CAMBRIDGE
UNIVERSITY PRESS

REPLIES

# Doing Harm: A Reply to Klocksiem

Erik Carlson [ID], Jens Johansson* [ID] and Olle Risberg [ID]

Uppsala University, Uppsala, Sweden
*Corresponding author. E-mail: jens.johansson@filosofi.uu.se

**Abstract**

In a recent article in this journal, Justin Klocksiem proposes a novel response to the widely discussed *failure to benefit* problem for the *counterfactual comparative account of harm* (CCA). According to Klocksiem, proponents of CCA can deal with this problem by distinguishing between facts about there being harm and facts about an agent's having done harm. In this reply, we raise three sets of problems for Klocksiem's approach.

## 1. Introduction

In a recent article in this journal, Justin Klocksiem proposes a novel response to the widely discussed *failure to benefit* problem for the *counterfactual comparative account of harm*.[1] This account can be formulated as follows:

> *The Counterfactual Comparative Account of Harm (CCA)*: An event *e* harms a person S if and only if S would have been better off had *e* not occurred.[2]

The failure to benefit problem is that CCA classifies some actions that intuitively merely fail to benefit a person as harming her. Klocksiem focuses on the following case:

> *Golf Clubs*: Batman buys some golf clubs for Robin, but the Joker convinces Batman to keep them. Batman would have given the clubs to Robin had he not decided to keep them.[3] (p. 430)

Klocksiem writes:

> On the plausible assumption that Robin would have been better off had Batman given him the golf clubs as planned, CCA implies that Batman's keeping them for himself harms Robin. But this seems incorrect; although Robin is not

---

[1]Klocksiem (2022). Parenthetical page references will refer to this article.

[2]CCA is defended in e.g. Bradley (2009); Feit (2015), (2019), (2022); Klocksiem (2012); Timmerman (2019).

[3]The case is taken from Bradley (2012, p. 397). For previous discussions of the failure to benefit problem, see e.g. Feit (2019); Hanna (2016); Immerman (2022); Johansson and Risberg (2020); Purves (2019).

benefitted, this does not mean he is harmed, and the claim that Batman has harmed him seems preposterous. (p. 430)

However, Klocksiem argues that proponents of CCA can deal with this problem by distinguishing 'between facts about the existence of harm and facts about attributions of having done harm to persons or other responsible entities' (p. 432). According to Klocksiem, while CCA does imply that Robin is harmed, this is acceptable so long as we can avoid the result that Batman does harm to, or harms, Robin. By showing that Batman is neither causally nor morally responsible for the harm that Robin suffers, Klocksiem argues, we can indeed avoid the latter result.

   We shall argue that Klocksiem's approach is unsuccessful. After presenting its main features (section 2), we shall criticize it on three grounds. To begin with, it arguably fails to properly address the problem it was supposed to solve, and it presupposes an implausible view of what it is to do harm (section 3). Moreover, while Klocksiem emphasizes that CCA does not have the unacceptable result that a person is harmed every time she is not benefitted, his own approach has precisely that result (section 4). Finally, Klocksiem's approach has implausible implications in cases where an agent makes someone well off but would otherwise have made her even better off (section 5).

## 2. Klocksiem's approach

According to Klocksiem, earlier discussions of the failure to benefit problem have suffered from a neglect of the distinction between what it takes for there to *be* harm and what it takes for an agent (or other object) to *do* harm. CCA, he contends, is solely concerned with the former issue – and what it says about it is, moreover, entirely correct. The correct view about the latter issue, by contrast, is the following:

   *Harming is Responsibility (HiR)*: An agent (or other object), $S_1$, does harm to $S_2$ if and only if there exists an event, $e$, such that $e$ is a final harm for $S_2$, and $S_1$ is causally or morally responsible for $e$.[4] (p. 436)

A final harm, Klocksiem explains, is the same as a *non-instrumental* harm. Assuming CCA, he says (p. 429), an event $e$ is an instrumental harm for a person just in case (i) she would have been better off had $e$ not occurred, and (ii) it is because $e$ causes or otherwise produces some *other* event $e^*$, such that the person would have been better off had $e^*$ not occurred, that (i) holds. An event $e$ is a final harm for a person just in case it satisfies (i) but not (ii). As these claims strongly suggest, Klocksiem contends that any instrumentally harmful event owes its harmfulness to its connection with some finally harmful event – that is, at least one relevant $e^*$ is a final harm. Klocksiem illustrates:

   For example, stubbing one's toe is an instrumental harm. Although I would have been better off had I not stubbed my toe, this is because stubbing my toe produced an episode of pain, and I would have been better off had I not suffered that episode of pain … The episodes of pain that are associated with or caused by toe stubbings, by contrast, are non-instrumental harms. Events such as these are, in some sense,

---

[4]Though he does not say so explicitly, Klocksiem apparently treats the expressions '$S_1$ does harm to $S_2$' and '$S_1$ harms $S_2$' as equivalent. We shall follow him in doing so.

directly harmful – I am better off if I do not suffer the episode of pain, and this is not because the episode of pain is connected to some further event such that I would have been better off had *that* event not occurred. (p. 429)

Despite what this passage may seem to suggest, however, it is not the case that an event is a final harm for a person in Klocksiem's sense just in case it is intrinsically (or finally) bad for that person. For example, Klocksiem suggests that a person's death can harm her in virtue of the fact that it results in 'deprivations of well-being or other goods' (p. 438, n. 26), with those deprivations being the relevant final harms. On all standard views of well-being, such deprivations are not themselves intrinsically (or finally) bad for the person who dies.

Klocksiem proposes the following principle of causal responsibility:

*Causal Responsibility (CR)*: An agent (or other object), $S_1$, is causally responsible for a final harm, $e$, suffered by $S_2$ if and only if $S_1$ initiates or makes a sufficiently significant causal contribution to a causal chain that results in $e$. (p. 437)

While Klocksiem leaves many key notions in CR unspecified, he stresses that making only a minor causal contribution to a final harm – as opposed to, for instance, causing someone pain by punching her in the face – is not enough to be causally responsible for that harm on CR.

Klocksiem's principle of moral responsibility is the following:

*Moral Responsibility (MR)*: An agent, $S_1$, is morally responsible for a final harm, $e$, suffered by $S_2$ if either $S_1$ is causally responsible for $e$ in a way that violates one or more of $S_1$'s moral obligations, or $S_1$ failed to satisfy a moral obligation to prevent $e$, or to reduce its severity.[5] (p. 438)

According to Klocksiem, these three principles – HiR, CR, and MR – together allow proponents of CCA to avoid the result that Batman does harm to Robin in *Golf Clubs*. To begin with, CR is not satisfied, as 'Batman's role in bringing about the harm to Robin is passive; Batman does not make an active contribution to a causal chain that affects Robin's well-being' (p. 439). According to Klocksiem, even if Batman does play an active and substantial role in causing various events that harm Robin on CCA – such as Batman's decision to keep the clubs, and his keeping the clubs – those events are only *instrumental* harms for Robin. Due to CR's restriction to *final* harms, then, Batman's playing an active and substantial role in causing those events does not imply that CR is satisfied. Furthermore, Klocksiem suggests that on a natural understanding of *Golf Clubs*, MR is not satisfied – there is no particular reason to think that Batman violates any moral obligation. By HiR, then, since Batman is neither causally nor morally responsible for any event that is a final harm for Robin, Batman does not do harm to Robin.

---

[5]We shall assume that although MR as formulated by Klocksiem provides only a sufficient condition, it should be taken to state also a necessary condition. Obviously, Batman's not satisfying a merely sufficient condition for being morally responsible for a final harm does not rule out his being morally responsible for it.

## 3. Klocksiem's solution to the failure to benefit problem

It is not entirely clear how Klocksiem's approach is supposed to solve the failure to benefit problem. Those who find CCA's verdict implausible in *Golf Clubs* are unlikely to object only or primarily to the conclusion that Batman, the agent, does harm to (or harms) Robin. The failure to benefit problem is usually framed in terms of an *event* harming a person.[6] Thus, what critics of CCA typically find implausible in *Golf Clubs* is the conclusion that Batman's *decision* harms Robin. This conclusion remains true on Klocksiem's approach, since his approach includes CCA. It may thus seem that his approach fails to address the main part of the failure to benefit problem.

Klocksiem could, however, be interpreted as offering a debunking explanation of the intuition that Batman's decision does not harm Robin. This intuition is, he might claim, the result of a failure to distinguish between facts about there being harm and facts about doing harm. He might argue that once we are made aware of this distinction, and realize that it can be true that Batman's decision harms Robin even though it is false that Batman does harm to Robin, the intuition is dispelled.

An immediate worry about this debunking strategy is that it rests on a questionable explanation of the intuition that Batman's decision is harmless. We suspect that many would still object to the conclusion that Batman's decision harms Robin, even if they became convinced that Batman does not do harm to Robin.

A more serious problem for Klocksiem's strategy, however, is that it assumes a highly problematic view of the relation between an event's harming a person and an agent's doing harm to (or harming) that person. The reason is that it is implausible that $S_1$ may *do* something which is a *harm* for $S_2$, without thereby *harming* or *doing harm* to $S_2$.[7] This can be brought out in two closely related ways.

First, for any term 'X' that is used both as a noun and as a verb, it seems obvious that if there is an X which S performs, then S thereby Xs. Thus if there is a kick (run, shout, etc.) which S performs, then S kicks (runs, shouts, etc.). And it would be very surprising if 'harm' turned out to be an exception to this rule. Admittedly, since the verb 'harm' is normally used transitively, the phrase 'S harms' is somewhat anomalous, without a specified object of the harm. But the same is true of, for example, 'S insults' and 'S praises', and presumably nobody would consider this as evidence that 'insult' or 'praise' are exceptions to the general rule stated. Note also that 'S does harm' seems quite regular, even without a specified object.

Second, the inference from '$S_1$ does something which is a harm to $S_2$' to '$S_1$ does a harm to $S_2$' seems clearly valid, regardless of whether the harm in question is final or instrumental. The same holds for the inference from '$S_1$ does a harm to $S_2$' to '$S_1$ does harm to $S_2$'. From '$S_1$ does something which is a harm to $S_2$', then, we can infer '$S_1$ does harm to $S_2$'.

The upshot of these considerations is that it cannot plausibly be maintained both that Batman's decision harms Robin in *Golf Clubs*, as CCA implies, and that Batman

---

[6]For example, Hanna's (2016) discussion of the problem focuses on whether the event of Batman's changing his mind harms Robin (Hanna 2016, p. 253), and Purves's (2019) discussion focuses on whether Batman's action of keeping the clubs harms Robin (Purves 2019, p. 2632). While Johansson and Risberg (2020) discuss slightly different examples, they also focus on CCA's implications about certain intuitively harmless actions. Bradley's (2012) presentation of the problem involves both claims about whether Batman harms Robin and claims about whether certain events harm Robin (Bradley 2012, p. 397).

[7]Klocksiem mentions this objection (p. 442), but does not consider the support for it that we provide below.

does not do harm to Robin, as Klocksiem takes HiR to imply. This combination of views is inconsistent with the validity of the rules just cited. Since this combination of views is essential to Klocksiem's solution to the failure to benefit problem, we conclude that this solution fails.

Indeed, the foregoing also provides more direct reasons to think that CCA proponents are, after all, forced to accept that Batman does harm to (and harms) Robin. For, again, CCA implies that Batman does something – deciding against giving Robin the clubs – that is a harm to Robin. By the rules cited, Batman thereby does harm to (and harms) Robin.

## 4. Negative events

We have argued that Klocksiem's approach fails to provide a plausible solution to the failure to benefit problem in its standard form. In this section and the next, we shall argue that Klocksiem's approach, to a much greater extent than CCA, also leads to two even more serious versions of the failure to benefit problem.

According to Klocksiem (p. 434), it is important to note that CCA does not have the unacceptable implication that a person is harmed every time she is not benefitted. Clearly, any view that has that implication thereby faces an extreme version of the failure to benefit problem. Ironically, however, Klocksiem's own approach seems to have precisely this implication.

The point that CCA lacks this implication – and thereby avoids colossal overgeneration of harm – is a familiar one.[8] In order for CCA to imply that a person is harmed, there has to be some event such that the person would have been better off had it not occurred. In many cases in which a person is not benefitted, however, there seems to be no event that satisfies this condition. Indeed, even events such as decisions not to benefit a person often fail to satisfy this condition. Klocksiem offers the following illustration:

> *Fifty Dollars*: Clark has a $50 bill in his pocket, and for no particular reason, contemplates whether to give it to Jimmy. Clark decides not to give the $50 bill to Jimmy. (p. 434)

Assuming that Jimmy would have been better off if Clark had given him the $50 bill, CCA may seem to obviously entail that Jimmy is harmed – in particular, by Clark's decision not to give him the $50 bill. But it might well be false that if Clark had not made this decision, then he would have given Jimmy the $50 bill – perhaps he would instead have remained indecisive or forgotten about the whole thing.[9] Moreover, there is no particular reason to think that any other event in the case leaves Jimmy worse off than he would have been had it not occurred. As a result, Jimmy might well fail to be harmed on CCA.

In our view, this is a plausible line of reasoning. It is important to note, however, that it presupposes the non-existence of 'negative' events consisting in a person's *not* having

---

[8]See e.g. Feit (2015, p. 385), Hanna (2016, p. 252).

[9]See especially Hanna (2016, p. 252). Klocksiem himself says that '[i]n ordinary circumstances, the nearest possible worlds in which it is not the case that Clark decides not to give Jimmy the $50 bill are worlds in which Clark does not even consider giving it to him' (p. 435). This seems to presuppose objectionable backtracking.

a certain property, and such that the person would have had that property if the negative event in question had not occurred. (It may seem obvious that any putative negative event of the form *S's not being F* is such that S would have been F if that event had not occurred. But this is not entirely clear. For example, a possible view is that any negative event of the form *S's not being F* consists in some positive – perhaps very complex – event of the form *S's being G*, where the fact that S is G is what explains why S is not F. And it may not always be true that if S had not been G, then S would have been F. There are of course further questions about how the distinction between positive and negative events is to be drawn in a principled way, but for present purposes, these questions can be set aside.) Suppose, for example, that there is an event such as *Jimmy's not receiving the $50 bill*, and that Jimmy would have received the $50 bill (and hence would have been better off) if this event had not occurred. On this supposition, CCA entails that this event harms Jimmy. More generally, given the existence of negative events like this one, CCA does imply, after all, that a person is harmed every time she is not benefitted. In fact, given the existence of such events, CCA also has the even more implausible implication that a person is harmed virtually every time she *is* benefitted; for example, even if Jimmy had received the $50 bill, there would have been an event such as *Jimmy's not receiving a $1000 bill*, which would have harmed him according to CCA.

The problem for Klocksiem is that negative events of this sort seem to be required in order for his approach to avoid another problem. As noted in section 2, Klocksiem recognizes that CCA implies that Batman's decision to keep the clubs, as well as various other events, harms Robin in *Golf Clubs*. As Klocksiem apparently also recognizes, Batman obviously makes a very significant causal contribution to a causal chain – in particular, to his own process of deliberation – that results in his decision. Hence, if Batman's decision were a *final* harm for Robin, HiR and CR would jointly imply that Batman does harm to Robin – the verdict that Klocksiem's approach is primarily designed to avoid. Of course, Klocksiem contends that Batman's decision is *not* a final harm, but an *instrumental* harm (see section 2). In that case, Batman's significant role in causing his decision does not imply that he does harm to Robin on HiR and CR.

Here, however, it is crucial to return to Klocksiem's view of instrumental and final harm (see again section 2). To repeat, according to him, every instrumental harm owes its harmfulness to its causing or otherwise producing some final harm. Assuming CCA, moreover, an event *e* is an instrumental (final) harm for a person just in case *e* satisfies CCA's condition, where this is (is not) because *e* produces some other event *e\** which satisfies CCA's condition. Now what could be a relevant final harm for Robin in *Golf Clubs*? No 'positive' event seems to be a promising candidate. To begin with, there is no reason to suppose that Batman's decision produces any event that is intrinsically bad for Robin, such as a painful experience. If it did, *Golf Clubs* would not even initially appear to constitute a problem for CCA. What about an event such as *Robin's continuing to use his old golf clubs*? Even supposing that this event occurs, and that it is in some sense produced by Batman's decision – we might assume, at least, that it would not have occurred had Batman not decided to keep the clubs (as Robin would have used the new ones instead) – there is no reason to suppose that Robin would have been better off had this event not occurred. In particular, it would presuppose objectionable backtracking to claim that if Robin had not continued to use his old clubs, then Batman would have given him the new clubs earlier. It is much more natural, and requires a considerably smaller divergence from the actual past, to suppose that in the nearest possible world in which Robin does not continue to use his own clubs, he uses some third set of

clubs instead, or refrains from playing golf (and Batman keeps his clubs). Similar remarks apply to other positive events to which Klocksiem might try to appeal – such as, for example, *Robin's being happy to degree 5* or *Batman's playing golf with his new clubs* (even assuming that they occur).

Given the existence of *negative* events of the sort described above, on the other hand, it does seem possible for Klocksiem to provide a relevant final harm – that is, one that the harmfulness of Batman's decision could derive from, on his approach. Suppose that Robin would have been happy to degree 10 had Batman not made his decision. Given the existence of negative events of the pertinent sort, one event that occurs is $e^* =$ *Robin's not being happy to degree 10*. Plausibly, while $e^*$ satisfies CCA's condition, this is not because it causes or otherwise produces some other event which satisfies CCA's condition. Moreover, it seems plausible to say that Batman's decision in some sense produces $e^*$ (at least, $e^*$ would not have occurred had Batman not made this decision) – and also that this is why Batman's decision satisfies CCA's condition. In this way, Batman's decision avoids being a final harm: it owes its harmfulness to its connection with $e^*$, which *is* a final harm.[10] Consequently, Batman's active and substantial role in causing his decision does not imply that he does harm to Robin on HiR and CR. Obviously, however, this line of reasoning presupposes the existence of negative events of the pertinent sort. And as already explained, given the existence of negative events of this sort, Klocksiem is, in the end, committed to the unacceptable claim that a person is harmed every time she is not benefitted – and, indeed, at virtually all times at which she *is* benefitted.

## 5. A further problematic case

Klocksiem's approach also faces yet another version of the failure to benefit problem – a version that, like the one in the previous section, seems even more serious than the standard one (involving cases like *Golf Clubs*), and more serious for Klocksiem's approach than for CCA. This version concerns the implications of Klocksiem's approach for a case that is very similar to *Golf Clubs*. Consider:

> *More Golf*: Batman has bought two sets of golf clubs – one that is of extremely high quality and one that is of slightly lower quality, but still very good. Now he has three options: (A1) to give Robin the slightly inferior clubs; (A2) to give Robin the better clubs; and (A3) to keep all the clubs for himself. Batman performs (A1), which makes Robin's well-being level increase significantly. If Batman had not done so, he would have given Robin the even better set of clubs, whereby Robin's well-being level would have been even higher.[11]

Klocksiem's approach seems to imply that Batman does harm to Robin in *More Golf*. To begin with, CCA implies (counterintuitively) that Batman's action of giving Robin the slightly inferior clubs harms Robin, as Robin would have been better off if Batman had not performed that action. Moreover, Batman undeniably makes a very

---

[10]One might worry that Batman is causally responsible for $e^*$ itself, in which case HiR implies that he does harm to Robin after all. As we understand Klocksiem, however, he takes the causal connection that needs to hold between instrumental and final harms to be considerably weaker than the one that CR's condition requires to hold between agents and final harms.

[11]This case is taken, with minor modifications, from Johansson and Risberg (2020, p. 1543).

significant causal contribution to a causal chain that results in this action. Hence, if this action is a *final* harm for Robin, HiR and CR together imply that Batman does harm to Robin.

Of course, Klocksiem might suggest that Batman's action in *More Golf* is only an *instrumental* harm for Robin. Given his view of instrumental and final harm, this requires the harmfulness of Batman's action to derive from some final harm which the action causes or otherwise produces. One candidate for such a final harm is $e^* = Robin's\ becoming\ the\ owner\ of\ the\ slightly\ inferior\ golf\ clubs$. Plausibly, Robin would have been better off if $e^*$ had not occurred (as he would then have received the even better clubs instead). Moreover, this does not seem to be because $e^*$ causes or otherwise produces any further event that harms Robin according to CCA.[12] By taking $e^*$ to be a final harm for Robin, then, Klocksiem avoids having to say that Batman's action is a final harm for Robin. However, his view still implies that Batman does harm to Robin. Since Batman undeniably makes a very significant causal contribution to a causal chain that results in $e^*$, HiR and CR together imply that Batman does harm to to Robin.

The implication that Batman does harm to Robin in *More Golf* is problematic for Klocksiem's approach for several reasons. For one, this result is counterintuitive in its own right. Moreover, because Klocksiem's approach implies that Batman does harm to Robin in *More Golf*, it provides no help for proponents of CCA in dealing with the theory's problematic implication that Batman's action harms Robin in that case. In particular, the type of debunking explanation considered in section 3 – namely, that this intuition rests on a confusion of what it is for an event to harm someone with what it is for an agent to do harm to someone – is unavailable, since Klocksiem's approach and CCA together imply both that Batman does harm to Robin and that his action harms Robin. Klocksiem's defence of CCA thus fails to generalize.

Last but not least, *More Golf* reveals that Klocksiem's approach is committed to several implausible combinations of verdicts. It implies that Batman does harm to Robin in *More Golf* but not in *Golf Clubs* – a pair of views which seems very hard to support on independent grounds. (Indeed, if it is at all justifiable to treat the two cases differently, it would surely be more plausible to claim that Batman harms Robin in *Golf Clubs* but not in *More Golf*.) It also implies that although Batman actually does harm to Robin in *More Golf*, by giving him the slightly inferior golf clubs, Batman would not have done harm to Robin if he had kept all the clubs for himself. This is absurd – especially from the perspective of a counterfactual comparative view of harm – since the former action leaves Robin much better off than the latter would have done. Proponents of CCA, on the other hand, easily avoid commitment to these combinations of views – unless, of course, they adopt Klocksiem's attempt to defend it.[13]

**Competing Interests.** The authors declare there are no competing interests.

---

## References

**Bradley B.** (2009) *Well-Being and Death*. New York: Oxford University Press.

**Bradley B.** (2012) Doing Away with Harm. *Philosophy and Phenomenological Research* **85**(2): 390–412.

**Feit N.** (2015) Plural Harm. *Philosophy and Phenomenological Research* **90**(2): 361–88.

**Feit N.** (2019) Harming by Failing to Benefit. *Ethical Theory and Moral Practice* **22**(4): 809–23.

**Feit N.** (2022) How Harms Can Be Better than Benefits: Reply to Carlson, Johansson, and Risberg. *Australasian Journal of Philosophy* **100**(3): 628–33.

**Hanna N.** (2016) Harm: Omission, Preemption, Freedom. *Philosophy and Phenomenological Research* **93** (2): 251–73.

**Immerman D.** (2022) The Worse than Nothing Account of Harm and the Preemption Problem. *Journal of Moral Philosophy* **19**(1): 25–48.

**Johansson J., and O. Risberg** (2020) Harming and Failing to Benefit: A Reply to Purves. *Philosophical Studies* **177**(6): 1539–48.

**Klocksiem J.** (2012) A Defense of the Counterfactual Comparative Account of Harm. *American Philosophical Quarterly* **49**(4): 285–300.

**Klocksiem J.** (2022) Harm, Failing to Benefit, and the Counterfactual Comparative Account. *Utilitas* **34**(4): 428–44.

**Purves D.** (2019) Harming as Making Worse Off. *Philosophical Studies* **176**(10): 2629–56.

**Timmerman T.** (2019) A Dilemma for Epicureanism. *Philosophical Studies* **176**(1): 241–57.