Ethic Theory Moral Prac
https://doi.org/10.1007/s10677-017-9840-z

When Good Things Happen to Harmed People

Molly Gardner[1]

Accepted: 2 October 2017

Abstract The problem of justified harm is the problem of explaining why it is permissible to inflict harm for the sake of future benefits in some cases but not in others. In this paper I first motivate the problem by comparing a case in which a lifeguard breaks a swimmer's arm in order to save her life to a case in which Nazis imprison a man who later grows wiser as a result of the experience. I consider other philosophers' attempts to explain why the lifeguard's action was permissible but the Nazis' action was not. After arguing that principles having to do with consent, expected utility, and the types of harms and benefits at issue do not fully solve the problem, I argue for a causal solution to the problem. The causal solution includes both a causal account of harming and a distinction between causes and mere conditions. It then distinguishes between the lifeguard and Nazi cases with following principle: 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.

Keywords Problem of justified harm · Causal account of harming · Non-comparative harm · Counterfactual comparative account of harming · Harmful omissions

1 Introduction

Sometimes we harm other people in order to benefit them. A parent might vaccinate his child against polio, even though the shot has unpleasant side effects. A lifeguard might break an unconscious stranger's arm in order to save her from drowning. A surgeon might cut into a patient's abdomen in order to remove a tumor. Commonsense morality suggests that in these kinds of cases, imposing harm for the sake of later benefits is morally permissible.

✉ Molly Gardner
mollyg@bgsu.edu

[1] Department of Philosophy, 312 Shatzel Hall, Bowling Green State University, Bowling Green, OH 43403, USA

Published online: 12 October 2017

In other cases, however, it seems morally impermissible to harm people, even if those harms also lead to benefits. It would be wrong for me to disable my child—perhaps by making him deaf or paralyzed—even though a life with a disability might turn out to be good for him.[1] It is wrong to rape someone, even if the rape leads first to a pregnancy and then to a child for whom the rape victim is ultimately grateful.[2] It is wrong to kidnap and imprison someone, even if, upon her release, the victim channels the trauma of the experience into a new reservoir of strength and wisdom.

A problem arises when we attempt to explain why it is permissible to impose the harms in the vaccine, drowning, and surgery cases but not in the disability, rape, and kidnapping cases. In every case, a person is both harmed and benefited, and we can suppose that in every case, the benefits are greater than the harms. Why, then, are the harms justified in the vaccine, drowning, and surgery cases but not in the disability, rape, and kidnapping cases? In this paper, I will argue that the most important difference between the two sets of cases is a causal one: the actions that cause the harm in the vaccine, drowning, and surgery cases also cause the later benefits, but the actions that cause the harm in the disability, rape, and kidnapping cases do not cause the later benefits. To justify this view, I will argue that we should reject what is commonly referred to as the "but-for test" as a sufficient condition, not only for harming and benefiting, but also for causation. I will begin in Section 2 with a more precise explication of the problem, which I call the "problem of justified harm," and I will survey some unsuccessful attempts to solve it. In Sections 3, 4, and 5 I will defend the claim that we should accept a causal account of harming and benefiting, and that if we do accept a causal account, we should pair it with a substantive account of causation that distinguishes between a cause and a mere condition. I will formulate a moral principle according to which a harmful action can be justified by only those benefits it actually causes, and I will conclude that this principle, in combination with the causal account of harming and benefiting and the distinction between a cause and a mere condition, best explains our intuitions about the two sets of cases.

2 The Problem

Consider 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.[3]
> **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

[1] See Elizabeth Barnes (2014) for an argument that having a disability can be good for someone.
[2] This particular example comes from Elizabeth Harman (2004).
[3] Cases like this are discussed in Parfit (1986), Woodward (1986, 1987), Shiffrin (1999), Harman (2004) and Bradley (2009).

> understanding of life" (Harman 2004, p. 99). These changes in him were largely responsible for his subsequent well-being.[4]

Intuitively, it seems that what the lifeguard did was morally justified, but what the Nazis did was wrong. Nevertheless, there is a puzzle about explaining this difference in judgments. On the one hand, it seems as though the benefits to the swimmer justified the harm of the broken arm. Avoiding death was a benefit to the swimmer, and avoiding death came to her as a greater benefit than a broken arm came to her as a harm. But if we say that, it's difficult to see why the benefits to the Nazi prisoner—benefits which were also greater than the harms he suffered—did not justify his imprisonment in the concentration camp. Granted, the case may be unusual; it is likely that most Nazi prisoners not only failed to flourish, but continued to suffer long after they were liberated from the camps. However, the prisoner we are imagining is a particular sort of person, such that having an enriched character benefited him more than being in the camp, awful as it was, harmed him. Even supposing that the prisoner was unusual in this way, it is still clear that what the Nazis did to him was wrong. I will use the term 'problem of justified harm' to refer to the problem exemplified by the Drowning Swimmer and Nazi Prisoner cases. The problem is one of identifying the feature or features that determine when it is permissible to harm someone who will later benefit, and when it is not.

In this paper, I won't advance a full solution to the problem. That is, I won't attempt to explicate *all* the features that determine whether any particular case of harming is justified. Some of these features have already been explicated by other philosophers, who have appealed, for example, to principles about consent, expected utility, types of harms and benefits, and rights violations. I hope to show, however, that whether they are considered singly or in combination, these other principles cannot satisfactorily answer all the questions we might have about why some harms are justified and some are not. Some of the aforementioned principles do not fully explain the relevant differences between cases, and others are simply false. In the rest of this section, I will show how each principle, in turn, runs into one or the other of these objections. At the end of this section, I will argue that, even in combination, these principles don't yield a satisfactory solution to the problem of justified harm; an additional principle is needed.

A first pass at a solution to the problem of justified harm appeals to the following principle: it is permissible to impose harm on someone when the harm is consented to, and it is impermissible to impose harm on someone when he or she does not consent.[5] According to this approach, The Nazis wronged the prisoner because he did not consent to be harmed. On

[4] This particular case is raised by Woodward (1986) and then discussed at length in Parfit (1986), Woodward (1987), Shiffrin (1999), Harman (2004), and Bradley (2009). It is also worth noting that a character in one of David Foster Wallace's short stories ruminates about this kind of case. At one point, the character says,

> Was the Holocaust a good thing? No way. Does anybody think it was good it happened? No way. But did you ever read Victor Frankl? Victor Frankl's *Man's Search for Meaning*? It's a great, great book. Frankl was in a camp in the Holocaust and the book comes out of that experience, it's about his experience in the human Dark Side and preserving his human identity in the face of the camp's degradation and violence and suffering total ripping away his identity. It's a totally great book and now think about it, if there wasn't a Holocaust there wouldn't be a *Man's Search for Meaning* (Wallace 1999, p. 98).

[5] Ben Bradley does not endorse this exact principle, but he suggests that consent can go a good way towards solving the problem (2009, p. 67).

the other hand, the lifeguard did not wrong the swimmer because, presumably, the swimmer consented to have her arm broken in the course of the rescue.

This principle will not fully solve the problem, for we can modify the cases so that there is the same presence or lack of consent in both of them. Suppose that the swimmer was unconscious when the lifeguard broke her arm, so she could not, at the time of his action, explicitly consent to it. We could also suppose that, for whatever reason, the Nazi prisoner was unconscious (or perhaps underage) when the Nazis captured him, and so he, too, was unable at the time of their action to explicitly consent. Even then, the lifeguard's action seems permissible, but the Nazis' action does not.

One might suggest that there is still a difference between the cases in terms of implied or hypothetical consent. Perhaps the lifeguard had the swimmer's hypothetical consent to break her arm, but the Nazis did not have the prisoner's hypothetical consent to capture him. We might then appeal to the principle that it is permissible to impose harm on someone when the harm is *hypothetically* consented to, and it is impermissible to impose harm on someone when he or she does not *hypothetically* consent.

This modified version of the consent principle, however, is not a full solution to the problem. In order to apply the modified principle, we need to know what it is about the swimmer's situation that grounds her hypothetical consent, and why it is absent in the prisoner's case. But whatever it is that establishes hypothetical consent in Drowning Swimmer and rules it out in Nazi Prisoner is, itself, a morally relevant difference between the two cases. Thus, if we wish to solve the problem of justified harm by appealing to the principle of hypothetical consent, we must also find at least one *other* moral principle that distinguishes between the two cases. The principle of hypothetical consent cannot, on its own, solve the problem.[6]

A second pass at a solution to the problem of justified harm endorses the following principle: it is permissible to impose harm on someone when and only when doing so maximizes the individual's expected utility.[7] According to this approach, the lifeguard acted permissibly because he maximized the swimmer's expected utility; he knew there was a good chance that he would save the swimmer's life, that the value of the swimmer's life to her was high, and that the harm of a broken arm, though certain, was relatively small. On the other hand, the Nazis wronged the prisoner because, although they maximized the prisoner's actual utility, they failed to maximize his expected utility. Although the prisoner flourished after his liberation, this was not something the Nazis could have expected when they imprisoned him; from their perspective, it was more likely that he would either die in the camp or suffer greatly for the rest of his life.

However, the principle that this approach appeals to is not true. As Elizabeth Harman (2004) points out, the Nazis would have acted wrongly even if they could have foreseen—and thereby *expected*—the resilience of this particular prisoner.[8] Contrary to the expected utility principle that I formulated above, it is sometimes wrong to maximize a person's expected

[6] Cf. Judith Thomson, who considers whether an appeal to hypothetical consent can solve the trolley problem. She rejects such a strategy, writing, "Why should we care about [hypothetical consent]? For my own part, I think we shouldn't. What I think we should care about is not that such and such people would consent if they were asked, but rather whatever it is about them in virtue of which they would consent, if they would" (Thomson 1990, p. 187).

[7] Ben Bradley does not endorse this exact principle, but he also takes expected utility to be relevant to whether harming someone is permissible (2009, p. 67).

[8] In support of this point, she cites Woodward, who writes, "[T]he usual view is that whether or not the benefit is intended or 'aimed at' (Parfit's phrase) makes very little difference to the justifiability of the action, although it may perhaps make a difference to the blameworthiness of the action if we decide that his action was unjustifiable" (Woodward 1986, footnote 8, cited in Harman 2004, endnote 32).

utility. There are some harms you ought not to impose on someone, even if you correctly foresee that those harms are the condition of greater benefits in the future.

A third approach to solving the problem appeals to various "non-comparative" conceptions of harm. A non-comparative harm is a condition or state of affairs that is bad for someone even if it doesn't make her worse off than she would have been in its absence. Elizabeth Harman appeals to a non-comparative conception of harm when she argues that "an action harms a person if the action causes pain, early death, bodily damage, or deformity to her" (2004, p. 93). This is a plausible step towards solving the problem of justified harm because it vindicates the judgment that the Nazis harmed the prisoner, even if he wound up better off on the whole. To show how the rest of the solution might go, Harman compares Nazi Prisoner to another case she calls "Surgery," in which a doctor cuts open her abdomen in order to remove her swollen appendix:

> First, the harms in …[Nazi Prisoner] are such awful, gruesome harms that it is much harder for other factors to render them permissible. Second, the consideration available to outweigh the harm in Surgery is the threat of worse harm of the same type. The considerations available to outweigh the harms in …[Nazi Prisoner] are substantive benefits that come along with the harms. It is much easier for the threat of worse harm to render harm permissible, than it is for accompanying benefits to do so (p. 99 – 100).

However, this approach is not wholly successful at solving the problem. For one thing, the same sort of problem arises in cases that involve less gruesome harms than those in Nazi Prisoner. Imagine a case where an innocent man is abducted, not by Nazis, but by small-time criminals. He is held captive in a basement for a week or so, then released. Like the Nazi prisoner, he flourishes later and was glad he was kidnapped, but it is still clear that he was wronged. Moreover, we can make the contrasting case more gruesome. Rather than a broken arm or abdominal surgery, we can imagine a case where a doctor inflicts a brutal regimen of surgery, chemotherapy, and radiation on a consenting patient in order to cure her cancer. In the latter case, despite how gruesome it is, the treatment is still permissible.

The second problem with Harman's approach involves her appeal to "harm of the same type." Certainly, in her surgery case, it seems plausible that the harm of having one's abdomen cut open is of the "same type" as the harm of a deadly infection; both of these are harms to one's body. However, not all cases in which we have the permissibility intuition are cases in which the harm avoided is of the "same type" as the harm inflicted. Suppose your wealth consisted of gold bars that you carried around in a backpack, but you were drowning in the ocean, and the backpack was weighing you down. It would be permissible for me to remove your backpack and let it sink to the bottom of the ocean if this were the only way to save your life. But in that case, the harm I inflict on you is the loss of your money, and the harm you avoid is the loss of your life; presumably, these harms are not of the same type.

Seana Shiffrin (1999) also attempts to solve the problem of justified harm by appealing to non-comparative conceptions of harm and benefit, but her proposal is slightly different from Harman's. Shiffrin appeals to what she calls "pure benefits," or "benefits that are just goods and which are not also removals from or preventions of harm" (p. 124). According to Shiffrin, when a person is unavailable to give consent, it is wrong to harm her in order to provide her with a pure benefit, although it may be permissible to harm her in order to prevent a greater harm. Supposing again that, for whatever reason, neither the Nazi prisoner nor the drowning swimmer were available to give consent, this approach is at least compatible with our judgments about both cases. The lifeguard yanks on the swimmer's arm in order to prevent her death, which is, indeed, a greater harm than the broken arm, so our intuitive judgment that

the lifeguard's action is permissible is consistent with the verdict Shiffrin's principle delivers, namely, that the lifeguard's action may be permissible. Moreover, it is plausible that the Nazis caused the prisoner years of suffering and did not prevent a greater harm to him at all; the improvements to his character were a pure benefit. Shiffrin's principle thus correctly implies that what the Nazis did was wrong.

Still, Shiffrin's principle has nothing to say about a case that is similar to Nazi Prisoner, but in which the benefits are not pure:

> **Kidnapping:** Before he was kidnapped by small-time criminals, a young man suffered from depression. He took to drinking regularly, and as a result of his intoxication, he was unable to consent to anything. Then he was kidnapped and held in a basement for many months, where he continued to suffer immensely and came very close to death. After he was rescued, he discovered that his depression had lifted, and he no longer craved alcohol. Being kidnapped and imprisoned had made him more stoic and wise. These changes in him were largely responsible for his ability to escape both depression and alcoholism later in life.

Intuitively, what the kidnappers did in this case was wrong. Nevertheless, Shiffrin's sufficient condition for wrongdoing is not satisfied. Shiffrin holds that when a person is unavailable to give consent, it is wrong to harm him in order to provide him with a pure benefit. But the benefit to the victim in this case is not a pure benefit, just as the benefit to the swimmer in Drowning Swimmer is not a pure benefit. If depression in later life would have been a greater harm for the man than being kidnapped and imprisoned, then the verdict that Shiffrin's principle delivers about the kidnapping is the same as the verdict it delivers about the lifeguard's action in Drowning Swimmer: both actions *may* have been permissible. Since Shiffrin's principle is unable to account for the judgment that the lifeguard's action *was,* indeed, permissible, while the kidnapping was not, it is not a complete solution to the problem of justified harm.

Even worse, Shiffrin's principle gets the wrong results in the following case:

> **PlayStation 4:** Felicity is in the hospital recovering from minor knee surgery. She is currently unconscious from the anesthesia and cannot consent to anything. As a special promotion, Sony is offering a PlayStation 4 to everyone in the hospital at that exact time who has a noticeable bruise on their upper body. Felicity's husband knows that Felicity has been wanting a PlayStation 4 more than almost anything else in the world, but the couple have been unable to afford one. Fred pinches Felicity on her upper arm, causing a bruise, so that Felicity can get the PlayStation 4.

Intuitively, the harm Fred causes his wife in this case is justified.[9] Yet a PlayStation 4 is a pure benefit. If so, then contrary to Shiffrin's principle, even when consent is unavailable, it is not always wrong to harm someone in order to give her a pure benefit.

[9] An anonymous reviewer questions whether the bruise really counts as a harm. However, I will stipulate that it's a painful bruise. Given this stipulation, the bruise (or at least the pain that goes with it) should qualify as a harm on Shiffrin's view; Shiffrin explicitly states that "pain counts as a harm" (p. 124). There may, of course, be a modified version of Shiffrin's view on which pain does not count as a harm. But on that view, it's not clear what Shiffrin's sufficient condition for wrongdoing would contribute towards solving the problem of justified harm. Presumably, such a view would instead attempt to solve the problem by holding that what I'm calling "justified harms" are not really harms at all. This is a version of the "moralized harm approach," which I discuss next. The upshot here is that if the painful bruise is a harm, then Shiffrin's principle is false, and if the painful bruise is not a harm, then Shiffrin's principle is irrelevant to solving the problem.

A fourth approach to the problem of justified harms is to hold that the difference between the two cases is that the Nazi prisoner was wronged, whereas the swimmer was not.[10] This approach is quite different from the other approaches I have so far considered because it rejects one of the assumptions I relied upon when I first formulated the problem: that in the cases where a child is vaccinated, a swimmer is rescued, or a patient undergoes surgery, the child, swimmer, and patient are harmed at all. Instead, what we can call the "moralized harm approach" holds that a necessary condition for something's *being* a harm is that it be wrongful: If the child, swimmer, and patient were not wronged, then they were not harmed. On this view, there is no problem of distinguishing justified harm from unjustified harm, for there is no such thing as "justified harm."

Nevertheless, this view faces an analogous problem, which we can call the "problem of justified action." The problem of justified action is motivated by exactly the same cases that motivate the problem of justified harm. To solve the problem of justified action, we must determine why the actions of, for example, the lifeguard and the surgeon were morally justified, whereas the actions of the Nazis were not.

Proponents of the moralized harm approach would most likely appeal to principles about rights. They might say that the Nazis violated the prisoner's rights, but the lifeguard did not violate the swimmer's rights. This seems plausible, but the explanation is still incomplete. We need to know the boundaries of these rights—why, for example, *didn't* the lifeguard violate the swimmer's rights? After all, if you were sitting safely on the beach, minding your own business, and a lifeguard ran over to you and broke your arm, he *would* have violated your rights. Why doesn't he violate the swimmer's rights when he breaks her arm in the drowning case?

The moralized harm approach thus faces a problem similar to the one I discussed earlier, in relation to the principle of hypothetical consent. Like the appeal to hypothetical consent, the moralized harm approach is not a full solution to the problem of distinguishing between cases like Drowning Swimmer and Nazi Prisoner. Additional principles—ones that can explain, for example, why some cases of breaking your arm are rights violations and some are not—will be needed.[11]

Let us now take stock. I have discussed a number of approaches to the problem of justified harm (or to what a proponent of the moralized harm approach might call the "problem of justified action"). I have argued that some of these approaches, such as the appeal to hypothetical consent or the appeal to a moralized conception of harm, are not full solutions: they cannot explain our intuitions about the cases unless they are supplemented with extra

[10] James Woodward (1986, 1987) explicitly endorses this view as a solution to the problem. There are also other philosophers who endorse moralized conceptions of harm, although they don't explicitly consider the problem of justified action as I have framed it. Such philosophers include Joel Feinberg (1984) and Joseph Raz (1986).

[11] An anonymous reviewer suggests that these additional principles can be derived from Kantian moral theory, from rule utilitarianism, or from any other plausible rights account, and that a solution that appeals to these theories might be preferable to the causal one I will eventually endorse. However, it's not clear to me that any of these moral theories have direct implications for the particular cases at issue; we still need to supplement them with interpretive principles that tell us how these theories may be applied in particular sets of circumstances. If we want to use Kantian theory, for example, we will need to know when cutting open someone's abdomen counts as treating her merely as a means, and when it does not. If we use rule utilitarianism, we will need to know what the specific rules are that govern harming someone for the sake of a future benefit. I suspect that the principles we appeal to when we interpret these theories will resemble the principles I have already discussed above. For example, many Kantians hold that you qualify as treating someone merely as a means in exactly those cases where she doesn't or wouldn't consent to your treatment.

moral principles. I also argued that other approaches, such as appeals to expected utility or to types of harms and benefits, got the wrong results in certain cases. It should now be clear why, *in combination,* the various approaches I have so far discussed won't work. If, for example, Shiffrin's principle gets the wrong results in certain cases, then it will get the wrong results in those cases even when it is combined with the hypothetical consent principle. Similarly, if the expected utility principle is false, then we won't make much progress by combining the expected utility principle with a moralized conception of harming. Even the combination of a moralized conception of harming and a principle of hypothetical consent will not fully solve the problem. Perhaps a principle of hypothetical consent can round out the moralized harming approach by, say, distinguishing rights violations from permissible rights infringements. Nevertheless, in order to use such a combined approach, we would still need to know what the conditions are that establish hypothetical consent.

Here is what will help: finding a morally relevant difference between cases that the other approaches to the problem have overlooked. And now notice this: *None of the previously discussed approaches challenged the claim that both the swimmer and the prisoner were benefited by the harmful action.* Indeed, I suspect that this failure to challenge such a claim is part of the reason that none of the previously discussed approaches were entirely successful. They were relying on a mistaken view of what it is to benefit someone.

Rather than hold that both the swimmer and prisoner were benefited by the very actions that harmed them, I think we ought to say that while the lifeguard's action both harmed and benefited the swimmer, the Nazis' action harmed but did not benefit the prisoner. The difference is a causal one: harming is causing a harm, and benefiting is causing a benefit. In Drowning Swimmer, the lifeguard's action caused the harm of the broken arm as well as the benefit of avoiding death. In Nazi Prisoner, the Nazis' actions caused all the harms associated with imprisonment in the concentration camp, but they did not cause the later benefits. The prisoner, himself, was the principal cause of those benefits, and imprisonment was a mere condition of them. Similarly, in all of the cases we have considered where later benefits are linked to changes in the agents' character traits or values, the agents' own decisions are the principal causes of those benefits. Harming an agent in such a case is not a way to cause those benefits, so *a fortiori,* it is not a permissible way to cause those benefits. In the next few sections, I will more carefully defend each piece of the view I have just articulated.

3 The Causal Account of Harming and Benefiting

The literature on harming and benefiting has recently undergone a fissure.[12] On one side of the fissure is the view I will refer to as the "counterfactual account," which can be formulated as follows:

> An action or omission *benefits* a victim if and only if she is better off given the action or omission than she would have been, given the absence of the action or omission. An

[12] For some of the recent work on the metaphysics of harming, see Norcross (2005), Hanser (2008), Harman (2009), Thomson (2011), Bradley (2012), Shiffrin (2012), Klocksiem (2012), Tadros (2014), Feit (2015, 2016), Hanna (2016), and Gardner (2015, 2017). With the notable exception of Shiffrin (1999), less work has been done on the metaphysics of benefiting; most philosophers seem to assume that an account of benefiting should simply be the mirror opposite of an account of harming. I find this assumption plausible, so I have formulated the two prominent accounts of harming I discuss as accounts of harming *and* benefiting.

action or omission *harms* a victim if and only if she is worse off given the action or omission than she would have been, given the absence of the action or omission.[13]

On the other side is what I will call the "causal account," which can be formulated this way:

An action or omission *benefits* a victim if and only if it causes a benefit for that victim.
An action or omission *harms* a victim if and only if it causes a harm for that victim.

Proponents of the counterfactual account tend not draw a sharp distinction between the concept of harming and the concept of a harm: they will often refer to an event that harms as either the "harmful event" or "the harm" (and likewise with benefit).[14] On the other hand, the causal account is usually supplemented by distinct accounts of harm and benefit. For most causal theorists, an action that harms is not, itself, the harm; rather, its upshot (usually a state of affairs) is the harm. A causal theorist's separate account of what makes a state of affairs a harm or a benefit may be non-comparative like the accounts discussed above, or it may be comparative.[15]

The counterfactual account was originally dominant in the literature.[16] However, philosophers have lately identified a number of advantages that the causal account has over the counterfactual account.[17] In this section, I will first review some of those advantages. I will then point out another advantage that, in my view, has not yet been fully appreciated.[18] Namely, the counterfactual account gets counterintuitive results in some of the *same cases* that have puzzled theorists—including metaphysicians and philosophers of law—who work on causation. However, the causal account can escape these counterintuitive results by availing itself of the innovations that these metaphysicians and philosophers of law have already developed to deal with such cases.

The first reason to favor the causal account over the counterfactual account has to do with linguistic parity. Just 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.[19]

A second reason to favor the causal account—at least with respect to harming—is that does not immediately rule out the possibility that you can be harmed in what is known as a "non-identity case," or a case where a seemingly harmful action is also the condition of your own worthwhile existence. Suppose, for example, that your mother's act of conceiving you at the age of 14 both brings you into existence and sets you up for a bad start in life.[20] The counterfactual account implies that as long as your life is worth living, your mother's act didn't harm you, since you would not have been better off, had she not conceived you. The

[13] This account of harming and benefiting is meant to be a generic statement of a view defended by Norcross (2005), Boonin (2014), Hanna (2016), and Feit (2015, 2016). Nevertheless, it is not an exact statement of any particular version of the view.
[14] I thank Neil Feit for helping me to appreciate this point.
[15] For example, Gardner (2015) combines a causal account of harming with a comparative account of harm.
[16] For example, before Shiffrin (1999) and Harman (2004) advanced a causal account of harming as a solution to the non-identity problem, almost all of the literature on the problem took a counterfactual account of harming for granted. For an overview of that literature, see Roberts (2015).
[17] See Harman (2004, 2009), Thomson (2011), and Gardner (2017).
[18] Harman (2009) incorporates a substantive causal principle into her causal account of harming, but she does not defend her choice of principle.
[19] See Thomson (2011).
[20] This is Derek Parfit's (1984) case, which he uses to illustrate the non-identity problem.

causal account does not have this implication, and is therefore a more promising starting point from which to justify the intuition that your mother has in some way harmed you.[21] Finding such a justification is a way to solve the so-called "non-identity problem."[22]

A third reason to favor the causal account is that it does not immediately rule out the possibility of accommodating our intuitive verdicts about benefiting or harming in preemption and overdetermination cases. Consider, for example, the following preemption case:

> **Preemption**: Jones is drowning, and Smith throws him a life preserver. If Smith hadn't thrown Jones the life preserver, then Brown would have.[23]

Intuitively, Smith benefits Jones: he saves Jones's life. However, the counterfactual account of benefiting implies that he doesn't, given that Jones is no better off than he would have been, had Smith not saved him. On the other hand, as long as it is conjoined with principles of causation sophisticated enough to handle preemption, the causal account of benefiting can avoid this result.

This third advantage of the causal account can be generalized. Part of what makes the causal account so attractive is that it can avail itself of all the sophisticated conceptual machinery that causal theorists have already developed. This is the advantage I want to emphasize, for it has so far not received due appreciation. Appreciating this advantage means acknowledging the wide number of harm- or benefit-related theoretical problems that can be solved, not by adding this or that epicycle to a moral theory, but by appealing to principles already available in the causation literature, which includes both the metaphysics literature and the philosophy of law literature. Such problems involve preemption and overdetermination, as is already recognized. They also involve cases in which we want to distinguish benefits or harms from epiphenomena. But what is most important for present purposes, such problems include the problem of justified harm. That is, if we accept the causal account of harming and benefiting, we can round out the solution to the problem of justified harm, not with a moral principle, but with a causal one.

4 Causes and Mere Conditions

In the previous section I suggested that metaphysicians and philosophers of law have already developed some causal principles that can help solve harm- and benefit-related problems in the ethics literature. In this section, I will argue that the kind of principle best suited to solving the problem of justified harm is a principle of *causal selection*: a principle

[21] Even if this intuition can be vindicated, it is a further question whether the harm in a non-identity case can be justified. There is thus a very close connection between the problem of justified harm and the non-identity problem. It follows from my solution to the problem of justified harm that only those benefits that are *caused* by the procreative action that brought you into existence can justify any harms that this action caused you. Because not all conditions are causes, the procreative action does not necessarily cause each benefit you experience in your life. Thus, even if you have a life worth living, it is still an open question—to be settled by the particular details of your case—whether the harmful action that brings you into existence is morally justified.

[22] See, for example, the harm-based solutions to the non-identity problem advanced by Shiffrin (1999), Harman (2009), and Gardner (2015).

[23] Parfit (1984) and Woollard (2012) also comment on the significance of preemption cases for counterfactual accounts of harming.

that distinguishes between causes and mere conditions. I will first explain how causal selection is accomplished by the theory Hart and Honoré (1985) develop in their book, *Causation in the Law*. I will then argue that Hart and Honoré's insights are roughly captured by a more general principle defended by Alex Broadbent (2007, 2008).

Hart and Honoré develop a theory intended to explicate the concept of causation that features both in ordinary thought and in the law. According to their analysis, ordinary people use the word 'cause' to selectively pick out either abnormal events or voluntary human actions or omissions that make a difference to later events, such as harms. However, not just any such event or omission will count as a cause of a harm; what would otherwise be a causal connection between one event and a subsequent harm can be severed by a second, intervening event or omission. In such a case, the second event or omission (which must either be an *abnormal* event or a *voluntary* human action or omission) becomes a cause of the harm, and the prior event or omission is relegated to the status of a mere condition. For example, if I voluntarily leave a pit in the road and you fall into it and break your arm, my omission has caused you the harm. But if I voluntarily leave a pit in the road and *someone else* deliberately pushes you into it, then my failure to fix the pit is not the cause of your broken arm. My omission is now a mere condition of the harm, and the actor who pushed you in is the cause.[24]

Hart and Honoré note that their theory of causation seems to be relied upon by the courts, not just in cases of harming, but also in cases where harms are followed by benefits. They write,

> [A] defendant cannot set off gain accruing to plaintiff unless it accrues in consequence of his wrongful act; and where the immediate source of the gain is an indemnity, compensation, or gift from a third person the rule is that it cannot be taken into consideration if the third person acted voluntarily (1985, p. 141).

This principle implies that if I wrongfully damage your property, but having heard of your plight, your friends raise some money to help you, the money they raise does not reduce the damages I owe you. It's true that if I *hadn't* damaged your property, your friends wouldn't have given you the money, but the truth of this counterfactual does not imply that I should get any credit for the money your friends raised.

There are clear similarities between Hart and Honoré's discussion of gains and the solution I will propose to the problem of justified harm. My solution turns upon the distinction Hart and Honoré also draw between a wrongful action that causes a later benefit and a wrongful action that is merely the condition of a later benefit. However, it's not clear that Hart and Honoré appreciated the implications their theory has for cases like Drowning Swimmer and Nazi Prisoner. In those cases, the matter of whether the benefit was caused by the earlier action determines, not just what an agent might owe in damages, but also whether the agent's action was justified to begin with.

There is also a second reason I am hesitant to endorse Hart and Honoré's view as a full solution to the problem of justified harm. My intention in this paper is to argue for a truly *causal* solution to the problem. And while Hart and Honoré frame their view as a theory of causation, I suspect that many metaphysicians take it to be a theory of something else: perhaps a theory of what our legal policies are or should be, or perhaps a theory of how judges,

[24] These pit cases are modeled on Hart and Honore's examples, which are based on actual tort cases; see Hart and Honoré 1985, p. 137.

historians, or lay people think or talk about causation.[25] To make it clear, then, that the solution I am arguing for is not merely an appeal to pragmatics or policy, I will defend my proposal for solving the problem of justified harm using Alex Broadbent's theoretical framework. Broadbent's principle for distinguishing causes from mere conditions has roughly the same implications as Hart and Honoré's theory, but Broadbent's principle is both more general and more unambiguously a principle of causation.

To distinguish between causes and mere conditions, Broadbent argues for the following necessary condition on causation:

> If *c* causes *e,* then if *e* hadn't occurred, *c* wouldn't have occurred (p. 355).

A unique feature of Broadbent's principle is that it appeals to a *backtracking counterfactual*, which asks us to evaluate what would have happened at an earlier time, had an event at a later time not occurred. Backtracking counterfactuals are not to be confused with the more familiar forward-tracking counterfactuals, which feature in many other legal and metaphysical accounts of causation, and which ask us to determine what would have happened at a later time, had some earlier event not occurred. Indeed, the "but-for test"—the main test that legal theorists rely on to establish what they call "causation-in-fact"—features a forward-tracking counterfactual, as do various Lewisian theories of causation, probabilistic theories of causation, and interventionist theories of causation.[26]

Broadbent's reliance on a backtracking counterfactual allows him to distinguish between causes and mere conditions in a way that those other theories cannot. Consider a case in which lighting strikes a barn, and a fire breaks out. Intuitively, we want to say that the lightning strike caused the fire, whereas the presence of oxygen in the air was a mere condition of the fire. The judgment that the lightning strike caused the barn fire satisfies Broadbent's principle: it is true that if the fire had not started, then the lightning would not have struck it. However, Broadbent's principle rules out the presence of oxygen in the barn as a cause of the fire. It is *not* true that if the fire had not started, then there would not have been oxygen in the barn.

Broadbent's necessary condition on causation is a consequence of a more general theory of causation that he defends at length. For obvious reasons, I won't recapitulate his whole defense here. However, this paper can be seen as offering one additional argument in support of his view. The argument is that, when it is conjoined with the causal account of harming and benefiting, Broadbent's theory supports a solution to the problem of justified harm.

5 Accounting for the Harms and Benefits

Recall the view I articulated at the end of Section 2. The view began with the claim that harming is causing harm and that benefiting is causing benefit. I then made the following

[25] Perhaps *because* of their attention to causal selection, Hart and Honoré's book—while highly influential among philosophers of law—has had much less influence in the metaphysics literature. For example, L.A. Paul and Ned Hall's (2013) book *Causation: A User's Guide* does not cite them at all and even uses a brief discussion of causal selection to illustrate how *not* to theorize about causation (see p. 35–36). Nor are Hart and Honoré cited in Ned Hall's (2006) *Philosophy Compass* overview of causation, nor are they cited in Jonathan Schaffer's (2016) *Stanford Encyclopedia of Philosophy* entry on the metaphysics of causation. Schaffer discusses causal selection, but he writes, "[S]election is now generally dismissed as groundless, and theorists seek to isolate some pre-selected, egalitarian conception of causation." I disagree, of course, with the notion that selection is groundless.

[26] For a helpful discussion of these other theories, see Hall (2006) and Paul and Hall (2013). By "Lewisian theories," I mean theories inspired by David Lewis's (1973) view.

claims about Drowning Swimmer and Nazi Prisoner: first, that the lifeguard and the Nazis harmed the swimmer and the prisoner, respectively; and second, that while the lifeguard also benefited the swimmer by saving her life, the Nazis did not benefit the prisoner by causing his character growth. In this section, I will defend each of these claims, respectively.

First, consider the claims that the lifeguard harmed the swimmer and the Nazis harmed the prisoner. To establish the truth of these claims, I would need to show that (1) the swimmer suffered a harm, (2) the lifeguard caused it, (3) the prisoner suffered a harm, and (4) the Nazis caused it. Claims (1) and (3) follow from any plausible causal account of harming. The swimmer obtained a broken arm, and the prisoner suffered from years of utter misery. If anything counts as a harm, a broken arm should count, and so should the prisoner's misery. Similarly, claims (2) and (4) would follow from any plausible theory of causation, Broadbent's included.

Next is the issue of benefits. First, notice that any plausible causal account of benefiting would likely affirm the presence of benefits in both Drowning Swimmer and Nazi Prisoner. Avoidance of death should qualify as a benefit, if anything does. In Nazi Prisoner, the prisoner has an enriched character and a deeper understanding of life. A plausible causal account of benefiting would affirm that these, too, are benefits.

Nevertheless, we must now determine whether the Nazis' actions were the cause of these benefits to the prisoner. Recall the "but-for test" that I mentioned earlier. The "but-for test" implies that the truth of the following counterfactual is sufficient in this case for causation:

(D) If the Nazis had not imprisoned the man, he would not have acquired such an enriched character.

D is undoubtedly true. And if it we accepted both D and the but-for test, we would have to conclude that by imprisoning the man, the Nazis not only benefited him, but they benefited him *more* than they harmed him. Indeed, I believe that an implicit acceptance of the but-for test and the consequent overestimation of the importance of D is responsible for most of the intuitions that fuel the problem of justified harm.

But notice that we can simultaneously hold that D is true while denying that it is as *important* as the but-for test would have it be. D is true, but so are these other counterfactuals:

(E) If the prisoner hadn't been born, he would not have acquired such an enriched character.
(F) If a meteor hadn't wiped out the dinosaurs, the prisoner would not have acquired such an enriched character.

Notice that we attach little significance to the truth of these other counterfactuals. We should similarly place little weight on the truth of D.

If we reject the but-for test as a sufficient condition for causation and accept Broadbent's necessary condition on causation instead, then we can see that the falsity of the following counterfactual implies that the Nazis *did not* benefit the prisoner:

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

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. Indeed, this case closely resembles the tort cases Hart and Honoré have in mind when they write that "the free, deliberate and informed act or omission of a human being, intended to exploit the situation created by defendant, negatives causal connection" (p. 136). In this case, the prisoner's free and deliberate *mental*

actions—reflecting on and learning from what happened to him—function to "negative" the causal connection that would otherwise hold between the Nazis' action and the prisoner's eventual character growth. Broadbent's principle explains why this is so. Consider some close possible worlds in which the prisoner does not have an enriched character. The closest of these worlds are not ones in which he isn't captured at all. They are worlds in which the Nazis imprison him, but instead of growing as a person, he succumbs to depression, anxiety, or death.

On the other hand, Broadbent's necessary condition on causation is satisfied in Drowning Swimmer, for the following counterfactual is true:

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

To see that this counterfactual is true, we can compare (a) the worlds in which the swimmer drowns and the lifeguard *doesn't* pull her to shore with (b) the worlds in which she drowns and the lifeguard *does* pull her to shore. Worlds of the former kind are less remote than worlds of the latter kind. Certainly, worlds of the latter kind are possible; there are scenarios in which the lifeguard pulls her to shore too late, after she has already inhaled a fatal amount of water. But I take it that (S) is true even if these other scenarios are possible; all it takes for (S) to be true is that the worlds in which the swimmer drowns and the lifeguard doesn't pull her to shore be slightly less remote than the worlds in which she drowns and he does.

Putting these considerations together, we can now see that the lifeguard's action had two effects: he broke the swimmer's arm, but he also saved her life. On the other hand, the Nazis' action had one set of effects: it caused the prisoner immense suffering. It is true that the prisoner later flourished, and it is also true that if the Nazis had not captured the prisoner, he would not have flourished in that way. However, his growth in character is to his own credit, and not to the Nazis'.

6 Conclusion

The best explanation for the difference in our judgments about the Nazi prisoner and Drowning Swimmer appeals to the following:

> *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.

In order to show that the principle applies to the two cases in question, I argued for a causal account of harming and benefiting. I also argued that if we accept a causal account of harming and benefiting, then we should also accept a principled distinction between a cause and a mere condition. The principle I endorsed, Broadbent's necessary condition on causation, says that an earlier event c causes a later event e only if it is true that if e had not happened, c would not have happened.

My view could, of course, be challenged at several steps. A number of those two whom I have presented this paper have challenged my argument that while counterfactual (R) in Nazi Prisoner is false, counterfactual (S) in Drowning Swimmer is true. Couldn't the empirical circumstances of the two cases be filled in differently, so that (R) comes out true and (S) comes out false?

In response to this objection, I grant that we could, indeed, alter the cases in ways that affect the truth values of the relevant counterfactuals. We can imagine, for example, the following variation on Drowning Swimmer:

> Drowning Swimmer 2: Lifeguard Larry doesn't like Swimmer Susan, so he swims over to her and breaks her arm, causing Susan to cry out in pain. Larry then exits the scene. Lifeguard Gary hears Susan's cries, and he starts swimming towards her. Just then, Susan starts to drown, and Gary is now close enough that he can grab Susan's other arm and pull her to shore.

In this case, we can suppose that Larry's action is a condition of Susan's benefit: if Larry hadn't broken her arm, she wouldn't have been saved by Gary from drowning. Nevertheless, the following counterfactual is false:

(S*) If Susan hadn't avoided death, then Larry wouldn't have broken her arm.

Those who take this false counterfactual to be an objection to my view will want to say one of two things. Either Larry's action *did* cause the benefit to Susan, in which case Broadbent's necessary condition on causation is false; or despite the failure of Larry's action to cause the benefit to Susan, Larry's action was still justified by that benefit, in which case (C) is false. However, I find neither of these claims to be plausible.

One might also object that the differences between Drowning Swimmer and Drowning Swimmer 2 are unnecessarily drastic. Perhaps, if we could get the relevant counterfactual to come out false without making so many changes, it *would* be plausible that we either ought to reject Broadbent's principle or (C). For example, we could add the detail that the swimmer inhaled a great deal of water. Why not say, given that detail, that she owes her life to her success at coughing it all up, and not to the lifeguard's pulling her to shore?

It is difficult to fully adjudicate this line of objection without resorting to a theory of the semantics for backtracking counterfactuals, and I am not aware that such a theory has yet been developed. Nevertheless, I am inclined to think that, if we had such a theory, it would still determine the relevant counterfactual to be true: that if the swimmer hadn't avoided death, then the lifeguard wouldn't have pulled her to shore.

Thus, despite the objections above, I maintain that (C) is still the best explanation of our moral intuitions about the relevant cases. Recall that (C) is not intended to be a *full* solution to the problem of justified harm; in some cases, we will have to appeal to other principles to explain why a particular harm is justified. In many of those cases, consent will undoubtedly make a difference. Other considerations like intentions and expected utility might have a role to play as well. However, there are some cases in which our judgments cannot be fully explained by factors like consent, intentions, expected utility, and so on. In those cases, I believe that the best explanation is a causal one.

Acknowledgments For helpful comments, I thank Stephen M. Campbell, Neil Feit, Duncan Purves, two anonymous reviewers, and audiences at the Workshop on Harm: The Concept and Its Relevance at Uppsala University 2016 and the Syracuse Philosophy Annual Workshop and Network 2016.

References

Barnes E (2014) Valuing disability, causing disability. Ethics 125(1):88–113
Boonin D (2014) The non-identity problem and the ethics of future people. Oxford University Press, Oxford
Bradley B (2009) Well-being and death. Clarendon Press, Oxford
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 15:169–189

Broadbent A (2008) The difference between cause and condition. Proc Aristot Soc 108:355–364
Feinberg J (1984) Harm to others. Oxford University Press, Oxford
Feit N (2015) Plural harm. Philos Phenomenol Res 90(2):361–388
Feit N (2016) Comparative harm, creation and death. Utilitas 28(2):136–163
Gardner M (2015) A harm-based solution to the non-identity problem. Ergo 2(17):427–444
Gardner M (2017) On the strength of the reason against harming. J Moral Philos 14:73–87
Hall N (2006) Philosophy of causation: blind alleys exposed; promising directions highlighted. Philos Compass 1(1):86–94
Hanna N (2016) Harm: omission, preemption, freedom. Philos Phenomenol Res 93(2):251–273
Hanser M (2008) The metaphysics of harm. Philos Phenomenol Res 77(2):421–450
Harman E (2004) Can we harm and benefit in creating? Philos Perspect 18:89–113
Harman E (2009) Harming as causing harm. In: Roberts M, Wasserman D (eds) Harming future persons: ethics, genetics and the nonidentity problem. Springer, Dordrecht, pp 137–154
Hart HLA, Honoré T (1985) Causation in the law, 2nd edn. Oxford University Press, Oxford
Klocksiem J (2012) A defense of the counterfactual comparative account of harm. Am Philos Q 49(4):285–300
Lewis D (1973) Causation. J Philos 70(17):556–567
Norcross A (2005) Harming in context. Philos Stud 123(1):149–173
Parfit D (1984) Reasons and persons. Oxford University Press, Oxford
Parfit D (1986) Comments. Ethics 86:832–872
Paul LA, Hall N (2013) Causation: a user's guide. Oxford University Press, Oxford
Raz J (1986) The morality of freedom. Oxford University Press, Oxford
Roberts MA (2015) The nonidentity problem. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Winter 2015 Edition). https://plato.stanford.edu/archives/win2015/entries/nonidentity-problem
Schaffer J (2016) The metaphysics of causation. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Fall 2016 Edition). https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/
Shiffrin S (1999) Wrongful life, procreative responsibility, and the significance of harm. Legal Theory 5(2):117–148
Shiffrin S (2012) Harm and its moral significance. Legal Theory 18:357–398
Tadros V (2014) What might have been. In: Oberdiek J (ed) Philosophical foundations of the law of torts. Oxford University Press, Oxford, pp 171–192
Thomson J (1990) The realm of rights. Harvard University Press, Cambridge
Thomson J (2011) More on the metaphysics of harm. Philos Phenomenol Res 82(2):436–458
Wallace DF (1999) Brief interviews with hideous men. Little, Brown, and Company, Boston
Woodward J (1986) The non-identity problem. Ethics 96(4):804–831
Woodward J (1987) Reply to Parfit. Ethics 97(4):800–816
Woollard F (2012) Have we solved the non-identity problem? Ethical Theory Moral Pract 15(5):677–690