Ethical Theory and Moral Practice (2018) 21:735–742
https://doi.org/10.1007/s10677-018-9912-8

# The Problem of Justified Harm: a Reply to Gardner

Jens Johansson [1] · Olle Risberg [1]

Accepted: 18 July 2018 / Published online: 1 August 2018


**Abstract**
In this paper, we critically examine Molly Gardner's favored solution to what she calls "the problem of justified harm." We argue that Gardner's view is false and that her arguments in support of it are unconvincing. Finally, we briefly suggest an alternative solution to the problem which avoids the difficulties that beset Gardner's proposal.

**Keywords** Problem of justified harm · Causal account of harming · Counterfactual comparative account of harming · Molly Gardner

## 1 Introduction

In a recent article in this journal, Molly Gardner proposes a novel solution to what she calls *the problem of justified harm*—the problem of explaining why it is morally permissible to inflict harm on someone, for the sake of greater subsequent benefits to that same person, in some cases but not in others (Gardner 2017). Gardner's discussion focuses on the intuitive contrast between the following two cases:

> **Drowning Swimmer:** A swimmer was drowning in the ocean. A lifeguard jumped into the water and pulled her to shore, breaking her arm in the process.
> **Nazi Prisoner:** A man was imprisoned in a Nazi concentration camp for many years, where he suffered immensely and came very close to death. But after he was liberated, he flourished. Being in the camp had "enriched his character and deepened his understanding of life" [citation removed]. These changes in him were largely responsible for his subsequent well-being. (Gardner 2017, pp. 2–3).

✉ Jens Johansson
jens.johansson@filosofi.uu.se

Olle Risberg
olle.risberg@filosofi.uu.se

[1] Department of Philosophy, Uppsala University, Box 627, 75126 Uppsala, Sweden

Intuitively, the lifeguard's action in Drowning Swimmer was morally permissible. This case seems to be a clear instance of justified harm, especially on the assumption that the subsequent benefits received by the swimmer were greater than the harm of her broken arm. Nazi Prisoner, by contrast, seems to be a clear case of *un*justified harm. What the Nazis did was intuitively morally impermissible, even if the prisoner's enriched character brought him benefits greater than the harms that he suffered in the concentration camp.

What explains the moral difference between the two cases? Gardner criticizes various possible answers to this question and puts forward a solution of her own, which appeals to the following principle:

> *The Causal Principle of Justified Harm (C):* A harmful action that causes greater benefits can sometimes be justified by those benefits, but a harmful action that does not cause greater benefits cannot be justified by any subsequent benefits that the action, itself, does not cause. (Gardner 2017, p. 14).

According to Gardner, whereas there is no reason to deny that the lifeguard's action caused the swimmer's subsequent benefits, we should deny that the prisoner's subsequent benefits were caused by the Nazis' imprisoning him. Gardner bases this asymmetrical judgment on the following "backtracking" view of causation (taken from Broadbent 2007, 2008):

> If *c* causes *e*, then if *e* hadn't occurred, *c* wouldn't have occurred. (Gardner 2017, p. 12)

Gardner argues that this requirement is satisfied in Drowning Swimmer. Consider this claim:

> (S) If the swimmer had not avoided death, then the lifeguard would not have pulled her to shore. (Gardner 2017, p. 14)

According to Gardner, (S) is true. For, she claims, while there are possible worlds in which the swimmer drowns although the lifeguard pulls her to shore, those worlds are more remote from the actual world than some world in which the swimmer drowns and the lifeguard does not pull her to shore.

Here is the corresponding claim about Nazi Prisoner:

> (R) If the prisoner had not acquired such an enriched character, then the Nazis would not have imprisoned him. (Gardner 2017, p. 13)

According to Gardner, (R) is false, "because the prisoner's path to an enriched character was not directly determined by what the Nazis did to him. Crucial to his growth was his own effort." (Gardner 2017, p. 13) By the backtracking view of causation, then, the Nazis' imprisoning him did not cause his enriched character. So, by (C), the lifeguard's action can be morally justified by the subsequent benefits but the Nazis' action cannot.

One may wonder why (C) should be accepted. In particular, why should it matter whether imprisoning the victim *caused* him benefits, so long as that action *itself* benefited him? That the Nazis' action did in fact benefit the victim is plausible on the well-known counterfactual comparative account of benefit and harm:

(CCA) An action benefits (harms) a person if and only if she would have been worse (better) off if the action had not been performed.

Gardner's response is that we should deny that the Nazis' action benefited the victim. She rejects (CCA) in favor of a causal account; let us call it "Gardner's Causal Account," or simply:

(GCA) An action benefits (harms) a person if and only if it causes a benefit (harm) for that person.

As Gardner points out, the notions of benefit and harm figuring on the right-hand side of (GCA) can be understood in various ways. The important thing to note right now, however, is that on (GCA), the counterfactual condition provided by (CCA) is neither necessary nor sufficient for an action's benefiting (or harming) someone. Instead, given that the Nazis' action did not cause the prisoner's subsequent benefits, (GCA) entails that their action did not benefit him.

## 2 Gardner's Arguments for the Causal Account of Harming

Before we turn to our arguments against Gardner's view, it is worthwhile to briefly consider her arguments in support of it. She provides three such arguments. One of them concerns the fact that (GCA), unlike (CCA), avoids problems to do with *preemption*. While we agree that preemption is a serious problem for (CCA), it should be noted that this fact provides reason to accept (GCA) only if it has been shown that other alternatives to (CCA) are likely to be unsuccessful.[1] Since that is an argument which Gardner does not seek to make, we shall focus on her other, more direct arguments in favor of her view.

Gardner's second argument concerns an analogy between verbs like 'clean', 'freeze', and 'kill', which Gardner calls "causal verbs," on the one hand, and 'harm', on the other.[2] While Gardner does not explain what she takes causal verbs to be, she writes that "[j]ust as causal verbs like 'clean', 'freeze' and 'kill' mean *cause to be clean*, *cause to be frozen*, and *cause to be dead*, so parity would suggest that 'harm' and 'benefit' are causal verbs, and that they mean *cause to be harmed* or *cause to be benefited*, respectively." (Gardner 2017, p. 9).

We have a hard time seeing what the supposed parity might consist in. Obviously, Gardner's point cannot be that all (transitive) verbs which intuitively have something to do with causation are semantically analogous to 'clean' and the rest; the sentence '*x* kicks *y*', for instance, does not mean that *x* causes *y* to be kicked. In light of that, however, it is far from clear why we should think that 'harm' is semantically analogous to 'clean' and the rest. It is of course true that 'harm' shares *some* grammatical properties with those verbs, but that is not motivation enough since the same thing may be said about 'harm' and 'kick'.

Perhaps it is tempting to respond that the verbs which Gardner lists correspond to adjectives that, in some intuitive sense, describe "states" of the objects to which they apply. In that respect, '*y* is clean' and '*y* is harmed' may seem to differ from '*y* is kicked'. Thus, might the

[1] We argue against various suggested solutions to the preemption problem for (CCA) in Johansson and Risberg (2017).

[2] The view that 'harm' is a causal verb is also proposed by Thomson (2011, p. 437).

argument be based on the assumption that whenever a verb corresponds to such a state-describing adjective, that verb must mean something like *cause to be in the corresponding state*? That line is also unconvincing. First and foremost, while Gardner takes 'kill' to be a causal verb, it is not plausible that '*y* is killed' always describes a kind of state of *y*. After all, there are at least some cases in which the person who is killed immediately ceases to exist, but the non-existent are in no states whatsoever. For similar reasons, the idea that '*y* is harmed' always describes a state of *y* requires the highly contentious assumption that death never harms its victim. But the argument is unsuccessful even if we disregard that complication. For although '*y* is wronged' also asserts that *y* is in a kind of state—the state of being wronged—'*x* wrongs *y*' does not mean, or even conceptually entail, that *x* causes *y* to be wronged. As an illustration, it may reasonably be said that the Nazis' action wronged the prisoner, but it is not the case that it also caused the victim to be wronged. (For one thing, that would require that the prisoner was not wronged until *slightly later* than the action was performed.) Rather, the Nazis' action, or some facts about it, *metaphysically explain* that the victim was wronged. And that fact is of no use for supporters of the causal view about harming. It is entirely open to opponents of that view to accept the corresponding claim about harm and harming, namely, that when someone is harmed, that fact is metaphysically explained by (some facts about) the event or events that harmed her.[3]

The upshot is that it is difficult to see why 'harm' should be analogous to 'clean', 'freeze' and 'kill'. In fact, that '*x* harms *y*' means '*x* causes *y* to be harmed' is precisely what opponents of the causal view deny, and the argument from linguistic parity does not seem to provide them with any reason not to do so.

Gardner's third argument for (GCA) is ethical rather than linguistic. The argument pertains to the *non-identity problem*, which roughly concerns whether and why it is impermissible to create a person who would occupy a low but overall positive well-being level rather than another person who would occupy a much higher well-being level. Gardner writes that in non-identity cases, "a seemingly harmful action is also the condition of your own worthwhile existence" (Gardner 2017, p. 9). Because she holds that a way to solve the non-identity problem is to show that the relevant procreative actions harm the conceived person (Gardner 2017, p. 10), she takes it to be a virtue of (GCA) that it allows for that result.

A problem with that argument, however, is that the non-identity problem arises even in cases in which none of the pertinent actions are plausibly harmful. Imagine a prospective parent who faces the choice of creating either Ash or Zola. Ash and Zola would live equally long lives if they were created, and neither of their lives would contain any negative well-being. However, Ash would be flourishing, living a life of immense positive well-being, whereas Zola would be subject to meager amounts of positive well-being only (muzak and potatoes, perhaps). The non-identity problem arises because even though Ash would not have been worse off if Zola had been created (and vice versa), the parent evidently ought to create Ash. However, since Zola's life would contain no intrinsic disvalue and the parent cannot do anything that is better for Zola than creating her, it is counterintuitive to think that creating Zola would harm her.

What this case suggests is that a solution to the non-identity problem will not rely on the assumption that the relevant procreative actions are always harmful to the conceived person. Contra Gardner, it is therefore difficult to see how any view of harm could be supported by reference to its alleged ability to solve that problem.

[3] Thanks to an anonymous reviewer for asking us to clarify the argument in this paragraph.

## 3 Problems for Gardner's View

We turn now to more direct problems for Gardner's view. The first one is very direct indeed—namely, a counterexample to her central thesis, the Causal Principle of Justified Harm. Here it is again:

> *The Causal Principle of Justified Harm (C):* A harmful action that causes greater benefits can sometimes be justified by those benefits, but a harmful action that does not cause greater benefits cannot be justified by any subsequent benefits that the action, itself, does not cause. (Gardner 2017, p. 14).

Now consider this variation on Drowning Swimmer:

> **Lifebuoy:** A swimmer was drowning in the ocean. A lifeguard threw a lifebuoy at her which landed on her arm, thereby breaking it. Despite the heavy waves, the swimmer managed to reach the shore, after which she went on to live a happy life. Although she would not have survived without the lifebuoy, her path to a continued life was not directly determined by the lifeguard's throwing it at her. Crucial to her continued life was her own effort.

It cannot be denied that the lifeguard's action—throwing the swimmer a lifebuoy—caused the harm of the broken arm. However, by the backtracking view of causation, the lifeguard's action did not cause the swimmer the benefit of avoiding death, because he would have thrown her the lifebuoy even if she had not avoided death. Instead, what caused her to avoid death were her own efforts—or at least this is what we should say if we agree with Gardner that in Nazi Prisoner, the prisoner's own efforts, rather than what the Nazis did to him, are what caused his enriched character (see Section 1). On the Causal Principle of Justified Harm, then, the swimmer's survival did not justify the lifeguard's throwing the lifebuoy. But it is clear that it did. So the Causal Principle of Justified Harm is false.

Gardner might reply that what justifies the lifeguard's action is the swimmer's having a lifebuoy.[4] The swimmer's having a lifebuoy, Gardner might suggest, is itself a benefit for the swimmer, and this benefit *is* caused by lifeguard's action: in particular, if the swimmer had not had a lifebuoy, the lifeguard would not have thrown the lifebuoy. This reply is unconvincing, however. To begin with, whatever we say about the swimmer's having a lifebuoy, it remains counterintuitive to deny that the swimmer's survival justifies the lifeguard's action. Furthermore, if the swimmer's having a lifebuoy is a benefit for her, this is presumably at least roughly because she would have been worse off if she did not have a lifebuoy—in particular, without the lifebuoy she would not have gone on to have a happy life. But similar things can be said about Nazi Prisoner. There must be several experiences that the Nazis' action caused the prisoner to have, ones without which he would have been worse off—in particular, experiences without which he would not have later acquired such an enriched character.

Gardner's Causal Account of harming is also susceptible to counterexamples. This may be illustrated through a pair of cases, the first of which runs as follows:

[4] Thanks to an anonymous reviewer for suggesting this reply.

**Hesitant Sniper, Lazy Victim:** A sniper hesitantly but intentionally fires a bullet towards a lazy victim. Due to her hesitance, the sniper almost decided not to fire the bullet. Once it was fired, however, the victim had no chance avoiding it. The bullet thus hits the victim, inflicting severe pain.

If Gardner is right that backtracking counterfactuals are quite often true, the following claim is true:

(1) If the lazy victim had not been hit by the bullet, the hesitant sniper would not have fired the gun.

Due to the hesitance of the sniper, the nearest world in which the bullet did not hit the lazy victim is one in which the sniper did not fire the gun (rather than one in which, say, the lazy victim suddenly gained the ability to dodge the bullet). (GCA) thus has the intuitively plausible implication that the lazy victim is harmed in this case. So far so good—but when the case is varied only slightly, the view has less promising results:

**Strong-Willed Sniper, Agile Victim:** A sniper with strong will intentionally fires a bullet towards an acrobat. The acrobat does her best to dodge the shot, and due to her agility, she almost manages to do so. Nevertheless, she is unsuccessful, and the bullet hits her to inflict severe pain.

We take it that even if backtracking counterfactuals are sometimes true, the claim about this case which corresponds to (1) is manifestly implausible. That is, we should reject:

(2) If the acrobat had not been hit by the bullet, the strong-willed sniper would not have fired the gun.

After all, due to her willpower, the sniper was strongly disposed to fire the gun. Any world in which she did not fire it is thus comparatively far away. By contrast, because of the acrobat's agility, the nearest world in which she manages to dodge the bullet is quite close—or at least close enough to ensure the falsity of (2). Thus, with regard to this case, (GCA) has the implausible result that the sniper's shot does not harm the acrobat.

Simply biting the bullet with regard to this problem would lead to further implausible results. As we have argued, (GCA) entails a "mixed" view about the two shootings: while the hesitant sniper's shot harms the lazy victim, the strong-willed sniper's shot does not harm the acrobat. The problem with the mixed view is not only that its implications with regard to the latter case are implausible in their own right. Another problem is that the *differences* between the two cases (concerning the abilities and dispositions of the agents involved) do not seem sufficient to justify different verdicts as to whether the victims are harmed by the respective shots. Finally, the mixed view also fails to track a difference in *moral* status between the relevant actions. More precisely, if actions can ever be impermissible on account of being harmful, it is evident that both shootings are such actions. This is significant because an important criterion of adequacy for theories of harm is that they "entail that harm is the sort of thing that it makes sense for there to be deontological restrictions about" (Bradley 2012, p.

396), and a theory which entails the mixed view thereby fails to satisfy that criterion.[5] It would thus be beside the point to reply that the strong-willed sniper's shooting was wrong for other reasons—the problem is that in both cases, the explanation of the action's wrongness should plausibly appeal to its harmfulness. Otherwise, the distinction between harming and not harming turns out to be as insignificant to ethics as the distinction between "incars" and "outcars" (cars in and outside of garages) is to metaphysics. Neither of those distinctions tracks something of theoretical importance.

(GCA) thus fails to satisfy Bradley's desideratum, and we think that this indicates a more general challenge for causal theories of harming. If there are deontological restrictions against harming, that is plausibly due to the fact that harm is the sort of thing which we should make sure that people do not suffer. However, from the point of view of a person who suffers harm, it makes no difference whether the harm was *caused* by some agent's action if that agent nonetheless could have ensured that the harm did not occur. But that suggests that the normative relevance of causing harm is not as great as causal theories of harm would have us think. To put the point another way, if the outcome of an action A1 is a harm for somebody and the outcome of an alternative action A2 is a benefit for them, that is surely a reason to perform A2 rather than A1 independently of whether A1 would also cause the harmful states of affairs for that person. But it is not clear how causal theories of harm can accommodate that fact.

## 4 Conclusion: an Alternative Proposal

The point just made leads us to a related and final remark. While our main aim in this paper is to show that Gardner's approach does not work, we shall conclude more constructively, by sketching a simple and straightforward view of the main moral difference between Drowning Swimmer and Nazi Prisoner.[6] In our view, the most important factor (which is notably absent from Gardner's discussion) concerns the set of *alternatives* available to the relevant agents. The lifeguard's action in Drowning Swimmer was morally permissible mainly because the harm inflicted upon the swimmer was a sort of necessary evil: he had to inflict that harm on her in order to save her life. That is, we take it that pulling the swimmer to shore without breaking her arm (or inflicting some equivalent harm on her) was not something that the lifeguard could have done. That assumption is not made explicit in Gardner's presentation of the case, but it is needed for the intuitiveness of the verdict that Drowning Swimmer is a case of justified harm. If we instead assume that the lifeguard could have pulled the swimmer to shore without harming her, it is no longer clear that his actual action was permissible, as we are now supposing it to bring the swimmer unnecessary harm. (Analogous claims apply to the variation on the case, Lifebuoy, which was introduced in Section 3.)

[5] Of course, not everyone agrees that the concept of harm must be able to play this role in deontological theories (see, e.g., Hanna 2016). Nonetheless, accepting that harm is ethically irrelevant seems to us to be a significant cost.

[6] There are also various other possibilities worth exploring, beyond the ones that Gardner criticizes. For instance, an anonymous reviewer points out that one potentially relevant difference between the two cases concerns the intentions of the agents: these are noble in Drowning Swimmer but evil in Nazi Prisoner. Gardner (2017, p. 15) briefly touches on this idea, but does not develop it.

Correspondingly, in Nazi Prisoner, the Nazis' action was morally impermissible mainly because the harm done to the prisoner was *not* a necessary evil: it was not something they had to do in order to bring him the subsequent (or equivalent) benefits.[7] There must have been several alternative actions available to the Nazis which would have resulted in a more favorable balance of benefit over harm for the prisoner; or at any rate, that assumption about their alternatives (while again not explicit in Gardner's presentation) is required for the intuitiveness of the verdict that Nazi Prisoner is a case of unjustified harm. If we instead assume, highly unrealistically, that the Nazis had no such alternative available—so that what they did actually made the prisoner's life as good as possible—it is no longer intuitively clear, at least not to us, that their action was morally impermissible.

Of course, the appeal to necessary versus unnecessary harm for the victim is only a partial solution to the problem of justified harm, not least because a full solution must also take into account how the action affects people other than the victim. Nevertheless, we think this view identifies the most important moral difference between Drowning Swimmer and Nazi Prisoner, and it does so without invoking any of the problematic elements in Gardner's proposal.

**Acknowledgments** We are grateful to Erik Carlson and two anonymous referees for their very helpful comments. Work for this paper was supported by grant P14-0212:1 from Riksbankens Jubileumsfond.



## References

Bradley B (2012) Doing away with harm. Philos Phenomenol Res 85(2):390–412
Broadbent A (2007) Reversing the counterfactual analysis of causation. Int J Philos Stud 15(2):169–189
Broadbent A (2008) The difference between cause and condition. Proc Aristot Soc 108:355–364
Gardner M (2017) When good things happen to harmed people. Ethical Theory Moral Pract. https://doi.org/10.1007/s10677-017-9840-z
Hanna N (2016) Harm: omission, preemption, freedom. Philos Phenomenol Res 93(2):251–273
Johansson J, Risberg O (2017) The preemption problem. Philos Stud. https://doi.org/10.1007/s11098-017-1019-x
Thomson J (2011) More on the metaphysics of harm. Philos Phenomenol Res 82(2):436–458

[7] Note that this is compatible with the reasonable assumption that bringing the prisoner such benefits is not what the Nazis *would* have done if they had not performed the action they in fact performed. For more on this sort of issue, see Johansson and Risberg (2017).