MUST CONSEQUENTIALISTS KILL?*

It is widely held that, in ordinary circumstances, you should not kill one stranger in order to save five. This applies to Organ Theft, in which you can provide life-saving transplants for five patients by killing an innocent stranger and using their organs. And it applies to Footbridge, in which you can stop a speeding trolley that will kill five strangers only by pushing a button that will drop a stranger off a bridge into its path; the person on the bridge would die, but the trolley would come to a halt.[1] There is dispute about how to formulate the principle that governs these cases, partly in light of Bystander, where it is allegedly permissible to switch the speeding trolley to another track, even though it will hit and kill an innocent stranger. But there is widespread agreement that this is the exception, not the rule.

If these claims are commonplace, it is even more widely held, first, that the only way to reconcile them with agent-neutral consequentialism is to insist that one killing is worse than five natural or accidental deaths; second, that there is no plausible way for consequentialists to deny that you should kill one stranger when that would prevent five others from being killed by someone else; third, that we should appeal instead to agent-centered restrictions, which give special weight to whether *you* kill anyone *now*.[2] Finally, it is often alleged that such restrictions are puzzling or paradoxical. If killing is so objectionable, and the objection has to do with the violation of victims' rights, shouldn't you prefer to minimize such violations?[3] What could explain your refusal to kill, even to prevent more killings, but a self-centered desire to keep your own hands clean?[4]

* For discussion of this material in earlier forms, I am grateful to Cian Dorr, Brendan de Kenessey, Caspar Hare, Richard Holton, Daniel Muñoz, Ryan Preston-Roedder, Quinn White, and participants in a seminar at MIT.

[1] The case is due to Judith Jarvis Thomson, "Killing, Letting Die, and the Trolley Problem," *The Monist*, LIX, 2 (April 1976): 204–17, drawing on Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect," *Oxford Review*, V (1967): 5–15.

[2] Representative expressions of orthodoxy include Samuel Scheffler, *The Rejection of Consequentialism* (New York: Oxford University Press, 1982); Shelly Kagan, *The Limits of Morality* (New York: Oxford University Press, 1989); and Judith Jarvis Thomson, *The Realm of Rights* (Cambridge, MA: Harvard University Press, 1990).

[3] On the puzzle here, see Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), chapter 3; and Scheffler, *The Rejection of Consequentialism*, *op. cit.*, chapter 4.

[4] For an especially clear account of this objection, see Caspar Hare, *The Limits of Kindness* (Oxford: Oxford University Press 2013), chapter 6.

0022-362X/18/1502/092–105

Despite believing that, in ordinary circumstances, you should not kill one stranger in order to save five, even from being killed by someone else, I reject this picture wholesale. As I will argue, the best account of reasons not to kill is both consistent with consequentialism and thoroughly agent-neutral. It does not imply that killings are substantially worse than natural or accidental deaths. And it agrees that you should not kill one in order to prevent five others from being killed by someone else. Because it is agent-neutral, this account avoids the paradox of agent-centered restrictions and the problem of clean hands. But it is subject to a paradox of its own. In the final section, I develop a new puzzle in the ethics of killing and saving lives, a puzzle that does not presuppose consequentialism or the reverse.

I. CONSEQUENTIALISM

In Samuel Scheffler's canonical formulation, act-consequentialists "specify some principle for ranking overall states of affairs from best to worst from an impersonal point of view. . . .After giving some principle for generating such rankings, act-consequentialists then require that each agent in all cases act in such a way as to produce the highest-ranked state of affairs that he is in a position to produce."[5] In short, among the actions available to you, you should perform an action whose consequences are best.

This raises several questions. First, why '*an* action', not '*the* action'? Because there might be several actions whose consequences are best: you should perform one of them. Second, what is meant by 'consequences'? Reply: 'consequences' must be understood inclusively. They consist in how the world would be if you performed an action, including the performance of the action itself, its relation to the past and present, and so on. Third, what if there is no fact of the matter about what would happen if you performed a given action? Perhaps the future is objectively indeterminate or chancy? Reply: that would call for risk-weighted values, a complication I will ignore. I will take a similar approach to ignorance and uncertainty. The formulation above is "objective" in that it looks to the consequences an action would have, regardless of whether you know about them or what your epistemic position might be. More plausibly, what you should do depends on what you know, what evidence you have, or what you believe. This introduces agent-relativity, but not the sort that interests us. In what follows, I will assume that the agent knows the consequences of the available actions, so as to avoid this distraction.

[5] Scheffler, *The Rejection of Consequentialism, op. cit.*, p. 1.

A further question is more difficult. What is meant by 'best consequences' or 'best from an impersonal point of view'? A number of influential philosophers have cast doubt on the intelligibility of 'good' as it is used above, finding a conceptual vacuum in the foundations of consequentialism.[6] We know what it means for something to be good for someone, or good for doing something, or good as an instance of a functional kind. We should not assume that we know what is meant by 'good consequence' or 'good state of affairs', which do not take any of these forms. Without endorsing such skepticism, I think it makes sense to avoid using 'good' and 'best' in stating consequentialism, introducing those terms only when their meaning has been defined. Accordingly, I will formulate consequentialism as the conjunction of two theses.

> Action-Preference Nexus: Among the actions available to you, you should perform one of those whose consequences you should prefer to all the rest.
>
> Agent-Neutrality: Which consequences you should prefer is fixed by descriptions of consequences that make no indexical reference to you.

The Action-Preference Nexus captures the act-consequentialist idea that the evaluation of actions (what you should or should not do) cannot come apart from the evaluation of their consequences (what you should or should not prefer to be the case). This is not a claim about the order of explanation but about the congruence of reasons. It is consistent with agent-relative consequentialism, on which reasons for preferring a given consequence may depend on what you do, or what happens to you, identified as such, in the relevant world.[7] It is a matter of terminological dispute whether such views deserve the label 'consequentialist'. Without entering that dispute, I will focus on forms of consequentialism that accept Agent-Neutrality. Indexical reference, here, is reference that identifies you as such. According to Agent-Neutrality, information of this kind is irrelevant to what you should prefer. It makes no difference whether the person who benefits, suffers, kills, or saves someone's life is you or a stranger. Having got this far, we can understand 'better' and 'worse' in terms of what you should

[6] See Philippa Foot, "Utilitarianism and the Virtues," reprinted in *Moral Dilemmas: and Other Topics in Moral Philosophy* (Oxford: Oxford University Press, 2002), pp. 59–77; and Judith Jarvis Thomson, "Goodness and Utilitarianism," *Proceedings and Addresses of the American Philosophical Association*, LXVII, 2 (October 1993): 145–59; both citing Peter Geach, "Good and Evil," *Analysis*, XVII, 2 (December 1956): 33–42.

[7] For this approach, see James Dreier, "Structures of Normative Theories," *The Monist*, LXXVI, 1 (January 1993): 22–40.

prefer, or disprefer, given descriptions of consequences that do not identify you as such.

Finally, I have stated consequentialism as a view about what you should do and what you should want. Some philosophers regard morality as a distinctive normative field, distinguishing what is morally wrong from what you should not do, all things considered, and what is right from what you should. They may interpret consequentialism as a view specifically addressed to moral right and wrong. I am wary of these concepts, with their allegedly distinctive content, so I do not employ them here. Others may substitute accordingly.

II. DEONTOLOGY

Scheffler goes on to define deontology as the belief in agent-centered restrictions:

> An agent-centred restriction is a restriction which it is at least sometimes impermissible to violate in circumstances where a violation would prevent either more numerous violations, of no less weight from an impersonal point of view, of the very same restriction, or other events at least as objectionable, and would have no other morally relevant consequences.[8]

As Scheffler observes, if there are agent-centered restrictions, "there is no non-agent-relative principle for ranking overall states of affairs from best to worst such that it will always be permissible to produce the best state of affairs so characterized."[9] In our terms, agent-centered restrictions involve reasons not to perform certain types of actions, even to prevent more actions of that type, that violate either the Action-Preference Nexus or Agent-Neutrality.

It will be essential for us to distinguish two kinds of agent-centered restrictions. On the one hand, there are restrictions that govern our treatment of everyone. These are "general restrictions." On the other hand, there are restrictions that govern our treatment only of some individuals, those with whom we have a special relationship. The relationship might be enduring and largely voluntary, like friendship, involuntary, like being someone's child, or transient, like having made an agreement or promise. Restrictions that depend on selective relationships are "special."

In my view, special restrictions conflict with Agent-Neutrality. Suppose I learn that, in the future, one of two things will happen. In My Neglect, I neglect my children and they suffer terribly; everyone else is a responsible parent. In His Neglect, I am a responsible parent;

[8] Scheffler, *The Rejection of Consequentialism*, *op. cit.*, p. 80.
[9] *Ibid.*

independently, a stranger neglects his children, who suffer terribly; everyone else is a responsible parent. If there are special restrictions that govern our treatment of our own children, I should not only give priority to my responsibilities over those of other people, I should prefer His Neglect to My Neglect, other things being equal. However, when these outcomes are described without indexical reference to me, they are indistinguishable, and it is not the case that I should prefer one to the other. This violates Agent-Neutrality: which consequences I should prefer is not fixed by descriptions of consequences that make no indexical reference to me. The same point holds when I compare My Broken Promise, in which I break a promise but someone else keeps theirs, with His Broken Promise, in which I keep mine but a stranger breaks his. I should be more concerned with whether I keep my promises than with promises made by strangers.

Since I believe in special restrictions, I reject Agent-Neutrality. For the sake of this paper, however, I want to set this fact aside. Our topic will be general restrictions, and in particular, restrictions on killing in order to save lives or in order to prevent killings. Do these restrictions conflict with Agent-Neutrality? Many have thought that they do, that restrictions against killing the innocent are agent-relative and therefore inconsistent with agent-neutral consequentialism.

It is difficult to formulate such putative restrictions properly. Do they appeal to the contrast between killing and letting die, between intending and foreseeing harm, to using as a means, redirecting threats, some combination of these, or something else entirely? In order to avoid these controversies, I will focus on particular cases. One we have met already. In Footbridge, you can stop a speeding trolley that will kill five strangers only by pushing a button that will drop a stranger off a bridge into its path; the person on the bridge would die, but the trolley would come to a halt. Murderous Footbridge is similar, except that a villain directed the trolley at the five, intending to kill them.

In both cases, I assume, you should not push the button.[10] We can account for this in Footbridge by insisting that killings are worse than accidental deaths, 'worse' being understood in terms of what you should disprefer, given descriptions of consequences that do not identify you as such. That is consistent with Agent-Neutrality and the Action-Preference Nexus. But whether or not this move is plausible, nothing like it can apply to Murderous Footbridge, which sets killings against killings. On the face of it, what matters here is whether it is the

[10] This view is orthodox; see Thomson, *The Realm of Rights*, *op. cit.*, pp. 137–41.

villain who does the killing or you. That looks like agent-relativity. But appearances can deceive.

III. AGENT-NEUTRALITY

The argument for agent-relativity in Murderous Footbridge misapplies the test for Agent-Neutrality. The test is this: does what you should prefer shift when a description identifies you as such? To apply this test correctly, as we did with My Neglect, we need to compare the original description of the case—Murderous Footbridge—with one that omits identifying information. In Someone's Murderous Footbridge, a villain has directed a trolley at five innocent people, who will die if it hits them. The only way to stop the trolley is to push a button that will drop another person off a bridge into its path; the person on the bridge would die, but the trolley would come to a halt. Someone stands by the button, knowing these facts about the case, deliberating. What should you want them to do? You should want the person to refrain from pushing the button, just as you should refrain from pushing it in Murderous Footbridge. Given the Action-Preference Nexus, you should prefer that you refrain in Murderous Footbridge. As we see in Someone's Murderous Footbridge, dropping the indexical reference to you does not affect what you should prefer. This is consistent with Agent-Neutrality.[11]

The point is not specific to Murderous Footbridge. In general, when you should not cause harm to one in a way that will benefit others, you should not want others to do so either. This fact is sometimes recognized, or near enough. Thus Philippa Foot writes: "In the abstract, a benevolent person must wish that loss and harm be minimized. He does not, however, wish that the whole consisting of a killing to minimize killings should be actualized either by his agency or that of anyone else."[12] This falls just short of saying that he should wish the whole in question *not* to be actualized, but it suggests as much. Foot also writes: "[It] is not true that a given act is worse when done by

[11] There are many variations on this case, some of which may generate uncertainty. Suppose, for instance, that the person at the button is the villain who directed the trolley at the five. Should he push the button? (The standard answer is no; see Thomson, *The Realm of Rights*, *op. cit.*, p. 139, describing a doctor who has given five patients a poison that will cause organ failure and can save them by harvesting the organs of a sixth.) Should you want him to? What if the person at the button is a would-be murderer, moved purely by malice toward the person on the bridge? He has no idea about the threat to the five or is indifferent to it. (This brings us closer to the comparsion of One Killing and Five Killings, discussed below.) I will ignore these complications. It is sufficient for the point about Agent-Neutrality, and the puzzle in section IV, that in the simplest version of Someone's Murderous Footbridge, where the person at the button is an innocent, well-informed, well-meaning stranger, you should want them to refrain.

[12] Foot, "Utilitarianism and the Virtues," *op. cit.*, p. 73.

oneself than when done by another, unless of course there is some relevant difference between us, as when only one of us is a doctor or a parent or a friend."[13] I should care about what happens to the potential victims of murder, not whether I am the one who does the killing.[14]

That the appropriate attitudes to killing in Murderous Footbridge may conform to Agent-Neutrality has been observed before. In an excellent essay, Tom Dougherty writes:

> A deontologist is free to say that [a bystander to Murderous Footbridge] should be opposed to your killing the single person, even though she knows that this will lead to more deaths overall. . . .Similarly, the deontologist can say that the bystander ought to prefer that you do not kill. Indeed, I suggest that these are rather attractive claims for the deontologist to make.[15]

Despite our pivotal agreement, I differ from Dougherty on three points. The first is terminological: I use 'deontology' for views that involve agent-centered restrictions, making Dougherty's 'agent-neutral deontology' an oxymoron. Second, I think it is vital to distinguish general restrictions, which lend themselves to agent-neutral treatment, from special restrictions, which do not. Dougherty does not track this distinction. Third, Dougherty sees a substantive contrast between the view that everyone should prefer the person to refrain from pushing the button in Someone's Murderous Footbridge (in the quotation above, he calls this an attractive claim) and the view that it is "impersonally worse" if they push the button, of which he writes: "I do not think such a theory is correct, but neither do I think it is off the wall."[16] Unlike Dougherty, I see no meaningful difference between

[13] Philippa Foot, "Morality, Action, and Outcome," reprinted in *Moral Dilemmas*, *op. cit.*, pp. 88–104, at p. 94.

[14] We can verify this claim against another comparison, by which it is easy to be misled. Suppose I learn that, in the future, one of two things will happen. In My Killing, I kill an innocent stranger; everything else goes well. In His Killing, someone else kills an innocent stranger; everything else goes well. When these outcomes are described without indexical reference to me, they are indistinguishable, and it is not the case that I should prefer either one. Does that change when I am identified as such? No. While I may prefer His Killing to My Killing, and this may well be rational, it reflects an agent-centered prerogative to care more about my own life than his, not an agent-centered restriction on killing people myself. Once we abstract from the effects of killing someone myself—for instance, in leading me to violate special restrictions, as when I am imprisoned and thus unable to care for my children—we can see that, while it is rational to prefer His Killing, I do not have decisive reason to do so. It would be rational to be indifferent here, as I am rationally indifferent when the outcomes are described without indexical reference to me.

[15] Tom Dougherty, "Agent-Neutral Deontology," *Philosophical Studies*, CLXIII, 2 (March 2013): 527–37, at pp. 530–31.

[16] *Ibid.*, p. 533, n. 17.

these views. To say that you should prefer that others refrain from pushing the button in Someone's Murderous Footbridge is to say that it is impersonally worse if they do.

In addition to these points of contrast, there are questions Dougherty does not address. I will emphasize two of them. One is about the value of killings and accidental deaths, which I take up in this section. This discussion sets the stage for a puzzle about killing and saving lives, which I develop in the next.

Think back to Footbridge, in which you must choose between killing one stranger to save five or allowing five accidental deaths. You should not push the button that drops the stranger off the bridge into the path of the speeding trolley. Given the Action-Preference Nexus, you should also prefer that this not take place. As in Murderous Footbridge, Agent-Neutrality holds. A bystander should equally prefer that you not push the button, and you should prefer that someone else refrain from pushing the button if they find themselves in the Footbridge case. Scheffler notes the possibility of such a view but claims that there is only one way for consequentialists to defend it: "they can suggest that one killing is a worse thing to happen than five deaths caused by accident or disease, and hence that killing the innocent person is prohibited because it actually produces the worse overall outcome."[17] That would get the results we want in Footbridge without agent-relativity. But it "requires a highly implausible account of the good. For a killing is not a worse thing to happen than *one* otherwise equally undesirable death, let alone a worse thing than five such deaths."[18]

Scheffler's picture is misleading. The claim that everyone should prefer that you not push the button in Footbridge does not require an implausible theory of the good on which killings are worse than accidental deaths. This comes out clearly when we translate from 'better' and 'worse' into what you should prefer. Suppose you learn that, in the future, one of two things will happen. In One Killing, there is a random murder. In Five Accidental Deaths, five strangers are killed by a runaway trolley. You should prefer One Killing. In that sense, one killing is not a worse thing to happen than five accidental deaths. Scheffler is right about that. But this is consistent with what was said about Footbridge above. There you compare quite different outcomes. In Five Accidental Deaths, five strangers are killed by a runaway trolley. In Killing One to Save Five, someone pushes a button that drops another stranger off a bridge in front of the trolley, killing them

[17] Scheffler, *The Rejection of Consequentialism*, *op. cit.*, p. 108.
[18] *Ibid.*, p. 109.

in a way that brings the trolley to a halt. You should prefer Five Accidental Deaths. One Killing is simply a different scenario than Killing One to Save Five.

In much the same way, what was said about Murderous Footbridge is consistent with the claim that you should generally prefer that there be fewer killings. Suppose you learn that, in the future, one of two things will happen. In One Killing, there is a random murder. In Five Killings, there are five. Of course you should prefer One Killing. This is simply different from One Killing to Prevent Five, as when the button is pushed in Someone's Murderous Footbridge. We must compare like with like. (What happens when we do? Suppose you are watching five instances of Murderous Footbridge unfold in front of you. You should prefer that no one push their button, and you should prefer that one person push instead of five.)

The moral is that, if we simply ask whether killings are worse than accidental deaths, or fewer killings better than more, our questions are unhelpfully coarse-grained. Should you prefer accidental deaths to killings, other things being equal? No, or not by much. Certainly, you should prefer a single killing to five accidental deaths. But should you always prefer this, no matter how the killing relates to the deaths? No. You should not prefer Killing One to Save Five to Five Accidental Deaths; in fact, the reverse. Similarly, you should prefer that there be fewer killings, other things being equal. But you should not prefer this when the one is killed in order to prevent the killing of the five, as in Murderous Footbridge.

IV. TRANSITIVITY

The idea that you should not push the button in Murderous Footbridge has proved contentious in part because it has been thought to depend on agent-centered restrictions, which philosophers have found puzzling or paradoxical.[19] If what I have argued is right, this dispute is beside the point. Paradoxical or not, agent-centered restrictions are not involved in the commonsense verdict that, in cases like Murderous Footbridge, you should not kill one in order to prevent five others from being killed. Nor does this verdict rest on a selfish desire to keep your own hands clean, refusing to kill even to prevent more killings. Since it is consistent with Agent-Neutrality, there is nothing self-centered about it. You should want others to act in just the same way.

[19] Nozick, *Anarchy, State, and Utopia*, *op. cit.*, chapter 3; Scheffler, *The Rejection of Consequentialism*, *op. cit.*, chapter 4.

The effect of focusing on agent-centered restrictions has been to divert attention from a deeper puzzle in the ethics of killing, which is brought into view by the arguments of section III. We can state the puzzle by appealing to the transitivity of 'better than', or in our terms, the transitivity of 'should prefer'.[20] Using '>' for this relation, we can summarize our results so far. To begin with, killings are not much worse than accidental deaths. In particular:

One Killing > Five Accidental Deaths.

We can imagine that the five are killed by a runaway trolley that could only have been stopped by pushing a button that will drop a stranger off a bridge into its path; the person on the bridge would die, but the trolley would come to a halt. In other words, the five deaths occur in a Footbridge scenario, but one in which the person at the button declines to push. Let us build that into Five Accidental Deaths. Turning next to Footbridge, we can add:

Five Accidental Deaths > Killing One to Save Five.

By transitivity:

One Killing > Killing One to Save Five.

Can this be right? If you learn that the future holds one or the other, should you really prefer that it hold a random murder than someone pushing the button in Footbridge, killing one but saving five?

We can raise a similar question in Murderous Footbridge, imagining that Five Killings involves a Murderous Footbridge scenario, but one in which the person at the button declines to push. We then have two results:

One Killing > Five Killings

and

Five Killings > One Killing to Prevent Five.

By transitivity:

[20] Transitivity has been questioned in recent work—see Larry Temkin, "Intransitivity and the Mere Addition Paradox," *Philosophy and Public Affairs*, XVI, 2 (Spring 1987): 138–87; Larry Temkin, "A Continuum Argument for Intransitivity," *Philosophy and Public Affairs*, XXV, 3 (Summer 1996): 175–210; Stuart Rachels, "Counterexamples to the Transitivity of Better Than," *Australasian Journal of Philosophy*, LXXVI, 1 (March 1998): 71–83; Alexander Friedman, "Intransitive Ethics," *Journal of Moral Philosophy*, VI, 3 (2009): 277–97; and Timothy Willenken, "Deontic Cycling and the Structure of Commonsense Morality," *Ethics*, CXXII, 3 (April 2012): 545–61—but the questions raised elsewhere are largely unrelated to the ones at issue here.

One Killing > One Killing to Prevent Five.

Can this be right? If you learn that the future holds one or the other, should you really prefer that it hold a random murder than someone pushing the button in Murderous Footbridge, killing one to prevent five killings?

In fact, the situation is more extreme. If you should prefer one killing to five accidental deaths, shouldn't you prefer two? And shouldn't you prefer two killings to five? In other words:

Two Killings > Five Accidental Deaths

and

Two Killings > Five Killings.

By transitivity, and the claims above, we can derive:

Two Killings > Killing One to Save Five

and

Two Killings > One Killing to Prevent Five.

If you learn that the future holds one or the other, should you really prefer *two* random murders to someone pushing the button in a Footbridge case, or in Someone's Murderous Footbridge, even though there is less killing if they do?

This is the puzzle promised above. From sensible verdicts on the cases framed in section III, we have derived a seemingly bizarre conclusion, that you should prefer two killings to a single killing that prevents five deaths. The implication is peculiar enough to elicit second thoughts. Did we somehow go astray? Perhaps our mistake was to insist on Agent-Neutrality. Perhaps we have discovered why the ethics of killing must appeal to agent-centered restrictions. But this is not an accurate description of what we have learned. Far from insisting on Agent-Neutrality as a theoretical constraint, I rejected it in section II, accepting special restrictions. I merely observed that Agent-Neutrality is consistent with the verdict that you should not push the button in Footbridge or Murderous Footbridge, given the fact, or what I take to be the fact, that a bystander should prefer that you not push the button in either case, and the fact that you should prefer Five Accidental Deaths to Killing One to Save Five and Five Killings to One Killing to Prevent Five. It is these claims about cases that create our puzzle, bottom-up, not theoretical structure imposed top-down.

The upshot is perplexing. We have four options. First, we could retract the initial claim that you should not push the button in

Footbridge or Murderous Footbridge. Second, we could drop the further claims about bystanders and third parties, what they should prefer and what you should prefer they do. Either way, we take a revisionary approach to the ethics of killing. These strike me as options of last resort.

Third, we could dispute the application of transitivity, the derivation of the troubling comparisons,

Two Killings > Killing One to Save Five

and

Two Killings > One Killing to Prevent Five.

The obvious way to do this is to claim that what you should prefer is sensitive to the outcomes being compared. That you should prefer *A* to *B* when you compare *A* and *B*, and *B* to *C* when you compare *B* and *C*, does not entail that you should prefer *A* to *C* when you compare *A* and *C*. 'Should prefer' is transitive only in relation to a fixed comparison class. In our case: when you compare Five Killings with Two Killings, you should prefer the latter; when you compare Five Killings with One Killing to Prevent Five, you should prefer the former. Nothing follows about what you should prefer when you compare Two Killings with One Killing to Prevent Five.

This way of resisting the application of transitivity may seem to have a precedent. According to Frances Kamm's Principle of Secondary Permissibility, permissibility is option-relative.[21] Suppose it would be wrong to push a stranger in front of a trolley, crushing his leg but stopping its progress, when it will otherwise kill five, but permissible to push the button in Bystander, switching the trolley to a side track where it will kill an innocent stranger. On Kamm's principle, if it is the same stranger and both options are available, it might then be permissible to take the first.

Whatever we make of this idea, the extension to our case is problematic. With the permissibility of action, options reflect one's circumstance in ways that may be ethically significant: what an agent can do affects how her actions relate her to others and thus, perhaps, what she is permitted to do. With preference, however, the question of what you can do does not arise: we are asking what you should prefer to happen, not what you should do. Why should it matter which outcomes you consider, if not because they are options? What you consider is a psychological fact about you, not a feature of your

[21] Frances Kamm, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (New York: Oxford University Press, 2007), pp. 169–71.

circumstance, or the object of preference, that might affect what you should prefer.

Even if preference *is* sensitive to the outcomes being compared, we can ask what you should prefer when you compare all three. The idea must be that, presented with this comparison, your preferences should shift. But why? It is one thing to contemplate in-principle violations of the independence of irrelevant alternatives, the idea that, if you (should) prefer *A* to *B* when you compare the two of them, you should prefer *A* to *B* when you compare *A*, *B*, and *C*. It is another to explain why it is violated in this case. Having no idea how to do that, I turn instead to option four.

Our final option is to accept the initially surprising claim that you should prefer Two Killings to Killing One to Save Five and One Killing to Prevent Five. Here is a way to think about these comparisons, focusing on the latter. Instead of asking what transpires at the end, the final body count, we evaluate outcomes by imagining them as they unfold in time. At a certain point in Two Killings, two lives are threatened. That is pretty bad. At a similar point in Five Killings and One Killing to Prevent Five, five lives are threatened. That is much worse. In Five Killings, things proceed as threatened. The resulting outcome is worse than Two Killings. Now consider One Killing to Prevent Five, in which things are going as badly as in Five Killings—which is to say, worse than Two Killings—up to the point at which the button is pushed. Do things improve at that point? No, they get worse still. Someone adds insult to injury by pushing the button that drops the stranger off the bridge. One Killing to Prevent Five starts out the same as Five Killings and then declines; that is why it is worse than Two Killings.

This description can be refined. The problem in One Killing to Prevent Five, before the button is pushed, is not simply that five lives are threatened. It is that the only way to save the five is by killing an innocent stranger. At the beginning of Five Killings, five people are going to be killed. In One Killing to Prevent Five, five people are going to be killed unless they are saved by the pushing of the button, which kills an innocent stranger by dropping him off a bridge into the path of the speeding trolley. The situation in which someone is going to be killed unless they are saved in this way is as bad as the situation in which they are going to be killed. Ethically speaking, the damage has been done. (Things may be different in Bystander, where the button switches tracks instead of dropping someone from a bridge.) It makes things worse, not better, that the button is pushed, so that the innocent stranger dies. That is why One Killing to Prevent Five is worse than Five Killings: it starts out the same and then declines. If we think through

the temporal unfolding of events in One Killing to Prevent Five, we can explain why Five Killings, and thus Two Killings, should be preferred.

This way of framing the comparison will not convince everyone. It may not convince you. But then you are forced to choose: accept that it is rational to kill in Murderous Footbridge, deny that others should prefer that you refrain from killing and that you should prefer that others refrain, or question the transitivity of 'should prefer'.

KIERAN SETIYA
Massachusetts Institute of Technology