being normal* and *being typical* are two quite different things. *Being typical* is a statistical concept, whereas *being normal* is normative. Abnormal noses, like Jones's, could well become typical. Does "abnormal murders are becoming typical" make sense?

[12] Malm exposes the error in her (1992).

[13] I am indebted to Luc Bovens, Steven Leeds and Michael Tooley for reading earlier drafts and making excellent criticisms and suggestions. I also thank the audiences of the Denver Metro Colloquium and the University of Miami Colloquium for comments and advice.

REFERENCES

Fisher, J. M. and Ravizza, M. (eds.) (1992): *Ethics: Problems and Principles*, New York: Harcourt, Brace Jovanovich.

Foot, P. (1978): 'The Problem of Abortion and the Doctrine of the Double Effect', reprinted in Fisher and Ravizza (1992), pp. 59–68.

Kagan, S. (1988): 'The Additive Fallacy', reprinted in Fisher and Ravizza (1992), pp. 252–271.

Malm, H. (1989): 'Killing, Letting Die, and Simple Conflicts', in Fisher and Ravizza (1992), pp. 133–145.

Malm, H. (1992): 'In Defense of the Contrast Strategy', in Fisher and Ravizza (1992), pp. 272–277.

Rachels, J. (1975): 'Active and Passive Euthanasia', reprinted in Fisher and Ravizza (1992), pp. 112–116.

Rachels, J. (1986): *The End of Life*, Oxford: Oxford University Press.

Thomson, J. J. (1976): 'Killing, Letting Die, and the Trolley Problem', reprinted in Fisher and Ravizza (1992), pp. 69–76.

Thomson, J. J. (1985): 'The Trolley Problem', reprinted in Fisher and Ravizza (1992), pp. 280–292.

Tooley, M. (1980): 'An Irrelevant Consideration: Killing versus Letting die', reprinted in Fisher and Ravizza (1992), pp. 106–111.

Department of Philosophy
University of Canterbury
Christchurch
New Zealand