Who Says We Should Never Stab Innocents?

As all the world knows, there is a big difference between the kind of normative theory pioneered by Bentham and Mill and the kinds of normative theories pioneered by Aristotle and Kant. It is not just that Bentham and Mill had distinctive views about the nature and extent of our moral obligations towards, for example, animals, poor people and people of other races. It is rather that they approached moral questions in a distinctive kind of way. They were, to speak anachronistically, consequentialists. For them, the moral status of an act was determined by whether it did or did not promote value.

And, as all the world knows, the consequentialist approach is very flexible indeed. There are as many kinds of consequentialism as there are theories of value. A consequentialist can say, contra Mill, that it is wrong to enhance someone’s pleasure at a cost to her freedom, because freedom is more valuable than pleasure. A consequentialist can say, contra Bentham, that it is not wrong to bait bears, because animal suffering has no value.

But, as all the world knows, there are certain things that a consequentialist cannot say. For example, she cannot say that it is always, everywhere, impermissible to stab innocents. Even if (very implausibly) she judges the stabbing of an innocent to be the very worst kind of event that can happen, an event whose disvalue cannot be outweighed by the occurrence of any number of events of a different kind, then she still must concede that there are circumstances in which it is permissible to stab an innocent so as to prevent two innocents from being stabbed – after all, no matter how terrible, dreadful, awful an event the stabbing of an innocent is, surely it is better that one such event occur than two.
This paper is a mild counterweight to common wisdom. I will argue that we can accept one very distinctive and appealing feature of the consequentialist approach to normative ethics and still deny that there are circumstances in which it is morally permissible to stab an innocent so as to prevent two innocents from being stabbed.

2. An Attractive Deontic Principle

Everything turns, of course, on what this ‘one very distinctive and appealing feature of the consequentialist approach’ is. Here is the basic idea: you have not done anything wrong unless you could have made things overall go in an impartially preferable way. More accurately:

(C) An act is morally impermissible only if some other act available to the agent has an uncentered world-historic outcome that is preferable in the morally relevant, agent-and-time-invariant sense.

The italicized terms need some explaining.

To say that what matters is preferability in ‘the morally relevant, agent-and-time invariant sense’ is to say that there is a single relation between outcomes that determines the moral status of an act, no matter who performs it, no matter when it is performed. This stands in contrast to saying that what matters is preferability in some agent-(and/or-time)-relative sense – that what determines the moral status of an action on the part of an agent \( a \) at a time \( t \) is whether alternative acts have outcomes that are preferable-relative-to-\( a \)-relative-to-\( t \).

---

1 As numerous authors (e.g. Broome 1991, Dreier 1993) have observed, you can very easily adopt this kind of ‘agent-relative consequentialism’ while remaining committed to the view that certain kinds of actions are
To say that what matters is a relation between *uncentered* outcomes is to say that the morally relevant outcome of an act does not have, encoded within it, information about who brought it about and when. This stands in contrast to saying that what matters is a relation between centered outcomes – saying, for example, that the morally relevant outcome of an action performed by agent, $a$, at a time, $t$, is the triple $\langle w, a, t \rangle$, where $w$ is the way world-history goes as a result of $a$’s acting as he does at $t$.\(^2\)

The simplest uncentered conception of outcomes takes them to be maximally specific ways for world-history to go. The outcome of an act is the way that world-history would go if the act were taken. Perhaps this is too simple. Is there really a precise way that world-history would have gone if I had raised my hand five seconds ago? One might think that there are a great number of ways world-history might have gone, though no single way it would have gone – so the outcome of my act should be taken to be a class of possible world-histories. Or one might think that of the great number of ways world-history might have gone, some would have been more likely than others – so the outcome of my act should be taken to be a class of probabilistically weighted classes of possible world-histories. No matter. For present purposes we will just pretend that the simple conception is right – for each action there is one possible world-history that would obtain if the action were to be taken – and we will call this the *uncentered world-historic outcome* of the action.

---

\(^2\) Again, you can very easily adopt this kind of ‘centered-outcome consequentialism’ while remaining committed to the view that (e.g) lying is everywhere impermissible – just stipulate that for any agent $a$, time $t$, any triple $\langle w_1, a, t \rangle$ such that $a$ does not lie at $t$ in $w_1$ is superior to any triple $\langle w_2, a, t \rangle$ such that $a$ lies at $t$ in $w_2$. Always, everywhere impermissible. To generate an inviolable constraint against stabbing, for example, just stipulate that for any agent $a$, time $t$, any outcome in which $a$ does not stab at $t$ is inferior relative-to-$a$-at-$t$ than any outcome in which $a$ lies at $t$. Whether it really deserves the name ‘consequentialism’ is a separate, surprisingly controversial matter.
I have called this principle ‘(C)’ because it is an idea that philosophers working in the consequentialist tradition are sympathetic to. But I don’t in any way mean to suggest that it captures the essential spirit of the tradition. Some central figures in the tradition (for example Moore\(^3\)) do not conceive of outcomes as uncentered world-histories. Many central figures in the tradition are maximizers – they regard it as sufficient for an act to be impermissible that some alternative to it has a preferable outcome. And, as we will see later, many central figures in the tradition are committed to a stricter necessary condition than that given by (C).

\[3\] A Dilemma

Though (C) may be a mouthful to state, the thought behind it is very appealing indeed. Our practical deliberation focuses on actions because actions make it the case that things go one way rather than another, and we care about the way things go. Actions are interesting, for practical purposes, because they are difference-makers. So, when you are deciding what to do, it makes sense to compare (appropriately accounting for uncertainty) the way things will go if you act the one way with the way things will go if you act the other way. Most of the time you will make this comparison in light of idiosyncratic interests of your own, but when someone asks you to think morally, to think about what is the right thing to do, you should make the comparison in a more impartial way. This involves (roughly) being swayed only by publicly accessible considerations, considerations whose weight every reasonable person is in a position to appreciate. You

---

\(^3\) Moore seems (albeit intermittently) inclined to consider the outcome of an act to include all and only its ‘causal consequences’ – events that would not have occurred if the act were not performed. See Moore (1903) section 89, for example.
should take an act to be morally wrong only if some alternative makes things go in a way that you favor, in light of such considerations.

This is first and foremost a picture of how moral deliberation works, but it naturally suggests a picture of what makes acts right or wrong. According to this picture there exists a relation between ways for the world to go (we can call it the relation – is preferable to –) that determines the moral status of acts in the manner given by (C). What I am doing when, in the throes of moral deliberation, I set about comparing this way for events to go with that way for events to go, finding weighty, publicly accessible considerations for favoring one over the other, is attempting to judge whether and how the relation obtains between them.

Many philosophers, even self-proclaimed ‘deontologists’, have recognized the appeal of this picture, but, they say, some unpleasant baggage comes with it. If you accept (C) then you must also accept that there are circumstances in which it is permissible (or even obligatory – if you are also committed to maximization) to do one terrible thing in order to prevent two terrible things from being done. So, for example, you must accept that, if all other things are appropriately equal (there are no precedent effects, all threats are entirely credible… etc.) it is permissible (or even obligatory) to accost an innocent person in the street and stab her to death, so as to prevent Jack the Ripper from accosting two innocent people in the street and stabbing them to death. But this is, to speak mildly, surprising. Imagine the act. Imagine selecting your victim. Imagine removing your knife from its scabbard. Imagine the blood and screaming. How can one credible threat from Jack the Ripper render this morally okay?

So it appears as if we are faced with a dilemma. We must either deny (C) or accept:
Sometimes it is permissible to stab an innocent person so as to prevent two innocent people from being stabbed.

And this dilemma really is the central tipping point in contemporary normative ethics. Roughly half of the world’s moral philosophers, the Deontologists, try to explain why (C) is, in spite of its prima facie appeal, false. Roughly half, the Consequentialists, try to explain why claims like (1-for-2) are not so surprising when you think about them carefully.

I find this way of setting things up suspicious. To me (C) seems almost like a platitude – my accepting it reflects nothing more than a commitment to bringing moral thinking within the sphere of instrumental rationality, and a commitment to the idea that morally relevant considerations are in the appropriate sense agent-neutral. But (1-for-2) seems like a substantive, questionable moral claim – accepting it reflects a commitment to substantive and questionable views about what agent-neutral considerations are morally relevant. How can a platitude entail something substantive? How can something so thin entail something so thick?

---

4 There are two basic strategies here. One involves denying that, for the purpose of moral assessing an act, we should focus on the way it makes world-history go. Some philosophers who adopt this strategy argue that we should focus instead on the attitudes it expresses. See Kamm (1989) for example. Others argue that we should focus on the character of the agent who performs it. See Williams (1981) and Darwall (1986) for example. Yet others (e.g. contractualists and rule-utilitarians) argue that we should focus on systems of rules that prescribe for or against it. See Scanlon (2000) and Hooker (2000) for example. The other basic strategy involves denying that there is one relation between outcomes that determines the moral status of an act, no matter who performs it, no matter when it is performed. In short: morally relevant values are not agent-neutral. See Nozick (1974) and Thomson (2001) for example.

5 See Kagan (1989) Chapter 1, for example.
4. (C) and the Denial of (1-for-2) are Formally Compatible

So, does (1-for-2) really follow from (C)? Not directly. It is formally possible for someone who accepts the doctrine to deny (1-for-2), in the sense that there exists some relation between uncentered world-histories such that, if that relation tracked preferability in the morally relevant, agent-invariant sense, then (C) would yield (e.g.) that it is always impermissible to stab an innocent. This is because we live in a world in which deliberative conditionals front-track, which is to say that, generally speaking, when I am deciding whether to do one thing or the other at \( t \), the world histories that would (in the deliberatively relevant sense of ‘would’) come about if I did one thing or the other are related in this way (the graph represents their respective evolutions, with time on the \( x \)-axis, and a one-dimensional measure of the state of the world at a time on the \( y \)-axis):

![Diagram](image)

They are qualitatively the same up to (or almost up to\(^6\)) time \( t \). The first difference between them comes at \( t \), when I do the one thing in the one world-history, the other in the other. Then they differ afterwards with respect to the causal consequences of my behavior at \( t \). For an advocate of (C) to say that stabbing an innocent is always impermissible, all

\(^6\) On the standard account, even in mundane, everyday situations, deliberative conditionals may back-track a little. See Lewis (1979) p. 39.
she needs say is that, for any pair of world-histories structurally like the pair above, if the first difference between them comes with the stabbing of an innocent in the one, but not the other, then the other is, in the morally relevant agent-and-time-invariant sense, preferable to the one. There we go – in a world like ours, where deliberative conditionals front-track, it is always impermissible to stab innocents.

This is a familiar observation. Indeed, Graham Oddie and Peter Milne have shown that it is formally possible for a consequentialist to align herself with almost any overtly ‘agent-biased’ normative theory by endorsing a suitably contrived ordering of outcomes. But it misses the point. When philosophers say that a consequentialist is committed to (1-for-2), I take it that they do not mean that it is formally impossible for her to deny (1-for-2). They acknowledge that world-history will proceed in different ways if, for example, Sarah does or does not stab one person so as to prevent Jack the Ripper from stabbing two:

So it is of course possible to say that world history will proceed in a preferable way if Sarah stabs the one. The point, I take it, is not that we cannot say this, but that this is a very implausible thing to say. Given that more nasty stuff happens in WH₂ than WH₁, why
on earth should we think that WH₂ is in the morally relevant, agent-and-time-invariant sense preferable to WH₁?

To answer this question, an advocate of (C) who believes in a constraint against stabbing must appeal to considerations that give us grounds for thinking WH₂ preferable to WH₁. And not any old considerations will do. If he is to have any chance at all of persuading us that his view is right, the considerations must be weighty.

So it is not enough to observe, for example, that the stabbing in WH₁ takes place earlier than the killings in WH₂. That’s not a weighty consideration. The stabbing takes place a little earlier, so what? Why should earlier stabbings, just in virtue of being earlier, be in any relevant sense more important than later stabbings? After all, nobody wants to say that, if Sarah has the power to determine whether Jack stabs one innocent at noon or two innocents a minute later, she is morally obliged to choose the latter.

Furthermore, given that it is preferability in the morally relevant, *agent-invariant* sense that we are after, the considerations must, at least, be publicly accessible. It must not be the case that you need to be a special person, in a special position, to appreciate their force. So it is not enough to observe, for example, that in WH₁ Sarah does the stabbing, while in WH₂ someone else does the stabbing. Perhaps Sarah is in a position to appreciate the force of that consideration, but other people are not.

What other differences do we have to work with? Perhaps we could try to make something out of the fact that, in WH₁, the one is stabbed *so as to prevent the stabbing of the two*. We could say that stabbings-to-prevent-stabbings matter more than mere stabbings. You suffer a special kind of violation, above and beyond the run-of-the-mill

---

³See Oddie and Milne (1991)
violation that comes with being stabbed, when you are stabbed as a means to saving others from being stabbed, and this makes such stabbings more than twice as bad. But this doesn’t seem like a promising way of generating the result that it is generally speaking wrong to stab one so as to save two from being stabbed, because we need to deal with cases where the two would be stabbed so as to prevent four from being stabbed. In such cases we would need to say that stabbings-to-prevent-stabbings-to-prevent-stabbings are more than twice as bad as mere stabbings-to-prevent-stabbings. But what special violation is associated with stabbing-to-prevent-stabbing-to-prevent-stabbing that is not also associated with stabbing-to-prevent-stabbing? The victim is no more or less stabbed as a means to the saving of others either way. Furthermore, to deal with cases where the four would be stabbed so as to prevent eight from being stabbed, we would need to say that stabbings-to-prevent-stabbings-to-prevent-stabbings-to-prevent-stabbings are more than twice as bad as mere stabbings-to-prevent-stabbings-to-prevent-stabbings… and so on. But that seems very implausible indeed.

5. A Proposal

My major object here is to present a way of comparing world-histories that yields that $\text{WH}_2$ is in the morally relevant, agent-and-time-invariant sense preferable to $\text{WH}_1$, and to argue that, if one occupies a certain frame of mind, this way of comparing world-histories makes sense.

Why might one think $\text{WH}_1$ preferable to $\text{WH}_2$? Well, let’s say that two world-histories differ in virtue of an action’s being taken when the action is taken in one of the
histories, and it is true in that history that if some other action had been taken, the other
history would have come about. A *constraints-consequentialist* believes:

(The Core Evaluative Claim)

For any pair of world histories, actions in virtue of which they differ have a special
bearing on which is preferable in the morally relevant sense.

So, for the purposes of comparing \( WH_1 \) and \( WH_2 \), Sarah’s stabbing the one (in \( WH_1 \)) has
a special significance, because \( WH_1 \) and \( WH_2 \) differ in virtue of this stabbing, because it
is true that, in \( WH_1 \), if Sarah had refrained from stabbing, \( WH_2 \) would have come about.
But Jack’s stabbing two has no special significance. It is not true in \( WH_2 \) that, if Jack had
refrained from stabbing, \( WH_1 \) would have come about. So (since we are assuming that the
world-histories differ in no other interesting respects) \( WH_2 \) is, in the morally relevant
sense, preferable to \( WH_1 \).

Well and good. But why might one believe the Core Evaluative Claim?

First notice that when we attend to a thing, there are two kinds of questions we can
ask – how is it? and – why is it that way? And notice that there is a certain kind of attitude,
where you care, for the purposes of judging a thing, not just about how it is, but about why
it is that way. Sometimes this involves caring about the *history* of the thing. For example,
it might be that you find irregularity in the seams of a pair of your trousers charming,
quirky, authentic, until you discover why the seams are that way – in virtue of being
stitched by the callused fingers of an eight year-old, trembling from lack of sleep and fear
of punishment. Sometimes it involves caring about the *underlying structure* of a thing. For
example, it might be that your opinion of an image on your computer – tessellated clouds swirling over a checkered sea – is enhanced by discovering that is generated by an elegant, two-line algorithm. Sometimes it involves caring about the *purpose* of a thing. For example, it might be that you are bothered by the height of the bumper of your car, until you discover that it is that way to protect the legs of pedestrians.

Likewise, when we compare two things, we can ask two questions: ‘How are they different?’ and ‘Why are they different?’ And, there is a corresponding kind of attitude, where you care, for the purposes of comparing things, not merely about the ways in which they differ, but also about *why* they differ in those ways. You look beyond the raw list of differences between them and care about the reasons why those differences exist.

Our constraints-consequentialist is taking this attitude towards world histories. When he compares $WH_1$ and $WH_2$, he is not merely concerned with how they are different (one stabbing in $WH_1$, two in $WH_2$), but also with *why*. They are different because of a stabbing in $WH_1$ that does not take place in $WH_2$. If Sarah’s stabbing had not taken place in $WH_1$, $WH_2$ would have come about. Her stabbing in $WH_1$ is, to speak suggestively, what *makes* the worlds different. Jack’s double-stabbing in $WH_2$, on the other hand, is not what makes the worlds different. To our constraints-consequentialist, then, for the purposes of comparing $WH_1$ and $WH_2$, Jack’s double-stabbing matters less.

Perhaps an analogy will help you get a grip on the attitude. Imagine that two brothers, co-directing a movie, fall out over the staging of a pivotal scene. One thinks that it should take place in a damp, dim basement. The other thinks that it should take place on an airy rooftop. Unable to resolve their creative differences, the brothers end up releasing different movies. These movies are just the same for an hour and a half, until one cuts to
the basement, while the other cuts to the rooftop – a difference that induces predictable
differences in the way they proceed (continuity demands that later scenes be shot in
different ways). Each of the brothers agrees that the other has tied up the movie in the
appropriate way, given the staging of the pivotal scene. But each remains committed to his
own version.

One thrifty film-buff, unwilling to purchase both DVDs, might go about deciding
which to add to her collection like this: ‘What I need to do here is to tot up the aesthetic
merits of each movie (which will involve a certain amount of adding up good moments
and subtracting bad ones, a certain amount of seeing how everything hangs together) and
see which comes out ahead.’ Another thrifty film-buff might think like this: ‘These
movies differ because the pivotal scene happens on a rooftop in the one, in a basement in
the other. All the differences between them are more or less determinate consequences of
this, the source of the differences between them. So I will pay special attention to this
scene. If (e.g.) the basement works, aesthetically, in a way that the rooftop does not, I will
favor the basement-movie, even though it may turn out that the rooftop-movie has,
overall, slightly more good bits in it, and hangs together slightly better.’

Now I don’t wish to suggest that we actually do or would go about comparing movies
like these in the manner of second film-buff. But it seems to me that her approach is at
least an intelligible one, and her judgment (e.g.) that the basement movie is preferable is
supported by weighty, publicly accessible considerations. It is not as if she thinks the
basement-movie preferable because, although the rooftop-movie has more good bits in it,
the good bits happen later in the rooftop-movie. That consideration is not weighty.
Nobody wants to say that, generally speaking, the aesthetic significance of what happens
earlier in a movie is lexically prior to the aesthetic significance of what happens later –
great movies do not always begin with a punch. Nor does she think it preferable because,
although it has fewer good minutes, it has a better 90th minute. Again, that’s not a weighty
consideration. Some terrible movies have great 90th minutes, and vice-versa. Nor does she
think it preferable because she owns basements all over Los Angeles, and has an interest
in promoting their use in film, or because she is an agoraphobe who just hates rooftops
and just loves basements, to the point where she judges any movie with a basement scene
superior to any movie with a rooftop scene. The considerations that support those
judgments are not publicly accessible – you would need to share her property-interests and
psychological oddities to appreciate their force. Rather, she thinks the basement-movie
preferable because of the superior staging of the pivotal scene, the scene that splits the
movies apart.

The constraints-consequentialist’s approach to comparing world-histories is very like
the second film-buff’s approach to comparing movies. The considerations to which he
appeals in judging WH₂ preferable to WH₁ seem, in just the same way, publicly
accessible. If they have any force at all, you do not need to be in a special kind of position,
to appreciate this force. And they seem at least prima facie weighty. The constraints-
consequentialist tells us that, for the purposes of comparing world-histories, actions that
make the difference between them have a special significance, and we proceed as if this
were true all the time. The intuitions that support deontology gain their strength from the
fact that we care more about the kinds of actions we perform than about the kinds of
actions we bring about. But, typically, my actions are the only ones that make the
difference between the world’s being the way it is and the world’s being the way it would
be if I were to act otherwise. So, typically, taking the approach of the constraints consequentialist and caring more about the kinds of actions I perform than about the kinds of actions I merely bring about, come to the same thing.

6. Limitations of Constraints-Consequentialism

Now let me point out three interesting features of the view.

Back-Tracking Conditionals

The ordering of the constraints-consequentialist generates an inviolable constraint against stabbing in circumstances, like our own, where deliberative conditionals *front-track*. But what about circumstances in which they don’t? What about circumstances in which, the world-histories that an agent is in a position to bring about by doing one thing or another at $t$ differ significantly before $t$?

There is a celebrated, ongoing disagreement about whether and when such circumstances arise. Some philosophers say never. Others say only in very exotic cases – when we make decisions in the vicinity of a time-machine, black-hole, quantum-tunnel or such like. Others say that deliberative conditionals may back-track here on Earth in conditions not unlike ones that some of us find ourselves in everyday – if I am faced with a kind of decision-problem that my distant twin faced yesterday, it may be appropriate for me to think ‘If I were to do this, then he would have done this yesterday, but if I were to do that, then he would have done that yesterday.’

But we do not want our normative theory to be hostage to controversial metaphysics, so let’s put the disagreement to one side and concede, for the sake of argument, that there
may be circumstances in which deliberative conditionals back-track. Will the ordering of the constraints-consequentialist render stabbing impermissible in all such circumstances?

No. Imagine that Jane and Jack are in a position to stab or not, Jane one innocent at $t_1$ and Jack one innocent at $t_2$. But this time their decisions are *conditionally inter-dependent* in the following way: If Jane were to stab, then Jack would later not stab. If Jane were not to stab, then Jack would later stab. If Jack were to stab then Jane would earlier not have stabbed. If Jack were not to stab then Jane would earlier have stabbed. Which is to say that, at $t_1$, Jane is in a position to bring about the world-histories below, and, at $t_2$, Jack is in a position to bring about *these very same world-histories*.

![Diagram](https://via.placeholder.com/150)

In this case, the constraints-consequentialist will observe that Jane’s stabbing in $WH_1$ makes $WH_1$ and $WH_2$ different, so, for the purposes of comparing $WH_1$ and $WH_2$, it is more significant than other, non-difference-making stabbings. But Jack’s stabbing in $WH_2$ also makes $WH_1$ and $WH_2$ different, so it has the same significance. This leaves a constraints-consequentialist with no grounds (all other things being equal) for judging either world-history preferable, either stabbing impermissible.
And this is not an unfortunate artifact of one way of ordering world-histories. No (irreflexive) ordering will render both stabbings impermissible. Either \( WH_2 \) is not preferable to \( WH_1 \) or \( WH_1 \) is not preferable to \( WH_2 \). In the first case Jane’s stabbing is permissible. In the second case Jack’s is.\(^8\)

So, according to constraints-consequentialism, the circumstances we are in matter. Sometimes, when conditionals are not well-behaved, it may be permissible to stab. One might think this a problem for the view – after all, why should it matter, morally, whether conditionals are well-behaved, whether our causal powers extend to the past as well as the future? But this is uncharitable. The constraints-consequentialist does not think that it matters, in itself, whether conditionals back-track. Rather he thinks that in some circumstances where conditionals back-track, something important happens. My actions cease to be the only ones that make the difference between the world being the way it is and the world being the way it would be if I were to act otherwise.

\textit{Too Much is Impermissible}

Second, if a constraints-consequentialist wishes to say that some \textit{composite} acts are always and everywhere impermissible, he must say peculiar things about the permissibility of the acts of which they are composed.

\(^8\) This might seem to contradict Oddie and Milne’s claim that, by adopting an appropriate axiology, you can generate any first-order moral view you like. They recognized a problem in this vicinity, and argued that their claim was restricted to first-order views about the moral status of \textit{autonomous} actions. In situations like this, Sally and Jack are not acting autonomously, because each is controlled by the other’s behavior. This strikes me as an implausible view of autonomy. I would not consider myself any less autonomous if I were to discover that, in some distant galaxy, there exists a person whose decisions are \textit{quantum-entangled} with mine – if I get a muffin, he gets a bagel, and if he gets a muffin, I get a bagel. Nor would I consider my actions to be any less interesting objects of moral assessment.
One kind of composite act is *shtabbing*. You shtab someone by shaking his right hand, then using your left to plunge a knife into the exposed side of his body. One might, conceivably, wish to be a moral theorist who enjoins people to be utilitarians with one proviso: never shtab innocents. One might wish to say: ‘shtabbing innocents is always impermissible, but all other acts are impermissible only if they fail to maximize utility, where there is a functional relationship between world-histories and the utilities they contain.’ But a constraints-consequentialist cannot say this, because no axiology will generate this result.

Imagine that Jack will shtab two innocents unless Beth shtabs innocent John, someone she *loathes*. Beth refuses to play Jack’s nasty game and the two innocents die, but she knows that if she had gone so far as to shake John by the hand, the idea that she would have an excuse for stabbing him would have overcome her better judgment, and she would have been unable to resist doing it.

If it is impermissible for Beth to shtab John, this must be in virtue of \(WH_1\) being in the morally relevant sense inferior to \(WH_2\). But then it must be impermissible for Beth to shake John’s hand – after all, \(WH_1\) and \(WH_2\) are precisely the world histories that she will
bring about by shaking or not shaking his hand. So shaking his hand is impermissible even though, by hypothesis, she maximizes utility by doing so. Even though there is no general prohibition against hand-shaking, there are circumstances in which utility-maximizing hand-shaking is impermissible.

If the constraints-consequentialist finds this result uncomfortable, there are ways for him to try to avoid it. He could, for example, deny that there are any composite actions (like shtabbing) that are always and everywhere impermissible. This seems like a stretch. Better, I think, just to concede that, when composite actions are always and everywhere impermissible, there may be circumstances in which the actions of which they are composed are impermissible, even though there is no general prohibition against actions of this type, and the token action is utility-maximizing.

Intransitivity

Finally, I generated the ordering of the constraints-consequentialist by stipulating that one world-history is preferable to another when a certain structural relation obtains between them. But can we generate the same ordering in more constructive way, by assigning numerical values to world-histories and then stipulating that one world is preferable to another when and only when its numerical value is greater?

The answer is no, because a constraints-consequentialist’s ordering must, on pain of implausibility, be intransitive while, of course, the relation greater numerical value than is not. The intransitivity will arise within triplets of world-histories that bear a certain kind of causal relation to one another. Here’s one way of embellishing such a relation.
Sheriff-Pete happens upon psychopaths Molly and Sally, with three bound victims. He runs away and Sally, at leisure, stabs Victims 1 and 2. If he had intervened, Molly would hurriedly have stabbed Victim 1. If he had intervened and Molly had refrained from stabbing Victim 1, psychopath-Sally, angered by Molly’s weakness, would have stabbed Victims 1, 2 and 3.

In this case the world histories that might have come about look like this:

In WH₁ there are two stabbings (on top of the countless stabbings that occur in all three world-histories). But, in WH₁, if Pete had intervened, WH₂ would have come about – a world history in which there is one stabbing. But, in WH₂, if that stabbing had not taken place, WH₃ would have come about – a world-history in which there are three stabbings.

When a constraints-consequentialist sets about ordering these world-histories, it is
natural for him to say that $\text{WH}_2$ is preferable to $\text{WH}_1$, since these worlds do not differ in virtue of a stabbing and more stabbings occur in the latter. Indeed, a constraints-consequentialist must say this if he is to say that Pete does wrong by running away. And he is committed to saying that $\text{WH}_3$ is preferable to $\text{WH}_2$, because these worlds differ in virtue of a stabbing in $\text{WH}_2$. What about $\text{WH}_1$ and $\text{WH}_3$? If the ordering is transitive it follows that $\text{WH}_3$ must be preferable to $\text{WH}_1$. But what grounds has he for saying this – given that $\text{WH}_3$ and $\text{WH}_1$ do not differ in virtue of any action taken in either world, and there are more stabbings in $\text{WH}_3$ than in $\text{WH}_1$? I can see no grounds. On pain of implausibility, a constraints-consequentialist must deny that $\text{WH}_3$ is preferable to $\text{WH}_1$, that the ordering is transitive.

In one respect this intransitivity is quite benign. A famous reason for someone who accepts (C) to be wary of adopting an intransitive ordering of outcomes, is for fear of creating situations in which each of finitely many actions available to an agent is impermissible. If, for example, I can bring about outcomes A, B or C, and the consequentialist tells me that A beats B, B beats C, and C beats A, then the consequentialist tells me that I cannot avoid doing something impermissible. But this ordering does not give rise to such situations. Pete is not in a position to bring about outcomes $\text{WH}_1$, $\text{WH}_2$, and $\text{WH}_3$.

But in another respect it is less benign. The view that certain kinds of actions are always and everywhere impermissible has much-discussed problems with uncertainty.\(^9\) Suppose that there is a constraint against stabbing innocents. How should we proceed when we can reduce the number of stabbings that ever take place by stabbing someone,

---

\(^9\) The problems are brought out most clearly in Frank Jackson and Michael Smith’s: “Absolutist Ethical Theories and Uncertainty” (unpublished manuscript).
but we aren’t quite sure whether he is innocent? How should we proceed when we are not quite sure whether, as a consequence of what we do, someone else will stab an innocent? You might think that, by adopting (C), the constraints-consequentialist helps himself an easy, general solution to these problems. He, like any other consequentialist, can recommend that we turn to expected value theory. We should assign numerical values to the various epistemically possible outcomes in a way that makes the relation *greater numerical value than* align with the relation *preferable in the morally relevant sense to*. Then we should assign conditional probabilities to each of the outcomes – tracking how likely it is, as far as we can tell, that we will bring about the outcome by doing one thing or the other. Then we should perform the action with highest expected value.

But the solution won’t work. Suppose that Pete is deciding whether or not to intervene, and isn’t quite sure whether or not, if he does, Molly will stab. If the constraints consequentialist is right, Pete cannot assign numerical values to outcomes in a way that makes the relation *greater numerical value than* align with the relation *preferable in the morally relevant sense to*. No matter what assignment he chooses, the relation *greater numerical value than* will always be transitive, but the relation *preferable in the morally relevant sense to* is intransitive. So Pete must respond to uncertainty in some other way.

This is not big news. No deontologist has ever suggested that we should respond to uncertainty by assigning numerical values to outcomes in a way that makes the relation *greater numerical value than* align with the relation *preferable in the morally relevant sense to*, and then applying the machinery of expected value theory. But it shows that, while adopting (C) may leave a deontologist no worse off than he was before, it does not confer all of the benefits of a consequentialist approach upon him either.
7. Value

So, to commit yourself to saying that we should sometimes stab so as to prevent stabbings, it suffices to adopt (C), plus the view that the relation is preferable to is transitive:

(C+) An act is morally impermissible only if some other act available to the agent has an uncentered world-historic outcome that is preferable in the morally relevant, agent- and-time-invariant, transitive sense.

And this is something that Bentham, Mill, Sidgwick, Moore and the great majority of their descendents would readily accept, because they assume, at least implicitly, that outcomes are partially ordered by preferability – the relation is preferable to is irreflexive, transitive and anti-symmetric.

There are many ways to build the transitivity assumption into your moral theory. One way is to state it explicitly. Another is to take it that there is some function from world-histories to intrinsic values¹⁰, and the relation is preferable to is the relation has greater intrinsic value than – transitive by design.¹¹ Another is to say that what matters is which outcome is ‘more valuable’, and assume a background theory of value that requires this to

---

¹⁰ This function may have a simple, additive character – you get the overall value of the world-history by summing up all the good and bad things that happen in it and subtracting the latter from the former. Or it may have a more complex character. See Kagan (1989). No matter.

¹¹ When Eric Carlson sets out to characterize consequentialism, for example, he writes that ‘the most fundamental feature of consequentialism is that it makes the moral status of an action depend entirely on considerations of intrinsic value.’ (Carlson (1995) sec. 2.2)
be transitive.\textsuperscript{12} Yet another is just to say that what matters is which outcome is ‘better’ – after all, it is plausibly an analytic truth that ‘if A is better than B and B better than C, then A is better than C’.\textsuperscript{13}

It is this idea, the idea that the appropriate way to compare world histories is to \textit{evaluate} them, where evaluating entities involves ordering them in a transitive way, that prevents a traditional consequentialist from embracing deontology. If you want to be a deontologist, if you cannot stomach the thought of Sarah’s bloody actions being right, then you must reject it. But you need not deny that the correct way to think about the morality of an action is to compare, in an appropriately impartial manner, the way the world will go if it is performed with the way the world will go if it is not. There is a perfectly intelligible, impartial approach to making such comparisons that yields that we should not stab so as to prevent two stabbings.

\textit{References:}


\textsuperscript{12} Some meta-ethical theories have this implication – for example, a view that value (or ‘goodness’) is a kind of stuff that inheres in entities, and whether or not one outcome is more valuable than another is a matter of whether it has more or less of the stuff. Others may not – for example, a dispositional theory of value according to which whether one outcome is more valuable than another is a matter of whether certain people in certain circumstances would rather that it came about. Such a theory may allow that the preferences of the certain people in the certain circumstances be cyclical. My own view is that the entire practice of \textit{evaluating} things presupposes transitivity, so, to the extent that a theory of value allows for intransitivity, it fails to be a theory of value.

\textsuperscript{13} John Broome gives a careful and convincing argument for this in Broome (2004). Even Larry Temkin, who famously argues that “all things considered better than” is not a transitive relation’ (Temkin (1996) p. 175) is willing to concede the point:

‘...I have little interest in terminological disputes. The crucial question is whether a central element of moral and practical reasoning involves an intransitive relation. If my arguments lead us to recognize that this question’s answer is “yes,” it matters not what we \textit{call} this element. If people want to use “all things considered better than” solely to express a transitive relation, so be it. We then need another expression...’ Temkin (1996), p.208


Dreier, James (1993): ‘Structures of Normative Theories’, *The Monist* 76


Kamm, Frances (1989): ‘Harming Some to Save Others’, *Philosophical Studies* 57


