Action Individuation and Deontic Cycling*

Stephen C. Makin

Tim Willenken argues that ‘commonsense morality’ is committed to intransitive deontic cycles; that consequentialism cannot countenance such cycles; and that, therefore, the project of compatibilism—making consequentialism and commonsense morality deliver the same moral verdicts, by way of an axiology—cannot succeed. I argue that the appearance of intransitive cycles is made possible only by an idiosyncratic method of action-individuation; when traditional methods are used, the appearance of intransitivity goes away. These results may reopen the door for the compatibilist project.

I

Tim Willenken argues that ‘commonsense morality’—a kind of prosaic deontological view—is committed to intransitive deontic cycles;¹ that consequentialism cannot countenance such cycles; and that, therefore, the project of compatibilism—making consequentialism and commonsense morality deliver the same moral verdicts, by way of an axiology—is dead in the water.²

I have my own suspicions about the compatibilist project; but my point in what follows will be chiefly to defend commonsense morality against the charge that it is committed to intransitive cycles. One upshot, however, will be that, if there is a problem with compatibilism, it will have to be found elsewhere.

Willenken’s argument turns on a trio of trolley scenarios, and (supposedly) a trio of actions (x, y, and z); his thought is that commonsense morality...

* I should like to thank Tim Willenken and two anonymous referees for their helpful comments.


2. The idea, roughly, is that we deploy a relativized assignment of values—an axiology—according to which one dead-by-me is worse, relative to me, than, e.g., one dead.
morality will endorse choosing action y over action x, z over y, and x over z, thus showing that our commonsense morality is committed to a fishy sort of ‘intransitive deontic cycle’. The three scenarios are these:

**Scenario 1:** There are two runaway trolleys: one en route to crush two people (Bob and Charlie) and the other en route to crush five (different) people. I can stop just one of these two trolleys; so I can either:

- (x) Let five people die (by getting crushed by a trolley); or
- (y) Let two people, Bob and Charlie, die (by trolley crushing).

According to familiar (if contested)\(^3\) moral principles, I should save the greater number, and therefore choose y. Next we have

**Scenario 2:** A single trolley on course to kill both Bob and Charlie; somehow or other, it transpires that I can push Bob (who will die anyway) into the course of the trolley, killing him, but sparing Charlie—or do nothing. That is, I can either:

- (y) Let two people, Bob and Charlie, die (by trolley); or
- (z) Kill Bob (by pushing him off a bridge), thus blocking the trolley from further harm.

And though not everyone would agree, many of us would say that z is permissible, perhaps even required: since Bob will die anyway, I should minimize the number of deaths, even if that will involve, unfortunately, killing Bob. Finally there is

**Scenario 3:** A trolley is again loose, this time heading for five innocents; Bob is again on a bridge—though this time out of harm’s way—from which I can push him into the path of the trolley, thereby saving the five—or I can do nothing. My choices, in other words, are:

- (z) Kill Bob (by pushing him off a bridge); or
- (x) Let five people die (by getting crushed by a trolley).

It is widely believed that I must choose x in this sort of situation. So x is better than z is better than y is better than x: an intransitive cycle. Or is it?

II

My complaint, in short, is that the relevant “actions”—x, y, and z—are fatally underdescribed; once they are fully described, the appearance of

intransitivity goes away. The underdescription is possible because of a quirk in the setup of the problem. The question will be: how do we individuate the actions to be evaluated? I come to that in Section III, below. First, though, it is useful to try to bring out how the appearance of intransitivity is produced in Willenken’s setup.

The immediate point is that the description “let two people die” is not sufficient to individuate an action. If it were, then the action

\[(j)\] “let two die, in order to avoid scratching my little finger”

would come out as the same action as

\[(k)\] “let two die in order to save five,”
as, after all, they are both instances of letting two die. But I think that it is clear that since \(k\) is, under any interpretation, a case of saving five, it is not the same action as \(j\), which is not so describable.

There is no intuitive ambiguity, however, in the descriptions of, for example, the actions labeled ‘\(y\)’ in scenarios 1 and 2, above: why not? This is a product, I think, of the setup: since we face a forced choice here between two exclusive options, it is possible to implicitly encode part of the description of some action (e.g., \(y\)) in the description of its alternative: since the only alternative, in Scenario 1, to letting two die, is letting five die, it is implicitly specified that letting two die is, somehow or other, also the saving of five. But once we state this explicitly as part of the content of the action \(y\), the ‘action’ loses its univocal character: in Scenario 1, \(y\) is at least a case of saving five; whereas the \(y\) in Scenario 2 is not the saving of anyone. They do not seem to be the same action at all; and this is enough by itself to upset the intransitivity charge.

III

Even if it is already clear that ‘\(y\)’ does not name the same action in Scenarios 1 and 2, the question remains just what belongs in a proper action description—that is, by what method we ought to be individuating actions. Now there may be more than one way for a deontologist to individuate actions; but, traditionally, she distinguishes actions by the intention with which they are done; or, as we could put it, by the “in order to” clause that will appear in them.\(^4\) Now she should also allow, I think, that some further

\[4.\] Aquinas, *Summa Theologica*, trans. Fathers of the Dominican Province (Wheaton, IL: Christian Classics, 1981), II-II, 64, 7: “Now moral acts take their species according to what is intended, and not according to what is beside the intention, since this is accidental as explained above.” This is also the leading theme of G. E. M. Anscombe’s monograph *Intention* (Cambridge, MA: Harvard University Press, 1963). As far as what I say in the text goes, then, it is an open and perhaps important question whether alternative action-
consequences—in the sense of: those things which will, as the agent foresees, come about should she choose some action—will also be relevant, if not to the identity of the action, then anyway to its moral assessment.\(^5\)

We have already seen that action \(y\) is not univocal. But how would a deontologist distinguish them? It would not be unnatural, I think, to describe the first of these (\(y_1\)) as

\[(y_1)\]  Stop a trolley in order to save five—and thereby allow two to die.

Whereas \(y_2\) would be naturally described as

\[(y_2)\]  Refuse to kill Bob—thereby letting two (Bob and Charlie) die.

These two are again clearly not the same action. Even if we allowed \(y_2\) to be written merely as “do nothing, and thereby allow two to die” or, more simply (and doubtfully) still, “let two die,” it remains clear that we are dealing here with two very different actions. In fact—although it is not presently to the point—the consequences of these actions are not even the same: under some plausible notion of the default-future, \(y_1\) results in \(+3\) lives, whereas \(y_2\) results in a very bad \(-1\). (I shall return to this.)

This point by itself is enough to upset the intransitivity charge; but, emboldened, we might go on to rewrite the whole thing:

\[(x_1)\]  Stop a trolley in order to save two—thereby allowing five to die.

\[(y_1)\]  Stop a trolley in order to save five—thereby allowing two to die.

\[(y_2)\]  [Refuse to kill Bob and thereby] let two (Bob and Charlie) die.

\[(z_1)\]  Push Bob in order to save Charlie (where Bob would die anyway).

\[(z_2)\]  Push Bob in order to save five.\(^6\)

\[(x_2)\]  [Refuse to kill Bob, i.e., do nothing, and thereby] let five die.

Whether or not we admit the bracketed phrases, it is quite clear that \(x_1 \neq x_2\), \(y_1 \neq y_2\), and \(z_1 \neq z_2\). So there is no intransitivity.

\(^5\) Are these consequences essential to the identity of the action? Whether or not this is so is central to the controversy surrounding the Doctrine of Double Effect. I do not try to settle the question here, but the main point does not hang on it: the intransitivity claim is destroyed as soon as we include the intention, and so does not require the inclusion of further consequences. I include them in my descriptions since the alternative, I take it, would be the more controversial.

\(^6\) I have written the ‘\(z\)’s as “Push Bob” rather than “Kill Bob” in deference to one, popular reading of the Doctrine of Double Effect: one is not aiming at Bob’s death in
If this is right, then commonsense morality is not committed to intransitive cycles after all—at least, by its own lights. It may be worth stressing, though, that even a consequentialist, with his metaphysics of action, need not regard this cycle as intransitive. With the right axiology in hand, the consequentialist will also dispute the equivalence of the action variants (see n. 2 on axiologies).

Now in the first place, it is not clear that a consequentialist can help himself to the action-individuation principle employed by deontologists—the “in order to” clause. Fortunately for him, he does not need to. He can use, instead, his time-honored technique of individuating actions by the consequences they produce. Using some fanciful numbers to represent outcome values, things look like this:7

\[ (x) \quad \text{Two saved } (+2,000) \text{ and five dead } (-5,000) = -3,000. \]

\[ (y) \quad \text{Five saved } (+5,000) \text{ and two dead } (-2,000) = +3,000. \]

\[ (z) \quad \text{Two dead } (-2,000) = -2,000. \]

\[ (y') \quad \text{One dead-by-me-but-doomed } (-0) \text{ and one saved } (+1,000) = +1,000. \]

\[ (z') \quad \text{One dead-by-me } (-100,000) \text{ and five saved } (+5,000) = -95,000. \]

\[ (x') \quad \text{Five dead } (-5,000) = -5,000. \]

It appears that, on one familiar notion of ’consequences’, \( x \) does not produce the same consequences as \( x' \) and hence cannot be the same action; and likewise for the \( y \)'s and \( z \)'s. This math gives us the original set of priorities, but does not produce any intransitivity. The fancifulness of the numbers is of course neither here nor there, since the whole point of the

---

7. Each life lost or saved is worth 1,000 (absolute) units; dead-by-me’s are worth -100,000. This is not, as I note in the text, the only way a consequentialist might assign values to actions. He might, for example, simply join us to maximize lives saved (where this doesn’t require dead-by-me’s). He would not then need to assign any positive weight to savings as such. For the matter of producing the right relative values of actions, this is fine. But including savings as part of the consequence of any action is, I think, the only sensible way that a consequentialist can individuate actions: if savings are not counted as essential parts of the action, then it would turn out that, for example, doing nothing (when nothing is at stake) and saving 1,000 lives (when a trolley is out of control) would count as “the same action”; and it would be trivial to produce intransitivities. (I thank an anonymous referee for pressing this point.)
axiological strategy, as I understand it, was to develop gimmicks which would show that commonsense morality and consequentialism are in principle compatible.

V

Now Willenken allows that compatibilism can be saved if we are allowed to “individuate actions by the available alternatives”—what he calls fine-graining; but he argues that such a move will produce nonsensical action-directives in the case of three-way choices. For that reason, we should ask whether the proposal just given is covertly an instance of fine-graining.

The idea is supposed to be this: if we can distinguish $z$ (when the alternative is $y$) from $z$ (when the alternative is $x$), then we can indeed evade intransitive cycles.$^8$ But problems are supposed to be raised when we transport this strategy over to three-way choices. Consider the following scenario:

Two trolleys are out of control: one is en route to crush five people, and the other is en route to crush Bob and Charlie. I may press exactly one of three buttons. I may press button $x$ and stop the trolley en route to killing Bob and Charlie (letting five die), press button $y$ and stop the trolley en route to crushing five people (letting Bob and Charlie die), or press button $z$, which will push Bob off of a bridge he is standing on (it is part of the track) onto an earlier portion of the track where his body will stop both trolleys before they crush anyone else.$^9$

Willenken is as baffled as I am about what the correct choice here is. But he suggests that it may be the third option: we have stronger reason to push button $z$ (and hence push Bob) than we do either to push button $x$ or button $y$. If that is right, however, then we get the following claim (among others) about what we have stronger reason to do:$^{10}$

$$z \text{ (when the alternatives are } x \text{ and } y) > x \text{ (when the alternatives are } y \text{ and } z)$$

and this is a nonsensical evaluation: it is not metaphysically possible, Willenken argues, to face such a pair-wise choice.$^{11}$ The general problem results from the attempt to evaluate two actions against each other, each of which was individuated in multiway choice scenarios.

8. The reasoning is parallel to, e.g., that in Sec. II above: once $z$ loses its univocal character, the intransitivity charge fails; pushing Bob (when the alternative is $y$) is now regarded as a different action from pushing Bob (when the alternative is $x$).


10. Notation: $\alpha > \beta$ means “you have a stronger reason to do $\alpha$ than $\beta$.”

11. Willenken, “Deontic Cycling,” 558–60. I am not sure that I understand Willenken’s argument. His claim, in effect, is that when we try to imagine such a situation, what we
The solution (if this is indeed a problem) is to individuate actions, not by alternatives, but, as we have done above, by outcomes (or alternatively, by the intentions with which they are done). That this method does not generate the same problems is clear enough from the application of our strategy to the three-way case; we get:

\[(x_1)\quad\text{Two saved (+2,000) and five dead (-5,000) = -3,000}\]
\[(y_1)\quad\text{Five saved (+5,000) and two dead (-2,000) = +3,000}\]
\[(z_3)\quad\text{One dead-by-me-but-doomed (-0) and six saved (+6,000) = +6,000}\]

So \(z_3 > x_1\) and \(z_3 > y_1\); that is perfectly compatible with our earlier results, and it is impossible to use this to generate any (putatively) nonsensical results.\(^{12}\)

Have we, perhaps, covertly individuated actions by alternatives? The answer, I think, is: not as such; but, given the examples, some alternatives end up built into the descriptions of the actions by way of the method of individuation—which is just what we should expect. As we have seen, the consequentialist will individuate actions by outcomes. In order to do that, she must make use, I think, of some notion of a default future: what would have happened in the absence of intervention. This will necessarily implicate, on one level, the available alternatives, but evidently not in a way that produces any trouble for the compatibilist project. I think this point is important in its own right, but another matter.

Parallel reasoning shows that the deontologist’s strategy (Sec. III) is also not a case of fine-graining. On the contrary, as we noted there, what we actually have is something like the reverse: the original description of Scenarios 1–3 left part of the description of each action out, and encoded it in the alternative; this was possible only because of a peculiarity of the setup: forced-choice binary evaluations.

---

Imagine instead is facing a choice between \(z\) and \(x\) simpliciter. But there seems to be nothing odd, anyway in terms of conceivability, in, for example, the wedding-menu claim that I have stronger reason to choose chicken (when the alternatives are pasta and seafood) than I have to choose pasta (when the alternatives are seafood and chicken). In any case, even if there is something dubious about this (perhaps I covertly just imagine choosing between chicken and pasta), it is still not clear that we have a serious problem here. The question would be how bad a result it really is that our theory should produce claims about what I have strongest reason to do in situations that are metaphysically impossible. Is this really a problem?

12. I am following Willenken here in treating Bob as doomed; in fact, it looks intuitively as though he is not: if we choose \(x\), then Bob does not die. In that case, the parallelism with the pair-wise cases above is further destroyed. (Note also that six are saved, and not five, in \(z_3\).) If this is accepted, then \(z_3\) should be rewritten as “One dead-by-me (-100,000) and six saved (+6,000) = -94,000,” and the correct choice will be \(y_1\). This is also compatible with our earlier results, although it is hardly good news for Bob. We get parallel results when we individuate by intention: \(z_3\) is “Push Bob in order to save six.”
So is compatibilism safe after all? As far as intransitivity goes, perhaps. If there is a problem with the project, it will be generated, I suspect, by the peculiar method of action-individuation we attributed to it, and in particular whether and how this method can be combined with the notion of a ‘dead-by-me’—matter for another note.