Solving the Trolley Problem

Shelly Kagan

One might despair of ever arriving at a principle adequate to capturing and accommodating our intuitions about the full range of cases that have come to be known as "trolley problems" (roughly speaking,[1] cases where one must choose whether to kill some to save others). But suppose there were such a principle, as indeed I imagine there probably is. For the moment, just call it Q.

Is there such a principle? As I say, I find it plausible to think there is. After all, *something* generates our intuitive reactions to cases. So the odds are there is *some* statement of a rule or law (or a set of rules or laws) that accurately predicts our intuitions. Properly reformulated, this rule could provide the proposed Q. (Here's the idea behind this talk of "reformulating" the rule: start with a rule that accurately predicts the precise circumstances in which we will have the intuition that a given act is permissible; restate it as a moral *principle*, one which correspondingly asserts that acts are permissible in precisely those circumstances. This principle will, by hypothesis, match our intuitions about cases; so that should be the desired Q.)

Admittedly, we might not *always* judge in conformity to Q: perhaps in some situations, or when thinking about certain cases, various psychological factors interfere with our ability to judge in perfect conformity with Q. Even if so, it might still be the case that Q is the best match for our various intuitions. But for simplicity, let us put this complicating possibility aside and suppose we can

indeed find a principle—Q—that really does match our intuitions *perfectly*.

More troubling is the possibility that people's intuitive reactions to trolley cases may not all be the same. Perhaps different people will judge some of the relevant cases differently. Then there may not be a single principle that matches everyone's intuitions perfectly; no single Q will fit all. I'll come back to this worry later. But for the time being, at any rate, let us suppose that people's intuitions *are* similar. So we should, in theory, be able to find a single principle, Q, that matches everyone's intuitions.

Suppose, then, that we had Q. Are we done? Have we solved the trolley problem? Far from it! For it might well be that the concepts and distinctions on which the principle Q turns are ones that we are not prepared to embrace, once they are so identified.

That might seem unlikely, since we typically have rather robust intuitions about trolley cases, and we certainly care deeply about right and wrong in matters of killing and letting die. How could it turn out to be the case that we don't care about the distinctions underlying our intuitive judgments?

The answer, of course, is that the distinctions and concepts that underlie our intuitions about the cases might not be ones that seem morally *relevant*, in and of themselves, when we think about them directly.

Suppose, for example, that Q involved elements like this: our intuitions (about which actions are permissible) depend on whether we are making the judgment on a Wednesday or a Sunday, or whether we make it standing up or sitting down, or whether we make it when we are warm or when we are cold (or holding a cup of coffee or a cold drink).

If something like this were the case—if these factors really were what our intuitions turned on—then I imagine that we would resist embracing these factors as morally relevant, and thus we would end up rejecting Q, even if it were a perfect fit for our intuitions

about those cases, and no other principle came close to matching our intuitions as well.

To be sure, in the examples I just gave the factors I've identified are about the *judges*, not about the *cases*, and one might hope that Q will take as relevant input only features of the trolley cases themselves (and not features of the person having the intuition). But the same thing could in principle happen even with regard to the cases themselves.

Thus, for example, it might be that Q involved elements like this: whether the act takes place on a Wednesday or a Sunday, whether the person being killed is standing up or sitting down, whether it is warm or cold outside when the agent acts (or whether the victim is holding a cold drink or a cup of coffee).

Here too, I imagine, we would reject Q—even if it matched our case specific intuitions perfectly. The distinctions on which it turns would not be ones we would see our way to endorsing as morally relevant.

To be sure, since we are stipulating that Q matches our intuitions about cases, any distinction important to Q will be "morally relevant" in at least *one* sense of that term: it will be relevant to generating our intuitions about certain important moral questions. But for all that, the distinctions in question might still be morally irrelevant in the sense that we simply cannot see why anything like that should *matter* morally. We may be simply—and appropriately—unwilling to embrace the factors in question as having any genuine moral significance.

What we want from an acceptable moral principle, after all, is not merely that it match our intuitions about particular cases; we also want to be able to see why the various factors appealed to by the principle should matter morally. So if we cannot see, directly, why the factors in question should matter in their own right, then at the very least we need to be able to see how they connect to still other factors, ones whose moral significance we can indeed

appreciate directly. In effect, we must be able to provide the principle with an attractive and plausible rationale.

Suppose, however, that neither of these holds for Q. Suppose, that is, that although Q matches our various case specific intuitions, we cannot see why the factors to which it appeals should matter morally, nor can we provide it with a compelling rationale. Then Q will be unacceptable—this, *despite* the fact that it matches our intuitions about the cases. What we will conclude, instead, is that at least some of the intuitions underlying Q are mistaken, influenced by morally irrelevant factors.

Of course, there is certainly no guarantee that Q *will* turn on such morally irrelevant elements. But I suspect that the more complicated Q turns out to be, the greater the danger that it will indeed appeal to features that do strike us as morally irrelevant. And if Q ends up having lots and lots of clauses, or depends on abstruse concepts—or both—it becomes proportionately less likely that a plausible rationale can be provided.

That's the situation I think we actually find ourselves in. I suspect that any principle actually capable of matching our intuitions across the entire range of trolley problems will be a messy, complicated affair, full of distinctions that strike us as morally irrelevant—or at least, will strike us that way once we directly face the question of whether something like *that* could indeed matter morally.

Of course, complexity itself needn't be a problem. An acceptable moral principle might have a large number of clauses and yet it might still be the case that each clause is independently attractive and plausible (or can be adequately motivated). And on the other hand, even a fairly simple principle might turn on a distinction for which no plausible rationale can be provided. Imagine, for example, a principle that permitted killing on all and only sunny days. This would be simple enough, but even if—somehow!—this perfectly matched our intuitions about the relevant cases, we would appropriately reject the principle as turning on a morally irrelevant

factor. Absent a compelling rationale, the principle would be unacceptable.

With all of this by way of background, let us turn, now, to Frances Kamm's two difficult but incredibly stimulating lectures on the trolley problem.

I think it fair to say that no one has worked harder to solve the trolley problem than Kamm has. Over the years she has probably examined hundreds of different cases, and she has struggled mightily to produce a principle that matches our intuitions about those cases—the elusive principle that I have been calling Q. In her two lectures, Kamm shares with us a summary of some of her recent thinking about the subject. And one of the most important things she does here is to sketch an outline of her proposed solution to the trolley problem.

Kamm calls her proposal the *Principle of Permissible Harm*. She both introduces the basic distinction lying behind the principle (see especially pp. 62–64, and pp. 66–67) and gestures toward some of the extra complications that she feels would be required for a full and complete statement of that principle. In neither lecture does Kamm provide us with a complete statement of the principle, though she does, I think, say enough to give us at least a rough feel for the basic idea. She also says enough to make it clear that a complete statement of the principle would be a complicated matter indeed. (An earlier work, *Intricate Ethics*, also discusses the Principle of Permissible Harm and gives a somewhat fuller statement of it.[2] But even that version is incomplete, though it takes more than half a page to write down.)

I don't have the space here to consider the various complications that Kamm thinks would need to be incorporated into an adequate statement of the Principle of Permissible Harm. But I do want to raise a question about the central distinction upon which the principle seems to turn.

Let me start, then, by trying to give a rough, intuitive gloss on what I take to be the basic idea (I am skipping many details):

Sometimes someone who is killed is killed by an event that is the very same event as the saving of a larger number of people (the greater good). Then the killing of the one may be justified. But in other cases, the one who is killed is killed by something that is merely a causal *means* to the event that is the saving of the larger number. In such cases killing the one is *not* justified. So the crucial question is whether the event that results in the killing of the one literally constitutes the saving or is merely a means to that saving.

As I understand it, then, Kamm's Principle of Permissible Harm makes essential use of the idea (or something close to it) that some events are themselves the very same event as the saving of the five, the existence of the greater good, while other events are not themselves the saving of the five, but only a causal means to it. Of course, the former events may not be *described* as the saving of the five, but they are—as Kamm puts it—"the noncausal flip side" of the saving of the five: they *constitute* the saving of the five. In contrast, other events are metaphysically distinct from the saving of the five; they are mere means to it. And Kamm's idea is that it makes a significant moral difference if the harm done in a given case is caused by the saving of the greater number *itself* (or its noncausal constitutive flip side), or if, instead, it is caused by something that is itself merely a causal *means* to the saving of the greater number.[3]

I should hasten to admit that this may be too rough an account of the precise distinction that Kamm has in mind. Indeed, this way of putting the idea is somewhat at odds with Kamm's remark (on p. 63) that being the noncausal constitutive flip side of saving the five is "very close" to being the very same event as the saving of the five. This implies, after all, that the event that is the noncausal constitutive flip side of the saving of the five is not, as I have just been suggesting, literally the very *same* event as the saving of the five (albeit under a different description). There is, to be sure, a very close relation between the two events (the latter is partially "constituted" by the former); but it isn't quite identity.

I have to confess, however, that the exact relation that Kamm has in mind—what exactly is involved in one event being the noncausal constitutive flip side of some other event—isn't altogether clear to me. I can think of a few different things that Kamm may have in mind, and she doesn't elaborate. Happily, however, she does imply that the difference between identity, on the one hand, and being the noncausal constitutive flip side, on the other, may not be a morally significant one. And as I have noted, she thinks the two relations (identity and being the noncausal flip side) are very close. So perhaps we won't be far off if we try to understand the key distinction in something like the terms I have offered.

(Whatever it is precisely that Kamm has in mind, there is obviously a fair bit of metaphysics being presupposed here about the proper individuation of events, and that metaphysics may or may not be correct. What's more, even if we grant the metaphysics, it isn't obvious to me that the noncausal constitutive relation always holds in all those cases where Kamm needs to find it. But I will let these two points pass.)

So here, again, is the basic idea underlying Kamm's proposed Principle of Permissible Harm: killing some may be permissible if they are killed as a result of an event that is the very same event as (or the noncausal flip side of) the saving of the many (the greater good); but it will not be permissible if they are killed as a result of an event that is merely a causal *means* to the saving of the many.

Now it should be borne in mind that much of what Kamm is doing here—in trying to lay out the Principle of Permissible Harm—is simply a kind of psychological reconstruction. She is doing her best to identify the various features that actually influence our intuitions about the different cases. From this perspective, presumably, it is not particularly important whether the concepts to which she appeals are familiar ones, or whether they are easy to articulate. Even if we don't recognize them, or find them difficult to grasp, they might still underlie our intuitions. As Kamm remarks

at one point, "people may not be able to articulate these proposals, which nevertheless underlie their judgments" (p. 61).

But, of course, Kamm is interested in more than psychology. She is looking for the *correct* moral principle, the one that *accurately* tells us when it is permissible to act. When considered from this perspective, however, it *is* troubling that the key concepts behind the Principle of Permissible Harm are so unfamiliar (in a moral context, at least), and so difficult to articulate. For the simple fact is that the key distinction to which Kamm appeals has no obvious moral significance. When we directly consider the difference between a harm being caused by the saving of the many (or its noncausal flip side) and its being caused by a mere means to the saving of the many, it isn't at all obvious—to me, at least—why a difference like that should *matter* morally. Viewed from this perspective, then, Kamm's proposed principle is in desperate need of a compelling rationale. Unfortunately, this is something that Kamm says very little about.

In fairness, of course, I do want to note that it isn't as though Kamm says nothing at all by way of offering a deeper rationale for the Principle of Permissible Harm. In an important passage (p. 69), she suggests that its central distinction may correspond to a further one—between subordination and mere substitution—which in turn directly connects to deeper moral questions about "persons and their status."

In our lectures, however, Kamm confessedly has little to say about this idea. So let me try to elaborate, I hope sympathetically and accurately, on her behalf.

Intuitively, some ways of treating people involve viewing them as being less valuable than others—as "subordinate" (in Kamm's term). The paradigmatic instance of this, I suppose, is slavery: we subordinate one person to another, treating the subordinate as a mere means to meeting the interests of the slave owner, the "superior." This is, of course, morally abhorrent. Kamm's thought, then, is that *some* ways—perhaps most ways—in which someone might

be killed so as to save others will involve this kind of subordination. And as such they will be unacceptable.

But in *some* cases, perhaps, someone can be killed in the course of saving others where this kind of subordination is *not* involved. Of course, the choice will still have been made to kill the one, say, rather than letting the many die (or be killed). So a kind of "substitution" will have been made: the death of the one (or the deaths of the few) will be chosen—substituted—for the deaths of the greater number. But this is "mere" substitution, as we might put it, not subordination. So the objection to killing that is present in normal cases will be absent here, and the permissibility of proceeding will be intelligible.

The thought, then, is that in the standard Trolley Case (where I turn the trolley from the five to the one), we have mere substitution, and so turning the trolley is indeed permissible; but in cases like Topple (where I topple the fat man onto the track, and his weight stops the trolley as he is killed by it), and in many other cases as well, we instead have subordination, and so the relevant acts are forbidden.

Although Kamm doesn't spell this out explicitly (though see pp. 75–76), the suggestion we're discussing is also closely related to a further point, one she emphasizes in the second lecture—namely, that what matters here may be "intervictim" relations, rather than, primarily, relations between the agent and the person killed. In cases of subordination, it isn't that the one killed is subordinated to the *agent*, but rather that the one killed is subordinated to the interests of the *others*, for whose sake the one is killed. It is the relation between the various potential "victims" that is key here, according to Kamm.

That's a promising idea, and surely we can all feel the appeal of the suggestion that something like this is going on in the various trolley cases: perhaps those acts that are intuitively unacceptable can be shown to involve subordination, while those that are intuitively acceptable involve mere substitution.

And obviously enough, Kamm thinks that this is what we actually have. More particularly, she thinks that this is what the Principle of Permissible Harm spells out for us. Not only does Kamm's principle (or a suitably elaborated version of it) sort the cases properly—as we can suppose it does—but Kamm suggests that it does so by way of distinctions and concepts that line up and interact so as to distinguish cases of subordination from mere substitution.

But it is exactly this that I do not see when I try my best to understand the principle. Consider, for example, what Kamm says about the standard Trolley Case as opposed to the *Two* Trolleys Case (where I deflect one trolley to bump a second trolley that would otherwise kill five, but the first kills someone else while it is en route to bumping the second; introduced on p. 61). The former, Kamm assures us, is mere substitution (see, for example, pp. 69, 75–76, and 97 n. 18), the latter, subordination (p. 76).

Why? Because in the former case, it is the noncausal flip side of the saving of the five (that is to say, the turning of the trolley) that is the cause of the death of the one, while in the latter case, the cause of the death of the one (that is, the turning of the *first* of the two trolleys) is the mere causal *means* to saving the five. And according to Kamm, when I kill you by doing something that merely causes the five to be saved, I subordinate you; but when I kill you by doing something that is itself the saving of the five (or is its noncausal flip side), I only substitute (pp. 75–76).

Well, that is what Kamm *says*. But why in the world should we *believe* her? Notice that in Two Trolleys the death of the one is not itself a *means* to saving the five. If it were, we might well agree that this was a case of subordination (one person being *used* to save others). But that isn't what we have in Two Trolleys. Rather, the death of the one is a mere side effect of saving the five. So why, then, is it subordination?

Because, Kamm says, in Two Trolleys the death of the one is caused by something that is itself a mere causal means of saving

the five, rather than something that is *itself* the saving of the five (or its noncausal flip side)!

Well, yes, there is that difference, but that wasn't our question. Our question was why this makes this a case of *subordination*—something that we can independently see the moral significance of. What is it that is particularly morally offensive (unacceptable subordination) about killing someone via something that merely causes the rescue, as opposed to killing via the rescue itself?

I just don't see it. Kamm herself offers nothing to support the idea that there is something intelligibly offensive (let alone worthy of the name *subordination*) in the former case (killing via the mere cause of the rescue) but not the latter (killing via the rescue itself). And what I suspect, of course, is that there is *nothing* to say. There is nothing to be said about why *that* difference should matter.[4]

To be sure, we ought to be able to point to a feature that is present in those cases of killing that strike us as permissible, a feature that is missing in the cases of killing that strike us as impermissible. Perhaps Kamm really has managed to identify the feature that explains our intuitions in this way. And having identified this feature, we can slap a label on it. We can *say* that the one kind of case involves "mere substitution," while the other unacceptably involves "subordination." But appearances to the contrary notwithstanding, that is not to offer an independently attractive and intelligible rationale for the principle we have articulated. It is simply to assume that since the principle in question matches our intuitions about specific cases, the factors on which it turns *must* be morally relevant. And that assumption, I believe, is highly suspect.

This, I believe, is the situation we find ourselves in with regard to Kamm's Principle of Permissible Harm. Perhaps the principle does a splendid job of matching our intuitions about cases. At any rate, I am not foolish enough to take Kamm on at her own game, to try to find a counterexample to her proposal, and to offer some alternative Q. No, my complaint is not that the principle doesn't match

our intuitions about cases. It is that it cannot be given anything remotely like a plausible rationale. And in the absence of such a rationale, we should reject it.

Of course, this is not to concede that the Principle of Permissible Harm *does* match our intuitions. Indeed, I doubt that it does, at least not for everyone. A very informal survey of students in my upper level normative ethics course at Yale in Spring 2013 suggests otherwise. Thus, to mention only two examples, half my students judged one case (described at the bottom of p. 42) in a manner contrary to the intuition that Kamm is trying to accommodate, and about three fourths judged a second case (described at the top of p. 40) in a contrary manner. While this hardly shows that Kamm's own intuitions are mistaken, it does reinforce the thought that it would be best to shore up the argument for the Principle of Permissible Harm by providing it with some independently attractive rationale.

Admittedly, in principle Kamm might try to dismiss the intuitions of my students—and others who disagree with her—perhaps arguing that her own intuitions are more reliable, having been shaped by years of training and reflection. I would not want to be dismissive of such a claim. But it should be noted, in any event, that Kamm does not in fact privilege her own intuitions in this way in our lectures. On the contrary, although she often simply reports her own judgments about cases, she also makes assertions about what, for example, "people would think" (p. 61) or what "we think" (p. 64), and she talks about "our intuitive judgments" as well (p. 13). So it does seem problematic for Kamm if others fail to share her intuitions; and this makes the failure to provide a plausible rationale for the Principle of Permissible Harm all the more significant.

Nonetheless, it may seem that I have been unfair to Kamm. Although she does offer a few remarks about the difference between subordination and substitution, and how this may provide a rationale for her proposed principle, Kamm is nonetheless quite

upfront about the fact that she isn't trying to develop or defend this idea in any detail. She explicitly says that "it is not the point of this lecture to investigate these deeper possible meanings" of the principle, although she certainly recognizes that it is important to do so eventually (p. 69; cf. p. 60). So isn't it unreasonable for me to complain—as I am complaining—that she hasn't spent more time on this issue?

I certainly hope that my doing this is not unfair. For my worry is precisely that the deeper connections that Kamm claims to find are not really there. (Or at least, somewhat more cautiously, it is far from clear that they *are* there.) It is one thing to establish that the Principle of Permissible Harm can be given a plausible rationale and then postpone more careful investigation of that rationale for another occasion. It is quite another matter to rest content, as Kamm does, with a mere gesture in the direction of a potential rationale—if, as I believe, the supposed "deeper meanings" cannot truly be established at all. If the principle really does turn on distinctions that have no independent moral significance—and this is what I take to be the case—then we have reason to worry about it, regardless of how well it matches our intuitions. And what Kamm has to say on this matter is simply not sufficiently reassuring.

Of course, even if I am right that the principle fails to align with a morally relevant distinction (whether subordination versus substitution, or some other), we might still hope that there is some other candidate for Q that meets the double challenge of matching our intuitions about cases (at least, doing so closely enough) and aligning with some morally relevant distinction (whether subordination/substitution, or something else).

But I am prepared to suppose, if only for the sake of argument, that the Principle of Permissible Harm—or at least a suitably modified and even more complicated version of it—*is* the principle that (best) matches our intuitions. That is, I am prepared to believe that Kamm (or future Kammians) will have met the first of these two tasks. And I will readily admit that, given the bewildering array of

trolley cases and variants that have been proposed over the years, this is no small feat. It would be a significant achievement indeed to find Q. Nonetheless, if that is all that is successfully accomplished it is not nearly enough. For if the Principle of Permissible Harm is indeed Q, that just makes my skepticism about the second task all the more worrisome. For if the principle is Q, and yet that principle lacks a compelling rationale, then we should come to wonder whether an adequate moral theory will really accommodate quite so many of our intuitions about the various cases.

In short, I suspect that our intuitions about trolley problems respond to factors that simply do not have any genuine moral significance. We cannot give them a rationale. And if that is right, then we cannot fit them into a larger moral discourse that we should be prepared to embrace.

If I am right about this, then any genuine solution to the trolley problem will be one that is more reformist than we might initially have hoped. Perhaps, as Judith Thomson has come to believe,[5] we can retain deontology, but we will have to abandon the intuition that it is permissible to turn the trolley in the standard Trolley Case—thus abandoning the very intuition that got us started thinking about the trolley problem in the first place! Perhaps we will settle for a principle that (roughly speaking) simply rules out killing the innocent. If such a principle can be given a plausible rationale, this may accommodate enough of our intuitions to satisfy us, even if it does not match all of them.

Alternatively, however, and even more radically, perhaps we will need to move somewhat further afield. Taking seriously the thought that an adequate moral theory must involve an adequate rationale, perhaps we will be led to a principle that matches even fewer of our case specific intuitions but which nonetheless can be provided with a rationale that is significantly more compelling. More particularly, perhaps we will ultimately want to abandon deontology altogether and embrace consequentialism instead.

I believe that the latter course is indeed the right choice to make. But that, of course, is a discussion for another day.

Notes

1. This is indeed rough, and in her lectures Frances Kamm periodically revisits the question of how, exactly, the trolley problem should be delimited; but for my purposes there is no need to try to arrive at a more precise answer.

2. F. M. Kamm, *Intricate Ethics* (New York: Oxford University Press, 2007), pp. 186 n.78 and 188 n.89.

3. Eventually (see pp. 83–91), Kamm suggests that, strictly speaking, the Principle of Permissible Harm doesn't require that the event in question be the same as (or the noncausal flip side of) the saving of the *larger* number (the *greater* good). She suggests, rather, that it may even be permissible to kill *more* than are being saved, provided that this is done as the result of the noncausal flip side of the saving (rather than as a means to the saving). This is a fascinating and important suggestion, but it isn't essential to the point I want to discuss, so I will put it aside.

4. Kamm does have a go at this question in *Intricate Ethics*, pp. 165–167; but if I read that passage correctly, even she acknowledges that the explanation she offers there isn't adequate.

5. See Judith Jarvis Thomson, "Turning the Trolley," *Philosophy & Public Affairs* 36 (2008): 359–374, and her comments in this volume.