Criminal Law, Philosophy (2019) 13:1–25
https://doi.org/10.1007/s11572-018-9456-y

ORIGINAL PAPER

Wild Goose Chase: Still No Rationales for the Doctrine of Double Effect and Related Principles

Uwe Steinhoff[1]

Published online: 23 February 2018

Abstract I focus on the question as to what *rationale* could possibly underlie the doctrine of double effect (DDE) or related principles. I first briefly review the correct critiques of the claim that people who intend some evil as a means to a good must be "guided by evil," and that this is allegedly always wrong. I then argue that Quinn's claim that violations of the DDE express certain negative attitudes of the agent and that agents violating the DDE must make an additional morally problematic presumption regarding their victims is mistaken. Tadros claims that an agent violating the means principle must force his victims to adopt his goals. I demonstrate that the difference Tadros tries to construe between an agent inflicting intended harm and an agent inflicting merely foreseen harm is non-existent. Sarch's official rationale for the DDE also fails to distinguish harming as a means from side-effect harming, and reformulations of his rationale that suggest themselves run into severe problems. Walen's defense of the means principle in terms of the "restricting claims principle" and Øverland's appeal to "moral obstacles" are susceptible to counter-examples and appear to be question-begging. Recently, Walen has offered a revised formulation of his Restricting Claims Principle, claiming that it overcomes counter-examples and explains the means principle. I will argue that it contradicts the means principle and does not overcome the counter-examples. Thus I conclude that so far we are still left without a reasonable rationale for the DDE or related principles.

Keywords Doctrine of double effect · Means principle · Moral obstacles · Warren S. Quinn · Rationale · Restricting claims · Alexander Sarch · Victor Tadros · Alec Walen

✉ Uwe Steinhoff
ustnhoff@hku.hk

[1] Department of Politics and Public Administration, University of Hong Kong, Pokfulam Road, Hong Kong, China

1 Introduction

Some philosophers think that our different moral intuitions about certain pairs of cases support the doctrine of double effect (DDE) or related principles, like the means principle. Here I will for the most part set aside this question, that is, whether there actually is such intuitive support. Instead I focus on the question as to what *rationale* could possibly underlie the DDE or related principles: what kind of deeper explanation could there be for the alleged correctness of the DDE or of related principles?[1] Not too many explanations have been offered. I will first, in Sect. 2, have a brief look at the claim that people who intend some evil as a means to a good must be "guided by evil," and that this is allegedly always wrong (or at least worse than allowing evil as a side effect of one's actions). I have nothing original to say about this rationale. It has already been successfully refuted by others, and I only review the results for the sake of completeness. Then, in Sect. 3, I turn to Warren S. Quinn's rationale for the DDE, which combines the ideas that violations of the DDE express certain negative attitudes of the agent and, relatedly, that agents violating the DDE must make an additional morally problematic presumption regarding their victims than agents who merely accept certain harms or evils as side effects of their actions. In my view, Quinn's account has also already been effectively refuted, namely by Jonathan Bennett, but here I do have something original to add to Bennett's critique. In particular, I show that Quinn's claim that agents who violate the DDE (or his version of it) have to make an additional problematic presumption compared to people who harm others in accordance with the DDE is both biased and mistaken. Then, in Sect. 4, I turn to Victor Tadros's recent defense of the means principle, which relies on the idea that an agent violating the means principle must force his victims to adopt his goals. I demonstrate that this argument fails. The difference Tadros tries to construe between an agent inflicting intended harm and an agent inflicting merely foreseen harm is non-existent. In Sect. 5, I argue that Alexander Sarch's official rationale for the DDE (which he only applies to culpability, not permissibility) fails to distinguish harming as a means from side-effect harming. Reformulations of his rationale that suggest themselves also run into severe problems. Some of those problems are similar to those faced by Quinn's account, but there are also additional problems. In any case, his "Two Strikes Arguments" is unsuccessful. Finally, in Sect. 6, I examine Alec Walen's defense of the "restricting claims principle" (RCP) and Gerhard Øverland's appeal to "moral obstacles." Both accounts are susceptible to counter-examples. Øverland is willing to bite the bullet but provides no argument why one should bite the bullet. Walen, in contrast, tries to explain away the counter-example, yet his proposed explanation is arbitrary and could be used not only to justify diverting a trolley away from five to one but also, very much against his wishes, to justify pushing one man off a bridge to stop the trolley using his body and save five. Thus the very difference Walen wants to explain would simply disappear. Moreover, his attempt to explain away the counter-example

[1] On occasion a defender of the DDE or the means principle admits that there might be no deeper rationale available. See Ramakrishnan (2016, esp. at 165).

violates his official rationale, which relies on the idea that the presence of people with "restricting claims" imposes "negative externalities" on other agents such that the restricting claims must be weaker to account for that. The potential victims in the counter-example, however, simply do not impose such externalities, a fact that Walen's forced reinterpretation of the example cannot change. Recently, however, Walen has offered a revised formulation of his RCP, claiming that it overcomes counter-examples and explains the means principle. I will argue that it contradicts the means principle and does not overcome the counter-examples. Thus I conclude that so far we are still left without a reasonable rationale for the DDE or related principles.

2 Aiming at and Being Guided by Evil

Thomas Nagel has suggested that "to aim at evil, *even as a means*, is to have one's action *guided* by evil … But the *essence* of evil is that it should *repel* us. … So when we aim at evil we are swimming head-on against the normative current."[2] This is supposed to support the DDE, since someone producing an evil in violation of the DDE would allegedly be guided by evil—and thus swim "against the normative current"—while someone who produces evil in compliance with the DDE would not be guided by evil.

This rationale for the DDE has been popular in the past, but meanwhile it seems to have fallen out of favor. As Bennett remarks:

> The force of Nagel's treatment comes from our thinking of being 'guided by evil' as being guided in a way which essentially involves the thought of evil [for its own sake]; that really would be swimming against the normative current; but it is not what the terror bomber [or whoever intends an evil to secure a greater *good*] is doing.[3]

Moreover, Dana Kay Nelkin and Samuel C. Rickless point out that "defenders of the aiming-at-evil rationale are caught on the horns of a dilemma depending on how they choose to understand the nature of evil." On the one hand, "if wrongness is part of the essence of evil, then it is circular to explain the wrongness of an action (or its tendency to be wrong) by adverting to the fact that, in performing the action, the relevant agent aims at something that is wrong." On the other hand, if an "evil" is understood as something that is bad (not morally bad, but bad for someone), we are confronted with the problem that "it does not seem wrong in itself to aim at something very bad (such as great harm): for example, it does not seem wrong in itself to

[2] Nagel (1980, 132).
[3] Bennett (1998, 224).

aim at harming people who were or are engaged in wrongful attacks on other people."[4] Thus, the aiming-at-evil rationale does not work.[5]

3 Quinn on Attitudes and the Number of Morally Problematic Assumptions

According to Quinn, the DDE should be understood in the following way:

> [I]t distinguishes between agency in which harm comes to some victims, at least in part, from the agent's deliberately involving them in something in order to further his purpose precisely by way of their being so involved (agency in which they figure as *intentional objects*) and harmful agency in which either nothing is in that way intended for the victims or what is so intended does not contribute to their harm. Let us call the first kind of agency in the production of harm *direct* and the second kind *indirect*. According to this version of the doctrine, we need, *ceteris paribus*, a stronger case to justify harmful direct agency than to justify equally harmful indirect agency.[6]

Basically, the idea is that using someone as a means—even in the attenuated sense of "harmful direct agency"—is more difficult to justify than harming them as a side effect of the pursuit of one's ends. But why should this be the case? Quinn explains:

> The agent of direct harm … sees [his victims] as material to be strategically shaped or framed by his agency. Someone who harms by direct agency must therefore take up a distinctive attitude toward his victims. He must treat them as if they were then and there *for* his purposes. But indirect harming is different. [Victims of indirect harm] may … be treated as beings whose harm or death does not much matter – at least not as much as the achievement of the agent's goal. And that presumption is morally questionable. But in a counter-

[4] Nelkin and Rickless (2015, 403). Sarch (2017a, 460) thinks that these two problems can be overcome if one uses the DDE not for assessments of permissibility but for assessments of the degree of culpability. He realizes that this is a rather restricted understanding of the DDE (ibid., 460–461). My concern throughout this paper is with permissibility—and that is how the DDE has been traditionally understood and intended. Moreover, Sarch's account actually fails even if only applied to culpability, as we will see below.

[5] A related "rationale" might be Wedgwood's (2011, 392–393). He claims that someone who acts intending certain outcomes is more "agentially involved" in the "intentional dimension" than someone who acts foreseeing certain outcomes, and that the more agential involvement an act (with bad effects) contains the worse the act is. Obviously, however, this does not explain anything but only combines a mere relabeling of the intending/foreseeing distinction with a mere claim that one is worse than the other.

[6] Quinn (1989, 343–344).

part case of direct agency there is the *additional* presumption that the victim may be cast in some role that serves the agent's goal.[7]

The appeal to a difference in attitude fails. Suppose a villain, on my cell phone, credibly threatens to kill 10 children unless I immediately slap the innocent man next to me, and hence I do slap that man in order to save the children. I have used him as a means and "involved" him as an intentional object in my saving the children. However, I did not thereby treat him as if he existed *for* my purposes.[8] I have not, to use the Kantian expression, treated him as "merely" a means. I would be doing this if I enslaved him and treated him as if he had no rights whatsoever. Nor need I think that he exists for the purposes of the 10 children. They may not enslave him either. Rather, what I thought was that he has a right not to be slapped by me, but that this right can be justifiably overridden given what is at stake. This is in no way different from a situation of "indirect harming," where the villain instructs me to immediately make a slapping movement to the right with my right hand, leaving it on the level where it is now (I just scratched my chin). Here the slapping movement itself, not the fact that the man gets slapped (that is, that he is harmed), would save the children. Moreover, in both cases, I can equally regret my action and apologize to the man afterwards, explaining to him why I did it. Thus "indirect harming" is *not* different.

It does not help, incidentally, to simply insist that "in some sense" I did treat the first man as existing for my purposes. Leaving aside the fact that no ordinary speaker would say that, it should be noted that if intentionally slapping the villain to save the children is to treat him as if he is there for my purposes, then intentionally asking someone for the time to find out what the time is is also to treat him as if he is there for my purposes—I do use him as a means to my ends, after all. But since that is not wrong (not all things considered, not *prima facie*, and not *pro tanto*), an appeal to treating someone as being there for my purposes does not produce the sought-after explanation for the particular wrongness of direct agency. Of course, Quinn only says that someone who *harms* by direct agency "must therefore take up a distinctive attitude toward his victims," but that would only explain why harming by direct agency is worse than harming by indirect agency if there were something intrinsically wrong in taking up that distinctive attitude; and that there is something intrinsically wrong in taking up that attitude is only credible if Quinn could show either that there is something wrong in asking someone for the time or that by asking for the time I am not taking up that distinctive attitude. Quinn shows neither; in

[7] Ibid., 348–349. Nelkin and Rickless (2014, 131–133) attribute to Quinn a "Kantian approach," which Quinn, allegedly, defends in an "dependent rights version" while they prefer an "independent rights version." This difference need not concern us—more important is that this attribution of the Kantian rationale to Quinn is problematic (as is their own invocation of the Kantian rationale as an explanation). They quote Quinn (1989, 350) referring to Kant and saying that "[p]eople have a strong *prima facie* right not to be sacrificed in strategic roles over which they have no say." Yet, Quinn (1989, 350, n. 25) recognizes that people will—including and perhaps especially on Kant's view—*also* have a strong *prima facie* right not to be sacrificed *collaterally* without having a say on that. The question is why one is *worse* than the other—and to explain *that* Quinn evokes the "additional distinctive attitude," not Kant.

[8] Compare Bennett (1998, 220–221).

fact, he is not even asking the question—which means that he is simply making an unwarranted stipulation.

The appeal to mathematics, to the counting of "presumptions," fails too. In fact, it is question-begging. After all, the question is precisely whether harming people as a means constitutes an additional *evil* beyond the evil of harming itself. If it does not, then the agent's presumption that he may cast the victim in some role that serves the agent's goal might well be additional but it would also be irrelevant.

Moreover, if one starts counting presumptions, one should note that one often hears something like this: "Let's win the war so that our soldiers did not die in vain." Dying in vain, it seems, is considered something bad. Being harmed in vain would also seem to be bad. In the first version of the slapping example, the man's being slapped saved 10 children. The harm inflicted on him served a noble purpose. The harm inflicted on the man in the second case, in contrast, did not serve any purpose. It was in vain. Thus, in the first case, the agent acted on the presumption that the victim's not being slapped does not matter more than the agent's goals and on the additional presumption (which also *can* exist in the second case, although it need not) that he may cast the victim in a role that serves the agent's purpose. In the second case, the agent acted on the presumption that the victim's not being slapped does not matter more than the agent's goals and on the additional presumption (which need not exist in the first case) that he may harm him without that harm serving any noble purpose. So we have the same number of presumptions in both cases. Quinn might consider the presumption he emphasizes to be more important, but that would require an independent argument. The appeal to attitudes is such an argument, but, as we saw, it fails.

Quinn also tries to further explain what he means when he says that the civilians in the *Terror Bomber* case (wherein civilians are killed to terrorize the population) serve the agent's goal but not the civilians who are collaterally killed in the case of the *Strategic Bomber* (wherein civilians are killed as a side effect of the bombing of a munitions factory):

> Suppose, for example, the civilians had effective bomb shelters … Then the bomber … could succeed only with the cooperation of the victims. The service exacted would then be voluntary. But in cases of indirect agency the victims make no contribution. If the civilians in SB [Strategic Bomber] had shelters …, the bomber … would see no point in their refusing to use them.[9]

However, if the civilians had anti-aircraft guns, the bomber would most certainly see the point of the civilians refusing to use them. (We will come back to this: Quinn is committing a mistake here that Tadros repeats, as we will see below.) If the civilians were actually able to shoot the bomber down, he could only succeed with their cooperation. So there is, again, no difference here. More importantly, the

[9] Quinn (1989, 349).

death of the civilians is involuntary in both cases.[10] Neither the victims of the *Strategic Bomber* nor the victims of the *Terror Bomber* volunteered to be bombed. And the question is why being involuntarily bombed is supposed to be less objectionable when one's foreseen and useless death is presumed not to matter as much as the bomber's goals than when one's intended and useful death does not matter as much as the bomber's goals. The only answer that Quinn gives us in the end is the appeal to attitudes. But again, that appeal does not work, for there simply need not be any difference in attitudes between the *Terror Bomber* and the *Strategic Bomber.*

4 Tadros on the Means Principle and Making Others Adopt One's Goals

Tadros's explanation of the alleged difference is inspired by Quinn, but nevertheless somewhat different:

> [T]he claim that it is wrong to use a person as a means is grounded in an independent moral judgment about a person's right to set her own ends, even if these are not impersonally best, and the relationship between this right and the duties of others.[11]

The part before the second comma in this quote is meant to account for the difference between *Strategic Bomber* and *Terror Bomber*, and the part on the relationship between rights and duties is to account for certain exceptions to the prohibition on using others: one may do so if the other person has a duty to serve the end or if the person consents (in which case, of course, she has voluntarily adopted the end in question).[12] Yet, while Tadros clearly thinks that *Strategic Bomber* (and *Trolley*, see below) is an example where the collateral victims' rights to set their own ends are not being violated or infringed and that *Terror Bomber* (and *Bridge*, see below) is an example where the relevant rights are thusly violated or infringed, it all depends on what one *means* by interfering with another person's right to "set her own ends." To wit, it is obvious that, in one pretty ordinary sense of the phrase, the *Strategic Bomber's* victims' right to set their own ends is, in fact, frustrated by his killing them, for it was certainly not their goal in life to perish at his hands—rather, it was their goal to live on. Moreover, once people are dead, they cannot set any ends anymore—that should certainly count as interference. So there is no difference here between the two cases.

[10] I have encountered here the somewhat mysterious objection that this observation of mine is "non-responsive" since Quinn recognizes that the death of the civilians is involuntary in both cases and he allegedly merely says that it *would* be voluntary *if* they cooperated with the bomber. In reply, first, recognizing something is not quite sufficient—one should also draw the logical conclusions; and second, yes, it would be voluntary if they cooperated with the bomber, but that is so in *both* cases (not shooting down the bomber would be a case of cooperation—which, as I pointed out in the main text, is evidently something Quinn does *not* recognize), and it therefore does not establish any difference between the cases.

[11] Tadros (2015, 68).

[12] Ibid.

Another interpretation of interfering with someone's setting his own ends would be expressed by the bombers saying to the victims: "Well, we really don't want to tell you to, or to make you, *adopt* certain ends: you may *set* yourselves the end of becoming loan sharks, of becoming professional boxers, or even of doing your best to shoot us down—that's your liberty-right, fair is fair. In fact, we don't ask you to do *anything*, and certainly not to set yourselves certain goals or to intend certain things—we are just going to *kill* you." However, *both* bombers can say this, and so again we find no difference between the two cases.

Given, therefore, that ordinary language interpretations of "interfering with (or violating) another person's right to set her own ends" do not generate the difference Tadros is after, one might reasonably suspect that he means the phrase in a more *technical* sense. He says, for instance, that he relies "on the idea that it is normally wrong to use a person in service of an end in a way that harms her without consent if that person is not required to serve that end at the relevant cost."[13] This formulation, used shortly after the previous indented quote, suggests that to use a person in service of an end is to violate or infringe her right to set her own ends. But if this is indeed what is *meant* by violating or infringing a person's right to set her own ends, then the appeal to such a right does not *explain* why it is wrong to *use* someone, but only makes the very same claim with different words. Accordingly, the appeal to a person's right to set her own ends, interpreted in this technical sense, does not provide a rationale for the means principle.

Yet Tadros does offer another rationale. This rationale does not concern itself so much with the *frustration* of the victims' ends or with robbing them of the ability to set their own ends (as in the first interpretation I offered), nor with the technical interpretation discussed in the previous paragraph, emphasizing instead the agent *making* his victim set certain ends, precisely as in the *second* interpretation of "violating a person's right to set her own ends" offered above (the interpretation adopted by the bombers two paragraphs ago). To wit, there is a straightforward way in which I can make someone *adopt* my goal and hence serve it as *an agent*: if I point a gun at someone and tell him: "Jump off the bridge," and he does, then, indeed, I have *made* him *adopt* the end of jumping off the bridge (of course, not as an end in itself) and made him serve my end as an agent. And Tadros's strategy for providing the sought-after rationale is now, to anticipate, to "rel[y] on the close relationship between forcing a person to act in service of a certain end and using that person's body against her will to serve that end" and to claim, further, that "harming a person as a side effect in pursuit of one's ends" is *not* "akin to compelling her to act in service of that end."[14] Yet I shall argue that Tadros fails to establish that there is such a close relation between using a person's body and forcing a person to act in service of a certain end; and, accordingly, he also fails to establish any moral difference between using and "side-affecting," as we may call it.

Let us look at this in more detail. We are already familiar with the strategy of connecting the idea of using people as means to the idea of making them serve one's

[13] Ibid., 68–69.

[14] Ibid., 67.

ends from Quinn, and we also already saw that it does not work. After all, that people are *used* does not mean that they are *made to serve*. To repeat the example: yes, if I point a gun at someone and tell him: "Jump off the bridge," and he does it, then, indeed, I have made him serve my end. If, however, as in the *Bridge* case, I simply throw him off the bridge (in order to stop a runaway trolley with his body and thus save five people on the tracks), then I have used him, yes, but I have not made him serve my ends—I merely made his *body* serve my ends. Tadros is well aware of this difference:

> V's body is used, but V is not coerced to act. Some might accept that it is normally wrong to compel a person to do something that she is not required to do but deny that D [the agent] acts wrongly in *Bridge*. Coercing a person to act exploits her agency, whereas using her body bypasses her agency.[15]

In fact, he admits that "[f]orcing a person to destroy herself might be thought especially bad," but then adds that it is nevertheless "surely normally wrong to bypass a person's agency to use her body to secure an end if she lacks a duty to serve that end."[16] That is correct, of course, but it is not the point as far as the DDE or the means principle is concerned. Rather, the question is whether bypassing a person's agency while *using* her body is worse than bypassing a person's agency while *discarding* her body. Therefore, Tadros's undeniably true statement that it would be wrong to throw an unconscious person off a bridge for no good moral reason[17] is beside the point. The question is whether it is *more* wrong for me to use my car to shove an unconscious person off the bridge in order to then have her body reserve my parking space below (*Car 1*) than to knowingly shove an unconscious person off the bridge as a side effect of parking my car right where the person is lying (*Car 2*). So far Tadros has not shown that it is more wrong, and intuitively it is not.

For what it is worth, however, let me note that I have come across the objection here that my parking space examples do not speak against the DDE since one can invoke another doctrine, the doctrine of doing and allowing (DDA), to explain why both cases are equally wrong (they are both cases of active harming, not of merely allowing harm to happen). This objection is, I dare say, silly. First, the DDE says that *all else being equal* intentional harming is worse than foreseen harming (so it would be methodologically incompetent to test the DDE by offering cases where the DDA applies to one case but not to the other), and second, the DDA *also* applies to *both* the original *Trolley* case *and* the original *Bridge* case: *both* are cases of active harming—so why doesn't the DDA "explain" *there* something that defenders of the DDE emphatically deny, namely that both cases are *equally* wrong? I have also heard the objection that to render an act permissible, the DDE also demands that the proportionality requirement be met. Yes, I know that. But again, the question is whether all else being equal it is more *difficult* to justify using my car to shove an unconscious person off a bridge in order to then have her body reserve my parking

[15] Ibid., 66.
[16] Ibid., 66–67.
[17] Ibid., 67.

space below than to knowingly shove an unconscious person off the bridge as a side effect of parking my car right where the person is lying. Suppose a millionaire will save *x* innocent people from starvation if I park at that exact spot. Must parking save *more* lives in the first case than in the second to be permissible? That seems to be counter-intuitive—there is no discernible difference in these cases, and that does speak against the DDE.

Therefore it is premature when after the example of the unconscious person Tadros declares that "[w]e are now well placed to explain the contrast between *Bridge* and *Trolley*." (In *Trolley* a trolley is diverted from five to one and his death foreseen.) His explanation, as already indicated, "relie[s] on the close relationship between forcing a person to act in service of a certain end and using that person's body against her will to serve that end." He adds: "The question is whether harming a person as a side effect in pursuit of one's ends is also akin to compelling her to act in service of that end. I think that it is not …"[18] Yet Tadros has still not established that there *is* such a close relation. His claim that there is is not credible in light of his own concession that "[f]orcing a person to destroy herself might be thought especially bad." (It should be added that it might not only be "thought" especially bad but that, all else being equal, it most definitely is especially bad. A person being killed is only being killed. A person being forced to kill himself is being forced *and* killed, where said force will have to rely on threatening something that the threatened person fears more than death itself.) While Tadros now tries to profit from the special badness of such an act by basically suggesting that it can be transferred ("the close relationship") to the completely different kind of act of using a person's body while bypassing her agency, the very difference in badness betrays the fact that this transference is not possible. There is no close relationship, and hence the crucial question remains: why is *using* an unconscious body by throwing it off a bridge worse than *discarding* an unconscious body by throwing it off a bridge, given that in both cases one most certainly has *not* done something that indeed might be worse than both, namely *coercing* the person to *jump* off the bridge?

So Tadros fails to establish that there is a close relation between using a person's body and forcing a person to act in service of a certain end. Accordingly, as already mentioned above, he fails to establish any moral difference between using and side-affecting. Does he have any other argument to establish such a difference? It does not seem so. To be sure, he claims that "co-opt[ing]" and thus using the "physical resources [of a person] to help me advance my goals when she would not be required to do this … is normally wrong … in virtue of the fact that a person is entitled to determine not only which ends to pursue but also which ends to use her body in service of," and then contrasts this with "harming others as a side effect."[19] However, this alleged contrast is supposed to have been explained by the alleged close relationship discussed above. Since this close relationship is fictitious, it is not surprising that the same is true for the supposed contrast. After all, *discarding* the physical resources of a person in the process of advancing my goal when she would

[18] Ibid.
[19] Ibid.

not be required to do this is also "normally wrong" in virtue of the fact that a person is entitled to determine not only which ends to pursue but also in pursuit of which ends her body is to suffer adverse side effects. To put this differently: throwing other people off bridges without their consent is "normally wrong," period. It certainly does not become *better* whenever throwing them off the bridge would serve no purpose whatsoever.

Tadros, however, claims, regarding side effects:

> I am not normally required to show the person would have an enforceable duty to serve the end that I am pursuing at the relevant cost. As long as the costs that I impose are proportionate to the importance of my goal, I need not establish that the person who bears the costs either does or must share my goal. V would not be required to turn the trolley away from the five toward himself in *Trolley*, for example.[20]

Again Tadros assumes a contrast here that does not exist. For exactly the same holds for an act involving the *use* of a person—if the act is proportionate and necessary, then it is justified. (Tadros admits that harming as a means can sometimes be justified, namely when the stakes are high enough.)[21] This is the very idea of a lesser evil justification, and to the extent that these two acts are justified at all, both the killing of the man on the side track in *Trolley* and the killing of the man on the *Bridge* in order to save the five can only be justified by a lesser evil justification. Thus the question whether side-effect killing is more difficult to justify than killing as a means is precisely a question about *proportionality*. Is it more difficult for killing as a means to be proportionate than for killing as a side effect? Accordingly, is it really *true* that in the case of the side-effect harming of a person "I am not normally required to show the person would have an enforceable duty to serve the end that I am pursuing at the relevant cost," while in the case of harming as a means I am so required? That it is true must be shown by argument; it cannot simply be assumed, but assuming it is all that Tadros does here.

Tadros's claim that the man on the side track would not be required to turn the trolley toward himself is of no help. It is the very same mistaken argument Quinn made about cooperation. The mistake lies in the fact that if the side-track man could stop the agent about to divert the trolley, then the side-track man's cooperation would be required. In fact, Tadros—unlike Quinn—does require the cooperation of the side-track man. According to him, this person "is not permitted to avert the threat at all."[22] It is therefore surprising when Tadros nevertheless claims that the rescuer "cannot be accused of imposing an end on V in *Trolley*, even the end of ensuring that V does not prevent D from saving the five, because D does not need V to serve any end in order to rescue the five."[23] This depends. If V could prevent the would-be rescuer of five and killer of one from killing him, and is, according to

[20] Ibid.
[21] Tadros (2011, 211).
[22] Ibid., 203.
[23] Tadros (2015, 73).

Tadros, required not to do so, then he clearly is required to serve the ends of the rescuer, and thus the rescuer imposes his goals on V by initiating the rescue. V would serve those goals by omission, but he would serve them. Of course, if V actually does not have sufficient means of defense, he need not serve the goal. But neither need people used as *means* serve a goal, as we already saw. To be passively *used* is simply not the same as to actively *serve*. The latter requires adopting other people's goals, and the former does not.

Thus, the very idea—that people are permitted to set their own ends and are not required to serve the ends of others—that would speak, according to Tadros, against using persons as means would also speak against requiring them not to prevent others from harming them as a side effect. In other words, as long as Tadros maintains that the side-track man must not interfere with the rescuer's efforts, he cannot also maintain that a person's entitlement to choose her own ends can distinguish the two cases. His position is incoherent.[24]

Yet Tadros's assumption that persons must have an obligation to serve a given end for it to be permissible to use them in the service of the end, while they need not have such an obligation to make it permissible to harm them as a side effect of the pursuit of that end, can be considered independently of the incoherence just mentioned. Obviously, however, it has to be considered not by comparing apples to oranges, for example *Trolley* to *Bridge.* To wit, in *Trolley*, the victim is simply run over by a train. In *Bridge*, in contrast, the victim is *non-consensually touched*, *kinetic force* is applied to him, he is *moved against his will*, *falls* from a bridge, *crashes* on the ground, and only *then* he is finally run over by the train. Offering such pairs of examples and our intuitions regarding them as "evidence" for the DDE or related principles is philosophically unhelpful, for the normative significance of the additional differences I noted should be obvious: non-consensual bodily contact is often considered to be already offensive by itself, and so is being man-handled (that is, being moved against one's will), the application of kinetic force will normally be painful or at least unpleasant, falling from a bridge will certainly cause anxiety, and crashing on the ground will hurt. All that does, indeed, seem to be significant. One might object here, however, that falling to one's death need not be more horrifying than being attached to the track and seeing the train hurtling down the track towards one. True, but this objection overlooks the fact that the man on the bridge is not *accidentally* falling but *made* to fall. Yet the difference between *Trolley* and *Bridge* evaporates if we keep this factor equal, that is, if the person diverting the trolley *actively traps* the man on the side track foreseeing that he will then be hit by the trolley—in fact, *Trolley* might now seem even *worse* than *Bridge*. Thus, comparing *Trolley* with *Bridge* might be *rhetorically* effective if it comes to motivating the DDE; however, it is philosophically not only useless but entirely misleading.

Thus, one must choose examples where everything else is equal. So let us compare *Trolley* with *Sensor Trolley*.[25] In *Sensor Trolley*, the rescuer can only divert

[24] Steinhoff (2014).

[25] I am, obviously, not against appealing to intuitions as such; I am merely against appealing to intuitions that are produced in methodologically inapt ways.

the trolley by pointing a sensor at the body of the man on the side track. Thus, the man's body is used as a means to divert the trolley.[26] Does this make any difference? Hardly. Intuitively, diverting the trolley by pointing the sensor at him is permissible, even though the man on the side track is not obliged to divert the trolley himself by pointing the sensor at himself. Of course, pointing a sensor at the person is not in and of itself a harm, but it does harm him because it causes him harm, and this is what matters on Tadros's own causal interpretation of the "*Using View*."[27] In addition, compare also *Robot Bridge 1* and *Robot Bridge 2*. In both examples, a runaway trolley threatens five people. In the first variation, the trolley can be stopped by pushing a red button on an instrument that has nothing to do with the normal operation of the trolley. As a side effect, a robot will be activated and throw the fat man off the bridge in front of the train, killing him. In the second variation, the trolley can only be stopped by having the robot throw the fat man in front of the train, which is achieved by pushing the red button. As far as my intuitions are concerned, these are both cases of impermissible killing. They would both become permissible if the number of people threatened by the trolley become large enough, but they would both become permissible at the same number and without it thereby becoming obligatory for the two fat men to press the buttons themselves.

In any case, Tadros has not given any *rationale*, any *explanation* as to why there should be a difference. As we saw, his supposed explanation in terms of a person's right to choose her own ends applies to *both* kinds of cases, to side-effect harming and to harming as a means. Being bombed by others in the pursuit of their interests contravenes one's own ends no less when it is useless to those others than when it is useful to them. And yes, harming a person as a side effect in pursuit of one's ends is not akin to compelling her to act in service of that end. But *neither* is harming her as a *means* to pursue that end. A person whose body is being used without her consent simply need not adopt other people's goals. Thus we still have not encountered any explanation as to why *using* people by throwing them off bridges or by blowing them up should be worse than *discarding* people by throwing them off bridges or by blowing them up. Given that the whole idea seems intuitively so bizarre—maybe we should stop looking.

5 Sarch's "Two Strikes Argument"

Yet people continue to look. Sarch is rightfully skeptical about the DDE making a difference for *permissibility*. Yet he claims that it does make a difference for *culpability*.[28] Sarch compares two cases of arson. In the first case, Alan is paid to burn

[26] Could Tadros simply refuse to "count" this as an instance of using as a means? Well, if pointing a sensor at a barcode in order to activate something amounts to using the barcode as a means—and it certainly does, at least in ordinary language—then pointing a sensor at a person to activate something likewise amounts to using the person as a means. Unless Tadros gives a technical definition of "using as a means" (and he has not), we are justified in taking him to be using the term as it is used in ordinary language.

[27] Tadros (2015, 57).

[28] Sarch (2017a, 458–461).

down a building, and he indeed burns it down in order to get the money, foreseeing with certainty the death of a victim who happens to be in the house. He regrets the victim's death, but the money is more important to him. In the second case, Bobby is paid to kill the victim (he would not get the money if the victim survived) by burning down the house, and in order to get the money he indeed kills the victim by burning down the house. He regrets the victim's death, but the money is more important to him.[29] Sarch claims that "it … seems that what Bobby did (intentionally kill without justification) is more culpable than what Alan did (knowingly cause death without justification)."[30] It does not seem so to me, and given that it is often thought that criminal law at least in Western democracies expresses widely shared moral intuitions and Sarch himself admits that "the difference between these cases does not matter from the legal perspective, since both Alan and Bobby would be guilty of murder,"[31] I have severe doubts that many people would share Sarch's intuition. In any case, Western jurisdictions do *not* share it.[32]

But let us set intuitions aside and consider whether Sarch can actually offer a convincing rationale, any explanation as to why Bobby's act should be worse than Alan's. Sarch's explanation relies on the "*Insufficient Regard Theory*," that is, on the premise "that one is culpable for an action to the extent it manifests insufficient regard for the interests of others (or perhaps more generally, for morally relevant interests)."[33] "[T]he rationale for DDE_{NACR} [his particular version of the DDE] is that intending … a given harm manifests more insufficient regard for others, all else equal, than merely foreseeing that one's action will cause that harm, without being committed to it."[34] This is so, says Sarch (and this reminds one of Quinn's approach) because someone intending to cause harm shows insufficient regard in not just one but two ways:

> Beyond being insufficiently repelled by the harm, you also display the further fault of taking it that there is a positive reason in favor of promoting the harm. That is, your act demonstrates that promoting the harm is something to which you are affirmatively attracted more than you ought to be, assuming the harm is unjustified. Since acting with the commitment to harm involves two manifestations of insufficient regard, while merely knowing or foreseen harm

[29] Ibid., 462.

[30] Ibid., 478.

[31] Ibid., 463.

[32] Sarch rightly points out that *some* crimes are legally defined with reference to intent or purpose. However, all the examples he gives (ibid., 456) are examples where *without* purpose there is *no crime at all*—which certainly does not correspond to the situation in the Alan/Bobby case—and where it seems to be rather clear that if there is a *moral* failure with intent in these cases, then there is also a moral failure with knowledge (consider, for instance, his example of falsely incriminating another). In other words, it would appear that in these cases—unlike in cases involving killing or otherwise physically hurting people—law does not even make the *attempt* to track morality but seems to be guided by other, perhaps pragmatic or evidentiary, concerns. That, in my view, severely undermines the probative value of *these* cases for our moral intuitions.

[33] Ibid., 465.

[34] Ibid., 466.

> involves only one, the insufficient regard theory entails that there is a respect in which the first actor will be more culpable for his conduct than the second is for hers.[35]

This is a strange "rationale" for the DDE. After all, according to Sarch, "X's doing A promotes p just in case it increases the likelihood of p relative to … the status quo."[36] Alan, however, increases the likelihood of the victim's death no less than Bobby, and he does so not on a whim but for a reason: he gets paid for it. Thus, *both* agents think that there is "a positive reason in favor of promoting the harm." *Both* of them commit *two* mistakes.[37]

This can also be seen by considering Sarch's statement that his "only claim is that in cases where the killing does remain instrumentally necessary to Bobby's getting paid … Bobby would feel motivational pressure to take steps to make Victor's death more likely."[38] However, that is only the case if Bobby believes that killing is instrumentally necessary to getting paid. But, likewise, when Alan believes that the killing is a necessary side effect of any course of action that will secure his getting paid, then he will feel motivational pressure to make Victor's death more likely, because he believes that any outcome without Victor's death will also be an outcome without him, Alan, getting paid. So *both* Bobby and Alan are committed to Victor's death: they both know that, in the actual world, they will not get paid unless Victor is dead.[39] Bobby knows that Victor's death is unavoidable for instrumental (that is, causally upstream) reasons and Alan knows that it is unavoidable for collateral (that is, causally downstream) reasons; and while both might in this sense be committed in different ways to the promotion of the harm, they are *equally* committed, *equally* attracted, and *equally* non-repelled. Accordingly, the commitment involved cannot explain the alleged difference between instrumental killing and side-effect killing.

Can Sarch's rationale be saved by reformulating it? As we saw, the talk about "promoting harm" does not help Sarch's case. But could we not say that Bobby, beyond being insufficiently repelled by the harm, also displays the further fault of taking himself to have a positive reason in favor of *the harm itself* (instead of merely promoting the harm)? Well, we simply cannot. As Sarch describes the case, it is very clear that neither Bobby nor Alan are attracted by the harm *itself*—if they were, they would not *regret* Victor's death but welcome it.[40]

---

[35] Ibid.

[36] Ibid., 464. Sarch affirmatively quotes here Schroeder (2007, 113).

[37] I have encountered the objection that from the fact that Alan thinks that he has a reason to burn down the house it does not logically follow that he thinks that he has a reason to promote harm. Well, that is true. However, unless Alan is demented (and I thought we were talking about rational actors), he will *know* to burn down the house in the example *is* to promote harm, but then he cannot think that he has a reason to burn down the house without also thinking that he has a reason to promote harm.

[38] Sarch (2017a, 465).

[39] This is also how Bennett would analyze the situation (1981, 101, point 2).

[40] To be sure, Sarch claims that if you do A with a commitment to a certain harm, then you also feel "some motivational pressure to affirmatively promote" the harm (2017a, 455). As I have already explained, however, both Bobby *and* Alan feel motivational pressure to affirmatively *promote* the harm—they will not get paid unless the harm ensues. They feel no motivational pressure, however, to celebrate the harm *in itself*.

Moreover, even if one accepted, for the sake of argument, the "harm itself" interpretation, this would only bring us right back to Quinn's arbitrary and question-begging suggestion regarding the counting of morally objectionable "presumptions." While Quinn conveniently overlooks presumptions that could *also* be counted but would not deliver the sum total he prefers, Sarch overlooks certain *manifestations of repulsion* that could also be counted. To wit, Alan manifests his disregard by being insufficiently repelled by the harm but he *also* manifests his disregard by being insufficiently repelled by the *uselessness* of the harm. Bobby, in turn, manifests his disregard by being insufficiently repelled by the harm and by being positively attracted to the useful harm. And again we have *two* manifestations on both sides. I see no non-question-begging way to show that one way of counting is more appropriate than the other, and if there is one, Sarch has certainly not shown it.

Yet I have come across the objection that I fail to explain why the uselessness of something should repulse an actor in its own right rather than simply consist in the mere absence of an affirmative basis for being attracted to the useless thing, and that it is also unclear how it could be true, as I allegedly assume, that one should be more repelled from (a) a harm with reasons equaling −10 against and +0 in favor, than one should be from (b) the same harm, with the same reasons for and against, which also is useless. Well, I do not assume the latter, since what the objection presupposes here is actually mistaken: that the harm has no reasons (+0) in its favor *means* that it is useless, so it makes no sense to say that something has no reasons in its favor *and* is *also* useless. What I am suggesting, instead, is the possibility that the uselessness of a harm could be an additional reason against inflicting it, so that, to illustrate, a harm with reasons −10 against and +1 in favor might become a harm with −11 or more against once the +1 in favor is removed. And I already motivated this possibility above, when discussing Quinn, pointing out that one often hears something like this: "Let's win the war so that our soldiers did not die in vain." Dying in vain, it seems, is considered something particularly bad. Last but not least, note the unwitting irony of the objection: allegedly, to repeat, I fail to explain why the uselessness of something should repulse an actor in its own right rather than simply consist in the mere absence of an affirmative basis for being attracted to the useless thing. However, a defender of the DDE makes a similar but by far stranger assumption: he assumes that the *usefulness* (+1) of something should *repulse* an actor in its own right rather than provide an *affirmative* basis for being attracted to the useful thing. If anything, it is this latter assumption that appears to be unmotivated, if not downright absurd, not the assumption that I suggested as a possibility.

Finally, the whole talk about manifested disregard is somewhat unclear. Is the thesis that one is more culpable the bigger the *manifestation* of disregard is, or the bigger the *disregard manifested* is? The latter thesis would certainly be significantly more plausible than the former. In any case, if it is the former, one would need to know how one measures the size of a manifestation of disregard. Sarch suggests that the *more* manifestations there are, *the bigger* the overall manifestation is. But that is clearly wrong. Showing another person one's tongue *and* one's middle finger are two manifestations of disregard, while shooting him in the head is only one. Yet the latter case seems to be a bigger manifestation of disregard than the former. If it is not, then "manifestations of disregard" can hardly matter morally. Alternatively,

Sarch might think that *all else being equal* more is bigger, so that showing someone both one's tongue and one's middle finger would be a bigger manifestation of disregard than just showing him one's tongue. Even if that were true (and I doubt it is), it has to be noted that Bobby and Alan manifest their disregard *in exactly the same way*, namely by their *act* of arson. Accordingly, there is the same number of manifestations in this case, namely *one*. Even if Bobby, as Sarch claims, displays "the further fault of taking it that there is a positive reason in favor of promoting the harm," this would be neither here nor there, since taking something as a reason might be a *form* of disregard but not its *manifestation*. In fact, Sarch himself stresses that it is *conduct* or *acts* that are at issue as manifestations of disregard, stating, among other things, that "it is a fundamental principle of the criminal law that we do not punish merely for bad attitudes or character traits one might possess, but only for conduct that manifests them."[41]

If, alternatively, the thesis is that one is more culpable the bigger the manifested *disregard*, it should be noted that Sarch states "that the amount of insufficient regard *manifested* by an action is equal to only the *minimum* amount that it is necessary to postulate in order to explain why the actor behaved as she did under the circumstances."[42] Yet to claim that in order to explain Bobby's act we need to "postulate" *more* disregard than in order to explain Alan's act would itself be nothing but a question-begging postulate, and an implausible one at that. *Why* should someone who gets paid to burn down a house and does so even though he is certain that he will thereby kill an innocent person necessarily manifest (let alone have) more disregard toward that person than someone who gets paid to kill the person by burning down the house and therefore burns down the house? To be sure, Sarch's "*Two Strikes Argument*"[43] is meant to answer precisely this question, but since it fails—there are, as we saw above, the same number of strikes on both sides, unless we use one-sided counting, which proves nothing—the source of the alleged difference in the amount of manifested disregard remains entirely mysterious. Indeed, this mysteriousness is so profound that it would give us reason to doubt the *Two Strikes Argument* even if we had not already seen *why* it is wrong.

I conclude that Sarch has failed to provide a rationale or explanation for the DDE.

6 Øverland and Walen on "Moral Obstacles" and "Restricting Claims," Respectively

Walen and Øverland have recently, independently of each other, offered two very similar accounts that are supposed to "transcend" (Walen) or provide an "alternative" (Øverland) to the means principle and the DDE, respectively. How plausible are these new accounts and the rationales offered for them?

[41] Ibid., 472–473. Sarch (2017b) further elaborates on this.
[42] Sarch (2017a, 466).
[43] Ibid., 467.

The basic idea, in the case of Walen, is that it is "*easier* to justify causing or allowing harms to those with restricting claims" than to those with non-restricting ones[44]; and according to Øverland there is a "reduced constraint against harming individual moral obstacles" as compared to harming people who are not "moral obstacles."[45] To understand these claims, one needs to know what restricting and non-restricting claims and moral obstacles are. Here is Øverland:

> *Moral Obstacle*: When an innocent person A is under threat of harm and has a defensive action available to her (independent of the presence of other people), another innocent person B, who poses no threat or physical hindrance, is a moral obstacle if his presence has the consequence that either (i) B will be harmed (if A performs her available defensive action) or (ii) A will be harmed (if A completely restrains her defensive action), or (iii) both A and B will be harmed in some measure (if A partially restrains her defensive action).[46]

To illustrate, Øverland considers a case where an innocent person A can only defend herself against an impermissible attack by some person C if she kills the attacker with a flamethrower.[47] However, if A kills C with the flamethrower, she will also kill the innocent and non-threatening person B standing behind C. Thus, she could only avoid killing B—and killing an innocent person is certainly a moral cost—by foregoing her defense against C's impermissible attack. B, therefore, is a moral obstacle to A's defensive action against C. In contrast, in a case where A can only defend herself against C by using another person B as a means, for example, by shooting her so that B falls from his scaffolding and crushes C standing below,[48] B is *not* a moral obstacle. The decisive difference between the two cases lies in the phrase in parenthesis "(independent of the presence of other people)": in the first example, A could defend herself with the flamethrower if B were not present, while in the second example A's defensive option *depends* on B's presence.

Here is Walen:

> [R]estricting claims, if respected as rights, would restrict an agent from doing what she could otherwise permissibly do for herself or others if the claimants were absent, or would require her to take an action that would make others worse off than if the claimants were absent; non-restricting claims, if respected as rights, would not in that way restrict an agent or require her to take an action that would cause others to be worse off than if the claimants were absent.[49]

This distinction seems to be able to account for the normative difference between *Trolley* and *Bridge*. In *Trolley*, respecting the right to life of the innocent bystander

[44] Walen (2014, 433).
[45] Øverland (2014, 491).
[46] Ibid., 486.
[47] Ibid., 484.
[48] Ibid., 485.
[49] Walen (2014, 446). It should be noted that Walen does not believe in permissible rights-infringements—he takes rights to be absolute. That is, to respect A's claim not to be killed as a right implies for Walen that one must not and will not kill A. See ibid., 440, n. 27.

would restrict the agent from doing what she could otherwise permissibly do, namely divert the trolley to the side track; in *Bridge*, however, respecting the right to life of the bystander on the bridge would not restrict the agent from doing what she could otherwise permissibly do, namely stop the trolley, for in that case she can only do that *because* of the presence of the bystander.

The underlying distinctions are clearly very similar (one could say that Øverland's moral obstacles have Walenean non-restricting claims). The *rationales* provided by Øverland and Walen are also similar. According to Øverland, "being a moral obstacle is typically relevant to determining whether they can be harmed permissibly because their presence typically gives rise to cost." The "crux" of the proposal is that "fairness dictates that those under threat should not be the only ones required to bear cost as a consequence of the presence of moral obstacles; the moral obstacles themselves should also bear some cost."[50] Walen, in turn, states that "[i]f a patient's being present with a claim on an agent makes others worse off than if he were not present, then his being present imposes something like a negative externality on them." Therefore, the strength of such a restricting claim must "reflect its impact on others so that it is not excessive."[51]

To their credit, both Øverland and Walen mention a potentially devastating counter-example to their accounts, namely a case where the rescuer can turn the trolley away from the five and towards the one on the side track only because of the very presence of the one. Somehow the switch would not work if the one were not present.[52] Øverland admits that the man on the side track gives rise to no cost in this case.[53] Yet this leads to a counter-intuitive result, namely that the man on the side track should be treated like the man in *Bridge* so that harming him by turning the trolley would be impermissible (or as difficult to justify as harming the man in *Bridge*).[54] Øverland says that he accepts this result "[o]n reflection,"[55] yet he does not report the contents of his reflection; that is, he provides no argument. Thus it seems that he is simply biting the bullet because he is already convinced of the theory of moral obstacles. This, however, will not move those who are not yet convinced. In fact, the example will in all likelihood induce them to reject the theory.

Walen seems to be aware of this danger and tries to explain the example away, or at least to mitigate its impact. He claims that this case "can be assessed in two ways." Allegedly it can be "analogized" to the *Bridge* case,[56] or it can be "analogized" to the original *Trolley* case.[57] Yet analogies are a mere distraction where we have actual implications. The question is what Walen's own official definition of non-restricting claims *implies*. And as Walen rightly notes about the man in the revised *Trolley* case

[50] Øverland (2014, 483–484).

[51] Walen (2014, 438).

[52] Walen (2014, 457). Øverland provides exactly the same example in "Giving Rise to Cost," ms. on file with the author, 117–118. He provides a similar example in Øverland (2014, 498).

[53] Ibid. See also "Giving Rise to Cost," 117–118.

[54] Ibid., 118.

[55] Øverland (2014, 498).

[56] Walen calls *Bridge* "*Massive Man*."

[57] Walen (2014, 457).

we are discussing now: "If he were absent, [the person pulling the switch] would not be able to save the five, and thus his claim not to be hit should count as non-restricting."[58] In fact, his claim not only "should count" as non-restricting, it simply is non-restricting by way of logical implication.

Yet Walen suggests a second interpretation of the case. According to this interpretation "the fact that he makes the switch work … is simply a given of the situation." And Walen states:

> The only thing the agent can do is take advantage of that fact or not. … [G]iven that she would take advantage of a fact that is true regardless of what she chooses, she should regard his claim as concerning not the fact that he plays a role in making the switch work, but only the harm that would befall him if she uses the switch. In other words, his claim should be taken to be restricting.[59]

First, what an agent guided by Walen's non-restricting claims principle should do is to take the implications of that principle seriously, and we already saw what those implications are. Second, *if* the agent in the revised *Trolley* case is permitted to simply take the fact that pulling the switch will save the five as "a given," ignoring the fact that it is given *by* the very presence of the man, why then should the agent in the *Bridge* case not also simply take the fact that making a pushing motion in a certain space–time continuum will save the five as a given? Both the pulling motion at the switch and the pushing motion on the bridge only divert or stop the trolley because of the fat man, and if one may ignore this in one case, one may ignore it in the other. In other words, Walen's way of dealing with these cases is arbitrary. This arbitrariness is also shown in the fact that the official rationale of the restricting claims principle, namely that the presence of those with restricting claims "imposes something like a negative externality" on other agents, is ignored in Walen's second interpretation of the revised *Trolley* case. Whether a person imposes negative externalities or not is a causal question. And the man on the side track does not impose a negative externality, since he causes the switch to work in the first place. To treat him as imposing a negative externality by ignoring the fact that one can turn the trolley only *because* of him is like treating a philanthropist's donation as a negative externality by taking the presence of the money in one's charitable account as "a given" and then deploring the burdens of having to be grateful.

Meanwhile Walen has admitted that his previous account begged the question,[60] and succumbed to counter-examples.[61] Allegedly, his recent reformulation of his RCP can overcome these problems. He states: "If a patient's claim pushes to restrict an agent's baseline freedom, and thereby pushes to make the agent or others worse off, then it is restricting"—otherwise not. Again, restricting claims are weaker, and non-restricting ones stronger.[62] The decisive revision lies in replacing

[58] Ibid.
[59] Ibid., 457–458.
[60] Walen (2016, 226).
[61] Ibid., 222–225.
[62] Ibid., 214.

the "counter-factual baseline" (referring to what would be the case if the patient were not present) with the "toolkit baseline":

> The alternative baseline is framed in terms of the things in the world that an agent can and cannot take herself to have a baseline freedom to use – *using* being particularly important to agency. If a patient owns what the agent wants to use, and if her baseline freedom does not for that reason include the use of that thing (including the patient's body), then the owner-patient's claim would be non-restricting. Rather than restricting the agent relative to her baseline freedom, the patient's claim sets the limit for what that freedom is. If, however, the things the agent needs to use are hers or otherwise available for her to use – if they are part of her 'toolkit' for action – then a patient's claim not to be harmed or to be helped pushes to restrict her relative to that baseline freedom. It counts, against this baseline, as restricting.[63]

Walen then claims that the new baseline "easily handles" the counter-examples.[64] Yet, it does not. To see this, note first that Walen has replaced the original counter-example with a new one, namely this:

> *Sidetrack Man Protecting Others*: An agent at a switch can throw the switch and thereby turn a trolley that would otherwise kill five innocent people onto a sidetrack where it will kill the sidetrack man. The sidetrack man's weight is sending a signal to another switch; if he were not there, then any trolley turned away from the track with the five would be sent to a third track where, as it turns out, ten people would be killed.[65]

Second, Walen has the side-track man in the famous *Loop* case (where the body of the man on the side track keeps the trolley from looping back to the five) say the following: "If my being hit would not stop the trolley from hitting the five, you would clearly have no reason to turn it; indeed, it would be impermissible to turn it because you would just add my death to theirs. Thus you must be relying on my serving as the means of stopping the trolley to justify turning it. But again, I am not in your toolkit." And Walen adds: "This response is, I believe, convincing."[66]

Obviously, however, the *Sidetrack Man Protecting Others* can say pretty much the same as the side-track man in *Loop*: "If my body's weight would not send the signal, you would clearly have no reason to turn the trolley; indeed, it would be impermissible to turn it because you would kill more people than you save. Thus you must be relying on my serving as the means of keeping the trolley from redirecting to ten people to justify turning it. But again, I am not in your toolkit." This reply is no less convincing than the reply of the side-track man in *Loop*. Accordingly, Walen's

[63] Ibid., 225.
[64] Ibid.
[65] Ibid., 222.
[66] Ibid., 243.

revised principle still does not work, since intuitively, as Walen admits, turning the trolley in this case is justified.[67]

Finally, Walen himself admits "that reframing the RCP in terms of the toolkit baseline may seem to be an exercise in circular reasoning," since "it may seem that its explanation of the MP [Means Principle] illicitly presumes the truth of the MP."[68] Yes, it indeed looks suspiciously as though a "tools principle" is now being sold as an "explanation" of a "means principle," but exchanging one word for the other is not really an explanation at all. Yet Walen thinks that such a reading is "superficial."[69] He might be correct since the problem is perhaps less that RCP has a circular relation to the MP and more that it has no relation to it at all, in particular no "explanatory" one. We can see this by considering Walen's objections to the circularity charge.

Walen says, first, that the RCP "has an explanatory structure that is different from the MP" since "it focuses on explaining why claims not to be harmed as a side effect are relatively weak, as well as on why claims not to be harmed as a means are relatively strong; and it explains why restricting claims are weaker than non-restricting claims by reference to a global balance of claims on an agent."[70] Actually, however, normative *principles* (and that is what the "P" in "RCP" and "MP" stands for) have no "explanatory structure" whatsoever; instead, they simply make certain normative *claims*. To wit, the means principle states that harming as a means is worse than harming as a side effect, but it does not give any *explanation* as to why that is the case. At best a philosophical theory *underlying* the means principle would give such an explanation. Likewise, Walen's RCP, in its revised form, simply states that "[i]f a patient's claim pushes to restrict an agent's baseline freedom, and thereby pushes to make the agent or others worse off, then it is restricting and should be considered weaker than it otherwise would be. If it does not push to restrict her baseline freedom in that way, then it is non-restricting and should be considered stronger than an otherwise analogous restricting claim."[71] Obviously, first, there is nothing in this—or any other—statement of the RCP that *explains* why restricting claims "should be considered" weaker than non-restricting ones. Moreover, second, since the principle does not even *mention* the means/side-effect distinction, it can hardly "explain" its normative significance.

One might object here that while, strictly speaking, the RCP itself does not explain the MP, at least Walen's underlying philosophical *theory* explains the MP

[67] Ibid., 223. Of course, the side-track man in the original counter-example could not argue that the turning of the trolley is only *permissible* because of his presence, but he could still argue that it would not be *possible* without his presence. So it would appear that the original counter-example cannot be handled by Walen's revision either: the side-track man is still used as a tool, and it is unclear—at least to me—why the "toolkit baseline freedom" of the agent should include the option of diverting trolleys to tracks where people without whom the turning of the trolley would not have been possible to begin with will be killed by them.

[68] Walen (2016, 240).

[69] Ibid.

[70] Ibid.

[71] Ibid., 214. This is Walen's definition of the revised RCP.

under an appeal to the RCP. Yet as shown a few paragraphs ago, the RCP still cannot handle certain counter-examples. It is simply mistaken. Moreover, given that, as we saw, the agent in *Sidetrack Man Protecting Others* is, *pace* Walen, relying on that man's serving as a means of stopping the trolley although the man is not in the agent's toolkit and thus *has* a restricting claim, but could nevertheless, even *according* to Walen, be killed justifiably and as a side effect, Walen's RCP actually *contradicts* the means principle, and consequently Walen can hardly explain the latter with the former.

Second, Walen claims that "those differences [between restricting and non-restricting claims] reflect the fact that it [the RCP] relies on the idea of a patient analog to an agent imposing a negative externality—an idea … with great explanatory force."[72] As the still unrefuted original counter-example and the *Sidetrack Man Protecting Others* show, however, there is no such explanatory force: in these counter-examples, the patient does *not* impose any negative externality but his claim not to be harmed is *still* weak, not strong.

Third, according to Walen, "the ground for using those ideas is *not* an appeal to the MP; it is an appeal to fundamental principles that shape the normative space of rights," namely "that agents have a fundamental right to pursue their own ends, that property rights are essential for their doing so, that agents must therefore conceive of the world's resources as divided between them, that the resources they have play an important role in defining their baseline freedom, and that patient-claims that do not reflect the agent-based division of the world have to register in a different way—this last point being the agent-patient divide."[73] Unfortunately, it is almost painfully obvious that these "fundamental principles" cannot be used to "ground" the claim that harming as a means is worse than harming as a side effect. After all, my fundamental right to pursue my own ends is not only infringed by destroying my property, including my body, as a means to an end, but also by destroying it as a side effect. Broke is broke, and dead is dead. Thus the appeal to such ideas, including the "agent-patient" divide, does not explain why harming as a means should be worse than harming as a side effect. Said ideas are entirely neutral on that point.

Fourth, Walen claims that his "account in terms of the agent-patient divide allows us to make sense of the moral significance of a patient's causal role, one of the mysteries undermining the plausibility of the MP. On the revised RCP, the relevance of a patient's causal role arises out of the relevance of an agent having the basic freedom to use her toolkit—something with obvious moral relevance."[74] Yet while it might be true that a patient's causal role might be sometimes relevant (I think it is), the question is whether it is worse to harm someone in a way that is causally useful for the achievement of one's goals than to harm him in a way that is causally useless for the achievement of one's goals. And, again, Walen has simply not shown that—as demonstrated by the previous point and the counter-examples, including the counter-examples already presented in Sect. 4, like *Car*, *Sensor Trolley*, and

[72] Ibid., 240–241.
[73] Ibid., 241.
[74] Ibid.

Robot Bridge[75]—nor has anybody else. Again, if anything, one would assume that uselessly throwing persons off bridges is *worse* than usefully doing so, not the other way around.

Thus, I conclude that both Øverland's moral obstacles principle and Walen's non-restricting claims principle might well "transcend" the DDE and the means principle, but they have severe problems of their own and are not supported by credible rationales.

7 Conclusion

The overall conclusion is that we are still left without any plausible rationale for the DDE and related principles.[76] This would not be a sufficient reason to stop looking for such rationales if only we had some independent reason to believe that they are normatively relevant. Yet we do not have such reasons. Testing our intuitions regarding pairs of cases that differ in *many* respects, like *Trolley* and *Bridge*,[77] is simply useless. Instead, one has to use examples that do keep all else equal, such as Sarch's arson examples. Yet such examples actually *fail*—somewhat subdued protests of defenders of the DDE notwithstanding—to elicit the intuitive responses on which defenders of the DDE or of related principles rely. Given that the "rationales" that have been offered for the DDE and related principles are, as we saw, arbitrary, contrived, and unsuccessful, this is not surprising. It is time to give up the wild goose chase.

Acknowledgements The research presented in this paper is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 17612817). I am very grateful for this support. I also thank an anonymous reviewer for very useful comments.

[75] Walen relies on the same methodologically inadequate examples—in particular *Trolley* and *Bridge*—that are also preferred by virtually all other defenders of the DDE, the means principle, or related principles (Sarch being the noteworthy exception—but his example does not elicit, at least not from Western jurisdictions, a response supporting the DDE). I have already explained that such examples prove nothing; see the last three paragraphs of Sect. 4. See also note 77.

[76] I did not make the attempt here to discuss every idea that has been suggested, but only some prominent and at least intelligible ones. Kamm (2008, 145–146 and 162–167), for instance, seeks to "explain" some of her deontological principles by an appeal to terms like "subordination" and "substitution." Some authors have criticized her distinction, which presupposes that they find it intelligible. I myself have to admit that I agree with Nye (2014, 449–450), who deems the distinction to be obscure. In any case, it is neither possible to discuss every suggestion in one article nor necessary in order to show that the "rationales" provided for the DDE tend to be unsatisfactory.

[77] The other usual suspects, like *Hysterectomy/Craniotomy* or *Tactical Bomber/Terror Bomber*, do not fare better. For a critique of the latter, see Di Nucci (2014, 177–187). For a recent general complaint about confounding factors in the typical hypotheticals employed by defenders of the DDE see also Cushman (2016).

References

Bennett, J. (1981). "Morality and Consequences." In S.M. McMurrin (Ed.), *The Tanner Lectures on Human Values II*. Salt Lake City: University of Utah Press.

Bennett, J. (1998). *The Act Itself.* Oxford: Oxford University Press.

Cushman, F. (2016). "The Psychological Origins of the Doctrine of Double Effect." C*riminal Law and Philosophy* 10(4): 763–776.

Di Nucci, E. (2014). *Ethics Without Intention.* London and New York: Bloomsbury.

Kamm, F.M. (2008). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

Nagel, T. (1980). "The Limits of Objectivity." In S.M. McMurrin (Ed.), *The Tanner Lectures on Human Values I*, 77–139. Salt Lake City: University of Utah Press.

Nelkin, D.K., and S.C. Rickless (2014). "Three Cheers for Double Effect." *Philosophy and Phenomenological Research* 89(1): 125–158.

Nelkin, D.K., and S.C. Rickless (2015). "So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short." *Nous* 49(2): 376–409.

Nye, H. (2014). "On the Equivalence of Trolleys and Transplants: The Lack of Intrinsic Difference between 'Collateral Damage' and Intended Harm." *Utilitas* 26(4): 432–479.

Øverland, G. (2014). "Moral Obstacles: An Alternative to the Doctrine of Double Effect." *Ethics* 124(3): 481–506.

Quinn, W.S. (1989). "Actions, Intentions, and Consequences: The Doctrine of Double Effect." *Philosophy & Public Affairs* 18(4): 334–351.

Ramakrishnan, K.H. (2016). "Treating People as Tools." *Philosophy & Public Affairs* 44(2): 133–165.

Sarch, A. (2017a). "Double Effect and the Criminal Law." *Criminal Law and Philosophy* 11(3): 453–479.

Sarch, A. (2017b). "Who Cares What You Think? Criminal Culpability and the Irrelevance of Unmanifested Mental States." *Law and Philosophy* 36: 707–750.

Schroeder, M. (2007). *Slaves of the Passions.* Oxford: Oxford University Press.

Steinhoff, U. (2014). "Why We Shouldn't Reject Conflicts: A Critique of Tadros." *Res Publica* 20(3): 315–322.

Tadros, V. (2011). *The Ends of Harm: The Moral Foundations of Criminal Law.* Oxford: Oxford University Press.

Tadros, V. (2015). "Wrongful Intentions without Closeness." *Philosophy & Public Affairs* 43(1): 52–74.

Walen, A. (2014). "Transcending the Means Principle." *Law and Philosophy* 33(4): 427–464.

Walen, A. (2016). "The Restricting Claims Principle Revisited: Grounding the Means Principle on the Agent-Patient Divide." *Law and Philosophy* 35(2): 211–247.

Wedgwood, R. (2011). "Defending Double Effect." *Ratio* 24(4): 384–401.