Reply to Klockslem on the Counterfactual Comparative Account of Harm

Erik Carlson¹

Accepted: 16 February 2020 / Published online: 26 February 2020
© The Author(s) 2020

Abstract

In a recent article in this journal, I claimed that the widely held counterfactual comparative account of harm (CCA) violates two very plausible principles about harm and prudential reasons. Justin Klockslem argues, in a reply, that CCA is in fact compatible with these principles. In this rejoinder, I shall try to show that Klockslem's defense of CCA fails.

Keywords Harm · Benefit · Counterfactual comparative account · Reason for action

The most widely held account of harm and benefit is the counterfactual comparative account:

CCA: A possible event or action would harm (benefit) a person if and only if she is worse (better) off in the nearest possible world, w , where the event occurs or the action is performed, than in the world nearest to w where the event does not occur or the action is not performed.¹

In a recent article in this journal, I claimed that CCA violates two very plausible principles about harm and prudential reasons.² Justin Klockslem argues, in a reply, that CCA is in fact compatible with these principles.³ In this rejoinder, I shall try to show that Klockslem's defense of CCA fails.

¹For discussion and defenses of CCA, see, e.g., Boonin 2014; Bradley 2009; Feit 2015, 2016, 2019; Klockslem 2012; Parfit 1984, p. 69; Timmerman 2019. Reasons for formulating the account in this slightly cumbersome way are given in Carlson 2019, p. 796.

²Carlson 2019.

³Klockslem 2019.

✉ Erik Carlson
erik.carlson@filosofi.uu.se

¹ Uppsala University, Uppsala, Sweden

1 Reason for Action

The first principle I discussed is as follows:

Reason for Action (RfA): If a and a^* are alternative actions open to you in a choice situation, and doing a would benefit (not harm) you, whereas doing a^* would not benefit (harm) you, then you have a prudential reason to do a , rather than a^* .⁴

I relied on an intuitive understanding of ‘alternative actions’, and did not provide a definition of this term. We may assume, however, that two or more actions are alternatives, in the relevant sense, just in case they are mutually exclusive and performable by the same agent in the same situation.

I find RfA undisputable. It seems evident that you have reason to choose an option that would benefit you, rather than one that would harm you, or at least not benefit you. Similarly, you have reason to choose an option that would not harm you, rather than one that would. To deny these claims would appear to reveal a lack of understanding of the notions of harm and benefit. Hence, an account of harm and benefit should, in order to be acceptable, be compatible with RfA.

In my paper, I took the following case to show that CCA violates RfA:

Darts: You can throw a dart at either of two boards. Board 1 is, unlike board 2, surrounded by a thin circle of a different colour. You will be best off if you hit the circle, second-best off if you hit board 1, third-best off if you hit board 2, and least well off if you neither hit one of the boards nor hit the circle. You are sufficiently good at darts to hit either of the boards. However, you are not good enough to hit the circle at will. If you aim for it, it is likely that you will miss both it and board 1. In the nearest possible world where you hit board 1, w_{b1} , it is nevertheless true that if you were not to do this, you would (accidentally) hit the circle. In the nearest possible world where you hit board 2, w_{b2} , it is true that if you were not to do this, you would miss both boards, as well as the circle.⁵

In w_{b1} it is true that if you were not to hit board 1, you would be better off. In w_{b2} it is true that if you were not to hit board 2, you would be worse off. CCA thus implies that hitting board 1 would harm you, while hitting board 2 would benefit you. But, since hitting board 1 would make you better off than hitting board 2, you have no prudential reason to hit board 2 rather than board 1. Hence, CCA violates RfA.

Klocksiesm claims, however, that distinguishing between actions and outcomes reveals that CCA does not, in fact, violate RfA in *Darts*. The alternative actions in *Darts* are, according to Klocksiesm, $a_1 = \text{throwing at board 1}$, $a_2 = \text{throwing at board 2}$, and $a_3 = \text{throwing at the circle}$, while *hitting board 1*, *hitting board 2*, and *hitting the circle* are possible outcomes of these actions. Whereas hitting board 1 harms you, according to CCA, throwing at board 1 benefits you, since “you are worse off in the nearest world in which you do something other than [a_1]”.⁶ Moreover, “[y]ou have reason to perform a_1 , which is the action that CCA classifies as a benefit; you have reason not to perform a_2 , which is the action CCA classifies as a harm”.⁷ Hence, Klocksiesm concludes, CCA does not violate RfA in *Darts*.

⁴ Carlson 2019, p. 797. I did not assume that this prudential reason must be conclusive.

⁵ Carlson 2019, p. 798.

⁶ Klocksiesm 2019, p. 676.

⁷ Klocksiesm 2019, p. 676.

Hitting a certain board is not an action, according to Klocksien, since “whether you hit it is not under your direct control, and [...] it is possible for you to throw at a board and miss”.⁸ This argument is problematic. First, the possibility of missing the board does not indicate that hitting it is not an action. Even attempts at arguably ‘basic’ actions, such as raising one’s arm, sometimes fail. Second, Klocksien does not explain what he means by ‘direct control’. If you are good at darts, why is it not under your direct control whether you hit the board? Granted, the success of your attempt to hit the board partly depends on background factors beyond your control. (The room must not be too dark, the board must not fall off its hook while your dart is in the air, etc.) But, again, this is true of almost any action. Success in your attempt to raise your arm requires that you do not suddenly become paralysed, that nobody grabs your arm and forces it down, and so on. A challenge for Klocksien is therefore to provide a definition of ‘direct control’ that excludes hitting a dart board from being an action without absurdly excluding raising your arm, as well.

However, we may for the sake of argument accept Klocksien’s distinction between actions and outcomes, and also grant that hitting a dart board is not an action. This does not really matter, since we can simply choose another example, to show that CCA violates RfA. Consider:

Buttons: On a board in front of you, there are four buttons, B1 to B4, any one of which you can easily press. Pressing B1 would be very good for you, and pressing B2 would be slightly less good. Pressing B3 would be very bad for you, and pressing B4 would be even worse. In the nearest possible world where you press B2, it is true that if you had not done so, you would have pressed B1. Further, in the nearest possible world where you press B3, it is true that if you had not done so, you would have pressed B4.

I trust that Klocksien agrees that pressing a button is typically an action. A distinction between actions and outcomes that implies the opposite has little plausibility. If this is granted, pressing B1, pressing B2, pressing B3, and pressing B4, are also *alternative* actions, since they are, by assumption, performable and mutually exclusive. CCA implies that pressing B2 would harm you, while pressing B3 would benefit you. Since pressing B2 would make you better off than pressing B3, however, you have no prudential reason to press B3 rather than B2. CCA therefore violates RfA.

To make the relevant counterfactuals plausible, suppose, for example, that you can reach B1 and B2 most easily with your left hand, while B3 and B4 are most easily reached with your right hand. Suppose you just pick a button, say B2. (Maybe you do not at the moment of choice care much about your well-being, or maybe you are unaware of the effects of pushing the buttons.) Had you not pressed B2, you would still have used your left hand and pressed B1. Had you pressed B3, on the other hand, it would have been true that if you had not done so, you would still have used your right hand and pressed B4.

We may note, moreover, that somewhat weaker counterfactual assumptions are sufficient for CCA to violate RfA in *Buttons*. If it is true in the nearest B2-world that you would press B1, were you not to press B2, it suffices that it is true in the nearest B3-world that you *might* press B4, were you not to press B3. Similarly, if it is true in the nearest B3-world that you would press B4, were you not to press B3, it suffices that it is true in the nearest B2-world that you *might* press B1, were you not to press B2. Under the former assumptions, CCA entails that pressing B2 would harm you while pressing B3 would not. Under the latter, CCA entails that pressing B3 would benefit you while pressing B2 would not. Either conclusion is a violation of RfA.

⁸ Klocksien 2019, p. 676.

An anonymous reviewer has suggested that it might be possible to revise either CCA or RfA, in response to my objection. In particular, the reviewer claims that, in *Buttons*, “[pressing] B2 harms you when compared to [pressing] B1; [pressing] B3 benefits you when compared to [pressing] B4”.⁹ If this remark is meant to indicate a way to revise CCA, it seems to suggest a “contrastive” version of the account, according to which harm is a three-place relation (involving an event, a person, and a contrast event), rather than, as the standard CCA assumes, a binary relation (involving an event and a person). On such a contrastive account an event never harms a person *simpliciter*; it can only harm her in comparison to some contrast event. I discussed contrastive CCA in my paper, and concluded that, although it may be compatible with (a contrastive version of) RfA, it nevertheless fails to capture the prudential relevance of harm.¹⁰ Moreover, Klocksiem has stated, in personal communication, that he, at least, is not proposing a contrastive version of CCA. My paper also discussed several other ways to revise CCA.¹¹ I do not claim to have shown that there is no possible revision of CCA that satisfies RfA and is otherwise plausible, but I think it is fair to conclude that the prospects for finding such a revision are not very bright.

As regards revising RfA, a proponent of CCA might claim that only the following, logically weaker principle is true:

Revised RfA: If a and a^* are alternative actions open to you in a choice situation, (i) doing a would benefit (not harm) you, whereas doing a^* would not benefit (harm) you, and (ii) you would do a^* if you were not to do a and vice versa, then (iii) you have a prudential reason to do a , rather than a^* .

The addition of clause (ii) to the antecedent makes Revised RfA compatible with CCA. Since it is false that you would press B3 if you were not to press B2, in *Buttons*, Revised RfA does not entail that you have a prudential reason to press B3 rather than to press B2.

The problem with this move is that RfA seems just as plausible as Revised RfA. The plausibility of RfA only requires that a and a^* are actions you *can* perform in a given situation. Unless the defender of CCA comes up with a reason to doubt RfA that is independent of the truth of CCA, the claim that RfA must be weakened is question-begging.

A third reply open to a defender of CCA is to claim that if a and a^* are alternatives for you in a certain situation, and it is true that you would do a^* if you were not to do a , then a and a^* are the only available alternatives. Or, equivalently, if there are more than two alternative actions available to you, then no ‘would’-counterfactual of the form “If you were not to do a , then you would do a^* ” is true. If this is correct, the counterfactuals in virtue of which CCA classifies pressing B2 as a harm and pressing B3 as a benefit are incompatible with the assumption that for each of B1, B2, B3 and B4, pressing it is an available alternative.

However, the suggested claim is implausible. Consider the following example. At the lunch restaurant on my campus they always serve a meat dish, a fish dish, and a vegetarian dish. Today, I chose the vegetarian option. And I am sure that if I had not, I would have chosen the fish. (I do not eat meat but I sometimes eat fish.) But I am also quite sure that choosing meat was an available option for me; i.e., something I *could* do, although it would have meant acting out of character.

⁹ Klocksiem 2019, p. 675, makes a similar remark regarding the *Coins* example (see section 2, below).

¹⁰ Carlson 2019, pp. 805–806.

¹¹ Carlson 2019, pp. 799–803.

2 Reason for Preference

The second principle I claimed to be incompatible with CCA is the following:

Reason for Preference (RfP): If possible events e and e^* are alternative outcomes in a present or future situation, and e would benefit (not harm) you, were it to occur, whereas e^* would not benefit (harm) you, were it to occur, then you have a prudential reason to prefer that e occurs, rather than that e^* occurs.¹²

I took e and e^* to be ‘alternative outcomes’ just in case they have the same temporal location, and the relevant e -world is identical to the relevant e^* -world, as regards particular facts, up to the time at which e or e^* may occur.¹³

Like RfA, RfP is surely a very plausible principle. I claimed that CCA violates RfP in the following case:

Coins: A coin-tossing mechanism selects and tosses one of two coins. If the mechanism tosses coin 1, you will be very well off if it lands heads, and slightly less well off if it lands tails. If the mechanism tosses coin 2, you will be very badly off if it lands heads, and slightly less badly off if it lands tails. Let ‘Heads₁’, ‘Tails₁’, ‘Heads₂’, and ‘Tails₂’, refer to the respective outcomes. In the nearest Tails₁-world, it is true that Heads₁ would occur, were Tails₁ not to occur. Similarly, in the nearest Tails₂-world, it is true that Heads₂ would occur, were Tails₂ not to occur.¹⁴

CCA implies that Tails₁ would harm you, whereas Tails₂ would benefit you. Clearly, however, you have no prudential reason to prefer Tails₂ to Tails₁. RfP is thus violated.

Klocksiem objects that Tails₁ and Tails₂ are not alternative outcomes, according to my definition:

As the case is described, the machine first selects the coin, then flips it. In order for coin 1 to land tails, the machine must have selected coin 1, and this is incompatible with its having selected coin 2. The coin-2 world therefore does not share its history up until the moment that coin 2 lands tails with any coin-1 world. No Tails₂ world is identical with any Tails₁ [world] up to the moment the coin lands, so Tails₁ and Tails₂ are not alternatives.¹⁵

This argument presupposes that the mechanism *first* selects a coin and *later* tosses it. *Coins* is, as I presented it, rather underdescribed, and Klocksiem’s interpretation is perfectly consistent with my description. But what I had in mind was a case where the selecting and the tossing of the coin are more or less simultaneous. Suppose, for example, that you hold a coin in each of your palms, and that somebody activates a mechanism that induces an electric current randomly in either your left or your right arm. The current causes your hand to jerk, and thereby to toss the coin you hold in that hand. Given this elaboration of the case, it is very natural to claim that Tails₁, Tails₂, Heads₁ and Heads₂ are alternative outcomes of activating the mechanism. Moreover, the counterfactual assumptions in *Coins* seem true.

Klocksiem might retort that since it takes a few milliseconds or so for the induced current to make your hand jerk, the shared-history condition in my definition of ‘alternative outcome’ implies that

¹² Carlson 2019, p. 800. As in the case of RfA, I did not assume that this prudential reason is necessarily conclusive.

¹³ Carlson 2019, p. 800.

¹⁴ Carlson 2019, p. 801.

¹⁵ Klocksiem 2019, p. 674.

Tails₁ and Tails₂ are not alternative outcomes. I did not, however, intend such a strict reading of the shared-history condition. And it does not really matter whether or not my definition, as I stated it, implies that Tails₁ and Tails₂ are alternative outcomes. The important issue is whether there is a notion of ‘alternative outcome’, which has this implication and makes RfP plausible. I believe that RfP is very plausible also if we weaken the shared-history condition to require only that the relevant *e*-world is identical to the relevant *e**-world, up to a time *shortly before* the time at which *e* or *e** may occur.

Independently of the truth of RfP, I think there are at least two interrelated reasons against a very strict shared-history condition. First, such a strict condition rules out even paradigmatic cases of alternative outcomes. If a dart is thrown at a board, its missing the board is surely an alternative outcome to its hitting the board. But given the dart’s position a millisecond before it hits the board (supposing this is the actual outcome), it cannot miss it, given that nearby possible worlds cannot contain large-scale violations of the actual laws of nature.

Secondly, a strict shared-history condition does not rhyme well with plausible assumptions about alternative actions and how they relate to alternative outcomes. Two possible actions can clearly be alternatives although they would, for instance, be preceded by different decisions. This means that a strict shared-past condition is misguided in the definition of ‘alternative actions’. Further, Klocksiesm apparently agrees that *a*₁ and *a*₂ are alternative actions in *Darts*. If, as he claims, hitting board 1 and hitting board 2 are outcomes of *a*₁ and *a*₂, respectively, we surely want to say that they are alternative outcomes. It would be strange to deny that such immediate outcomes of alternative actions are alternative outcomes. But the nearest worlds where you hit board 1 are worlds where you aim at board 1, while the nearest worlds where you hit board 2 are worlds where you aim at board 2. Hence, these worlds must differ already somewhat before the dart hits the board.

Finally, it is worth noting that we need not rely on *Coins*, to show that CCA violates RfP. *Buttons* is another case in point. The outcome that B3 is pressed would benefit you, according to CCA, whereas the outcome that B2 is pressed would harm you. But you have no reason to prefer the former outcome to the latter, and it cannot plausibly be denied, I think, that the two outcomes are alternatives.

3 Concluding Remarks

Despite Klocksiesm’s arguments to the contrary, it appears that CCA violates both RfA and RfP. Klocksiesm’s argument for the compatibility of CCA and RfA relies on a highly controversial way of distinguishing between actions and outcomes. I argued, moreover, that CCA violates RfA even if Klocksiesm’s distinction is granted. In trying to show that CCA and RfP are compatible, Klocksiesm presupposes an implausibly strict shared-history condition in the definition of ‘alternative outcomes’.

For CCA to be refuted, it suffices that one of RfA and RfP is correct and incompatible with CCA. Although I find both principles eminently plausible, I am inclined to put somewhat greater weight on RfA than on RfP. It is slightly more controversial that there are reasons for preferences than that there are reasons for actions, and we have, I believe, a firmer intuitive grasp of the notion of alternative actions than on the notion of alternative events or outcomes in general. Even if the cogency of RfP may for these reasons be questioned, RfA seems unassailable.¹⁶

¹⁶ I wish to thank Justin Klocksiesm, Jens Johansson, Olle Risberg, and two anonymous referees for *Ethical Theory and Moral Practice*, for very valuable comments on earlier drafts of this article. I am also grateful for financial support from the Swedish Research Council, Grants 2018–01361 and 2016–01531.

Funding Information Open access funding provided by Uppsala University.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Boonin D (2014) *The Non-Identity Problem and the Ethics of Future People*. Oxford University Press, New York
- Bradley B (2009) *Well-Being and Death*. Oxford University Press, New York
- Carlson E (2019) More Problems for the Counterfactual Comparative Account of Harm and Benefit. *Ethical Theory Moral Pract* 22:795–807
- Feit N (2015) Plural Harm. *Philos Phenomenol Res* 90:361–388
- Feit N (2016) Comparative Harm, Creation and Death. *Utilitas* 28:136–163
- Feit N (2019) Harming by Failing to Benefit. *Ethical Theory Moral Pract* 22:809–823
- Klocksien J (2012) A Defense of the Counterfactual Comparative Account of Harm. *Am Philos Q* 49:285–300
- Klocksien J (2019) The Counterfactual Comparative Account of Harm and Reasons for Action and Preference: Reply to Carlson. *Ethical Theory Moral Pract* 22:673–677
- Parfit D (1984) *Reasons and Persons*. Oxford University Press, Oxford
- Timmerman T (2019) A Dilemma for Epicureanism. *Philos Stud* 176:241–257

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.