Harm and Causation

ROBERT NORTHCOTT

Birkbeck College

Abstract

I propose an analysis of harm in terms of causation: harm is when a subject is caused to be worse off. The pay-off from this lies in the details. In particular, importing influential recent work from the causation literature yields a contrastive-counterfactual account. This enables us to incorporate harm's multiple senses into a unified scheme, and to provide that scheme with theoretical ballast. It also enables us to respond effectively to previous criticisms of counterfactual accounts, as well as to sharpen criticisms of rival views.

1) CONCEPTUAL GROUNDWORK: THE TWO FACES OF HARM

Begin by asking: what is the difference between harm and mere badness? The answer lies in the fact that harm carries an active connotation. It is something that is done to us, or that is caused by something. This idea seems to be common and is perhaps the biggest motivation for the various comparative theories of harm – because implicit in the active connotation is the notion of *change*, from a non-harmed or pre-harm state to the harmed or harmful one. Indeed, as will be discussed below, a fundamental worry about any *non*-comparative view is precisely that it must be ill suited to capturing this active connotation. Tellingly, for instance, the most recent non-comparative account of harm, namely Matthew Hanser's (2008) event-based one, is taken by him to encompass the notion of *loss* of a good, i.e. again endorsing a notion of change.

I argue that this entanglement of harm with change is best analyzed through the lens of causation. Indeed, such a move seems inevitable, for any change, at least in the

macro-world, is presumably caused. By importing influential recent work from the causation literature, new light is shed on several outstanding issues in the harm debate.

An initial objection is that often we speak of 'causing harm', which suggests that harm itself is distinct from the act of causing it. Hanser (2008) argues that the notion of harm is therefore prior to that of 'suffering' it. Judith Jarvis Thomson (2011), in the most recent comparative account, agrees. But everyday usage is blurry. We can both 'harm' someone and 'cause them harm'. This illustrates how harm has a curious dual usage – it can be either a verb or a noun, and correspondingly can be understood as either a discrete action or an ongoing state. Similarly, to 'suffer' harm can also refer to either (being on the receiving end of) a discrete action or an ongoing state. I conclude that linguistic practice alone does not tell us which notion is prior, the state or the action. We must delve further.

My own guiding approach will be this: at a first pass, to harm is to cause a bad. But, crucially, this first pass itself needs to be refined. The refined version will be that to harm is to cause an *increase* in badness – or, which I take to be equivalent here, a decrease in well-being. It follows that harming is a relation between such a decrease and some cause of that decrease. This will be given expression by a suitably relational semantics. To preview briefly, my eventual definition of harm will be derived from one of causation. In particular, harm will be analyzed as an instance of causation with a decrease in well-being as the effect term. This raises the technical issue of just how a decrease can be an effect term. The answer is, in a way to be elucidated below, to make the effect term *contrastive*: the effect is an actual level of well-being compared to a salient counterfactual level.

What of harm's verb/noun dual usage? A causal definition speaks immediately to the verb half. The noun half I address by defining harm (in the noun sense) to be the result of harm (in the verb sense). A clarification is required here: by the 'result' of harm I mean the actual resultant level of well-being, not the decrease in it. For example, suppose I was harmed by being blinded. The verb sense of harm is the causing of the decrease in my visual capacity; the noun sense is not this decrease, but rather just the blindness itself. The relation between the two senses is thus easily mistaken. The noun sense is *not* just the effect term of the verb sense, because strictly speaking that effect term is contrastive.

In this light, return next to our starting question: what is the difference between harm and mere badness? I take a 'bad' to be synonymous here with an actual low level of well-being, such as, in our example, being blind.[1] This raises an apparent difficulty, for in effect I am defining harm (in the noun sense) to be a level of well-being that is caused[2] – but surely all such levels are caused by something or other so, at least in cases in which the level is low, doesn't a harm therefore simply reduce to a bad? In reply, speaking of a harm instead of a bad in such circumstances serves the pragmatic role of emphasizing (certain aspects of) the bad's causal history. Indeed this emphasis on what *led* to a bad is arguably the whole point of distinguishing harm from mere badness in the first place – recall that our starting point was precisely harm's connotation of change.

Such an analysis of the two faces of harm is consistent with what we find generally when concepts possess this kind of verb/noun duality. 'Erosion', for instance, can refer both to a process and to the result of that process. (The same applies to 'natural selection'.) Thus, poor soil may be both an instance of erosion and also the result of a

process of erosion. Moreover, 'erosion' in the noun sense refers to an actual resultant state, just as harm does. That is, it refers to the actual soil. Finally, the pragmatic role is analogous too. Describing poor soil as 'erosion' rather than 'poor soil' serves precisely to emphasize (certain aspects of) the soil's causal history.

2) CAUSATION AND COUNTERFACTUALS

Most current theories posit a close relation between causation and *counterfactuals*. This reflects the common emphasis on causation's difference-making aspect – intuitively, to be a cause, something must make a difference to its effect, in other words the effect must be different to how it otherwise would have been. Such an understanding of causation is standard in law (Hart and Honore 1985). It is also standard throughout science.[3] Within philosophy, it has been a commonplace since Hume if not earlier. More recently, ever since Stalnaker and Lewis developed a semantics for it more than forty years ago, it has been widely agreed that counterfactual dependence is a sufficient condition for causation. Dispute has focused on whether it might also be necessary and thus constitute a *definition*, as first proposed in Lewis (1973). Either way, it is uncontroversial that in practice causation often connotes counterfactual dependence. Therefore, to say that an event causes us harm commits us (at least typically) to the counterfactual claim that we suffer more harm than we would have done had the event not occurred.

In addition, I will appeal more specifically to a *contrastive*-counterfactual account of causation. To digress briefly, the motivation for such an account is that the truth values of causal claims seem to be sensitive to contrasts, on both the cause and effect sides (see

e.g. Northcott 2008 or Schaffer 2005 for details). On this view, causation therefore takes the general form:

x_A-rather-than-x_C causes y_A-rather-than-y_C

x_C and y_C are the salient contrasts to, respectively, the actual events x_A and y_A.

A contrastive view is implicit in the contemporary Bayes net and causal modeling literatures (Pearl 2000, Spirtes et al. 2000), and arguably is endorsed by experimental practice (Woodward 2003) and by conceptions of causation in statistics too (Northcott 2012). It is also a commonplace in the mainstream literature on probabilistic causation (Hitchcock 1996), and on causation in the law (Hart and Honore 1985). To be sure, David Lewis (2004) and others reject an explicitly contrastive account. In Lewis's view, contrasts instead determine which possible world is nearest, in effect therefore appearing in the pragmatics rather than semantics of causation. This claim is the subject of current dispute. But all that matters for our purposes is that it is therefore accepted widely that contrastive concerns are salient to causal claims one way or another.

Moreover, the literature surrounding harm usually does not distinguish between causation and mere causal *explanation*. Does something cause harm to me or others, or does it merely explain any such harm? The distinction matters here because, at least according to current orthodoxy, causal explanation is much less controversially analyzed contrastive-counterfactually than is causation itself. Among other things, specification of contrasts serves to represent an explanatory context. That is, whether a particular cause is explanatory depends on the explanatory context, and this is represented by the relativization of explanation to contrasts.

The take-home point is that a contrastive-counterfactual approach does not leave us hostage to anything too contentious. For ease of explication, I will phrase matters throughout as being about causation. If desired, the reader may substitute causal explanation for causation, or perhaps interpret all points as bearing only on causal pragmatics rather than semantics. The important thing is that, whichever precise theoretical connection is preferred, one way or another the notion of harm is intimately bound up with contrastive considerations.

3) A DEFINITION OF HARM

Accordingly, our definition of harm will be a contrastive-counterfactual one. Roughly speaking, the cause term in the definition is the harming event (plus contrast), and the effect term the subject's level of well-being (plus contrast). For there to be harm requires that the harming event reduce that level of well-being. Such an understanding of harm is long familiar. The novelty here will come from presenting it in explicitly causal terms, and in particular in contrastive-counterfactual ones.

More formally, for a subject A, cause event c and salient contrast c*, effect event e and salient contrast e*:

A is harmed just in case:

1) c-rather-than-c* causes e-rather-than-e*
2) e leaves A in a worse-off state than e* would have done[4]

Understanding causation contrastive-counterfactually, c* and e* are interpreted as counterfactual events. In turn, this means that the definition is equivalent to:

A is harmed just in case:

1) c* counterfactually entails e*[5]
2) e leaves A in a worse-off state than e* would have

The above is a definition of being harmed. A definition of what it is for you to harm someone follows immediately: we further require that the cause in question is an action of yours (plus salient contrast). Following section 1, the noun sense of harm can also be derived straightforwardly: it is the worse-off state itself. The effect term e-rather-than-e*, meanwhile, represents a change in well-being rather than just some absolute level of it.

It follows from this definition that *harmed* is a relational property, and in particular is relativized to c* and e*. There is no absolute fact of the matter, independent of explanatory context. More formally, whenever we ask whether someone has been harmed, on my view a presupposition of the question is a particular specification of contrasts. That these relativizations are often not explicit does not show that they are not present, only that they are tacit. It also follows from the definition that the same person can be both harmed and not harmed by the same event, varying with explanatory context. The intuition against this thought is explained as being the result of a violation of pragmatic maxims dictating relevance to our conversational presuppositions (in particular, to the presupposition fixing only a particular specification of contrasts as salient). Similarly, explanatory claims in general often have a non-relational surface form even though really they are relational. Such a relational property is not arbitrary. In particular, once (but only once) given a specification of contrasts, the truth of whether someone has been harmed is perfectly objective – or anyway as objective as the evaluations of the relevant counterfactuals.

4) SOME PROBLEMS RESOLVED

Comparative analyses of harm have been subject to several objections in the literature. Armed with the above approach, turn to these objections now. Answering them will usefully flesh out several details of this paper's proposal.

Suppose we define harm to be relativized to some counterfactual world. One problem is that there are of course many such counterfactual worlds, so which do we choose? To adapt an example from Feldman (1992), suppose a person lives a happy life in country X, but they would have led an even happier one in country Y. Does living in country X rather than Y therefore harm them, even if they are never aware of this foregone opportunity? The problem with simply answering 'yes' is that it seems almost everyone is harmed compared to some alternative world or other, even if those alternatives only be science-fiction worlds without ageing or disease. That in turn suggests that a comparative theory must register everyone as suffering harm, which is taken to be implausible. How can we restrict the admissible alternatives only to those that are 'realistic' (Nagel 1979)?

In reply, first, relative to some contrasts even residents of happy country X are indeed harmed. If those contrasts are salient then it is desirable that our theory registers this harm. Fundamentally, which counterfactual alternatives are 'realistic' is itself something that will vary with context; it should not be stipulated in advance.

Second, there is another issue here too. As others have pointed out, when considering counterfactuals concerning particular events, we should consider only, in Stalnaker-Lewis terminology, the nearest possible world. For example, what would have happened had I not died in an accident aged fifty? Absent a complete modal skepticism, it

is entirely legitimate to say that some answers are objectively more likely than others – for instance, to say that it is more likely I would have carried on living another ten years than it is that I would soon after have been kidnapped and killed painfully by aliens. Given this judgement, it is in turn entirely legitimate to say that my death in the accident was (probably) an opportunity cost for me. This is especially so given the fact, emphasized by McMahan (2002), that the relevant counterfactual is clearly not 'if I was immortal' but rather is the much more tractable token one 'if the particular death had not happened'. (Counterfactuals that concern death do raise other issues, considered below .)

The lesson from this is that, in order to be tractable, a counterfactual must in a sense be particular – its antecedent must refer to the absence of a particular event. This is a constraint on our theory of harm. But most leading candidates already satisfy this constraint, so in practice it makes little difference. And once it is satisfied, evaluating the requisite counterfactuals is then no more problematic than evaluating counterfactuals generally.

Hanser (2008) raises several interesting new objections to a counterfactual-comparative account of harm.[6] Consider these now. The first is:

> [C]omparative accounts look only at the difference … between the subject's actual (or present) level of well-being and the level he would have enjoyed … in some relevant alternative state of affairs. The subject's absolute level of well-being is immaterial. It doesn't matter how high up or down the scale the levels being compared lie; all that matters is the size of the gap separating them. But perhaps this is a mistake. Suppose that I must either cause a very well off person to undergo a moderate decline in well-being or cause a much less well off person to undergo a decline of equal magnitude. And suppose

> that although the well off person would remain quite well off after his decline, the less well off person would be pushed below an important threshold: he would come to be not just worse off than he was before, and worse off than he otherwise would have been, but *badly* off. It is surely plausible to say that given this choice, I should cause the well off person to undergo the decline. (431, italics in the original)

In other words, there is more to harm than size of gap, implying that mere comparison is insufficient – we need to take some account of absolute levels too. In reply, I agree that in Hanser's example the less well off person is plausibly the more harmed; but I disagree that this tells against a comparative account. In causal modeling terms, all depends on choice of effect variable; causation is not even well defined until that has been specified. In this example, what exactly is the effect variable?

There are two possibilities. First, the effect variable is a (therefore poor) measure of well-being that is insensitive to how well-being varies non-linearly with the measure's absolute value; perhaps monetary wealth would be an example. The second possibility is that, by contrast, the variable is so sensitive; perhaps some philosophical construct of well-being, such as level of positive affect, would be such a measure. In the first case, a comparison of losses of monetary wealth would indeed fail to track the asymmetry between the rich and poor man cases. But this just shows that such a choice of effect variable is misplaced here. (It's hardly news that monetary wealth is a poor measure of well-being.) Any definition of harm should clearly take as its focus instead some philosophically well grounded measure of well-being. Comparisons of the rich and poor man would then indeed show the poor man's loss, and hence the harm done him, to be greater.

The apparent paradox arises from the framing of the example. First, the rich and poor men are said to suffer well-being declines 'of equal magnitude'. Next, it is then stated that the poor man's loss of well-being is in fact of much greater magnitude. But, if understood as being with respect to the same entity (i.e. well-being), these two claims are incoherent. They only make sense if we are implicitly switching effect variables between the two sentences. The remedy therefore is to be explicit about our effect variable from the start.

Move on now to a second objection raised by Hanser, this time concerning the distinction between long-run and short-run effects:

> Think of the title character of the 1970s television show *The Six Million Dollar Man*, whose legs were shattered in an accident but who was then given 'bionic' replacement legs enabling him to run faster and jump higher than he ever could before. Although in the long run … the shattering of his legs came to him as a benefit, I think we should grant that it at first came to him as a harm. (424)

The problem for a counterfactual account here is taken to be that it cannot capture the short-term harm, because the bionic man's life is by assumption better with the accident than without.

But to causalist eyes, the solution is straightforward. The key point is that we are free to choose whichever effect variable best reflects our prior investigative interests, and the impact of any given cause will inevitably depend on this choice. Consider an analogy: suppose an economic policy increased output this year but lowered it next year (compared to if we had left policy unchanged). Depending on whether we care more

about output this year or next, without contradiction the very same policy can therefore be deemed either beneficial or harmful.[7] Similarly, we can be interested in either short-term or long-term well-being.[8] Either is a perfectly legitimate focus of interest. It follows that the bionic man's accident could be either harmful or beneficial, depending on whether the effect variable of interest is his well-being immediately after the accident or his well-being after the operation too.

Turn next to a third objection that Hanser raises, namely that counterfactual accounts are unable to encompass the harmfulness of death. Indeed, this is one of Hanser's main charges against all comparative approaches. He writes: 'I assume that when someone dies, he ceases to have any level of well-being. The state of *being dead* has no value for a person, whether positive, negative or neutral.' (2008, 437, his italics) It follows that we cannot compare the levels of well-being of being alive versus being dead, and therefore that a comparative account cannot declare death a harm, no matter how young and thriving the unfortunate victim. Hanser's own account, in contrast, focuses just on the single event of dying, which consists in the loss of certain basic goods and thus counts straightforwardly as a harm.

I think the best response here is to deny Hanser's starting assumption, and to assert instead that we can assign a value to being dead – namely, a neutral or zero value. This is indeed the consensus view among philosophers (e.g. Nagel 1979, Bradley 2004, Thomson 2011). As many have argued, it is supported by linguistic usage, for instance the phrase 'better off dead' would seem to imply that death is ascribed at least an ordinal value as better than a bad life. It is surely supported also by our practice. The obvious desire of the happy not to die, as well as the occasional welcoming of death by the

unhappy, each implies that the state of being dead has some middle value between life's extremes. I think we should follow the lead of our words and actions here.

After the next section I will turn to yet another objection, namely cases of pre-emptive harms.

5) SOME PROBLEMS FOR RIVAL VIEWS

There is a rival comparative approach to the counterfactual one this paper favors, namely a temporal-comparative one. This holds that you suffer harm just when you are in a worse (actual) state than you were previously. But there is a general category of counterexample to this view, namely cases of what Hanser calls preventative harms.[9] For example, if someone prevents a surgeon from operating to cure your blindness they have done you a harm, even though there is no change to your actual state of blindness. A counterfactual version of comparativism handles such cases straightforwardly – in the surgeon example, for instance, by appealing to the counterfactual of the improved state you would have been in had the surgeon operated.[10]

A causalist approach offers theoretical underpinning to this rejection of temporal comparativism. In particular, the rejection is for the same reason that, truistically, causation is associated with counterfactual dependence rather than mere correlation. A focus only on actual states of affairs leads to fatal difficulties with spurious correlations, post hoc propter hoc fallacies, and the like. Fundamentally, if what matters is what *caused* you to be in your current state, that requires comparison with what would otherwise have been the case rather than with your previous actual state.

Another common view of harm is not comparative at all, seeing it instead as consisting entirely in a bad state itself (e.g. Shiffrin 1999). It seems to me that this approach inevitably misses the connection between harm and change, concentrating as it does entirely on the noun rather than verb sense of harm. As a result, it is vulnerable to cases in which our judgements of harm are sensitive to the verb sense, and standard counterexamples to it play on exactly this aspect. For example, to borrow an example from Hanser, consider two people with dim vision: one's vision got that way through an unfortunate accident, whereas the second's got that way thanks to a surgeon's innovative intervention – prior to that they had been blind. The usual judgement is that only the first person has suffered a harm, even though both, of course, end up in the same actual state. The difference lies in how they got there.

This difference is naturally captured via contrasts, exactly as we should expect given that it is fundamentally a difference in causal histories. In the blindness example, if c* = the accident never happened (e.g. the doomed car had continued normally), e* = the subject has full vision, then c* entails e*, and e leaves the subject worse off than does e*. Accordingly, we correctly conclude that the accident harmed them. If, on the other hand, c* = the surgeon never intervened, then matters are different. For e* = the subject is blind, it follows that c* entails e*, but now e would be better than e* and so we would conclude, as desired, that the subject was benefited not harmed by the surgeon's intervention. Different salient c*, different harm judgement.

Turn next to Hanser's own, highly original event-based account. This 'holds that to undergo a harm (or benefit) is to be the subject of an event whose status as the undergoing of a harm (or benefit) derives from its being the sort of event that it is,

independently of the badness (or goodness) of any resulting state.' (Hanser 2008, 440) The central idea is that we can define harm not in terms of a bad state or comparison between bad states, but rather entirely in terms of a particular harming event.

However, from a causal point of view this proposal seems unpromising. Causation is usually taken to be a relation between events, not to be an event itself. How therefore, on Hanser's account, do we ever *cause* harm? For example, suppose you were blinded by a villain. Normally, this would be analyzed in terms of a cause event of the villain's assault and an effect event of your eye being damaged. The best interpretation of Hanser's scheme would seem to be that his focal harm event is what I have labeled the effect event, here the damaging of the eye. But the problem is that this seems to neglect the noun sense of harm. One result of that is the inability to distinguish between short-lived and long-lived harms. In our example, there would be no distinguishing between temporary and permanent blindness, for instance. Thomson (2011, 456-457) presses several objections along these lines. A possible response is to interpret a state of blindness to be one extended event, the degree of harm corresponding to its duration. Hanser, replying to Thomson, himself suggests this possibility (2011, 468). But then the harm is no longer defined as the momentary event of the blinding itself, and so incorporating the noun sense in this way seems only to come at the cost of losing the verb sense. In a way, such problems are not surprising: the causation literature rejects any attempt to, in effect, shoehorn the causal relation into just a single event in this way, and for good reason.

In many respects, the nearest precursor to this paper's approach is Alastair Norcross's contextualist theory (2005). It too is counterfactual-comparative and explicitly

relativizes harm to contrasts. It arrives at this view via a rather different route than we did, engaging neither with the causation literature nor with most of the cases discussed in this paper. It also offers no positive account of pre-emption cases (see below). I do endorse many of its criticisms of previous utilitarian approaches though, and its analysis of related examples.

Finally, Judith Jarvis Thomson offers her own twist on a counterfactual-comparative account. She defines harm as follows (2011, 448):

> Y harms A just in case A is in a state s such that:
>
> Y causes A to be in s, and for some state s*,
>
> (a) Y prevents A from being in s* by the same means by which Y causes A to be in s, and
>
> (b) A is worse off in a way for being in s than he would have been if he had been in s*.

A crucial maneuver is mentioned only in passing one page before this definition: 'I don't supply an analysis of [causation] … I leave [it] to intuition.' (2011, 447) This in effect immediately insulates a counterfactual approach from its most famous criticism, namely pre-emption cases. Pre-emption cases are problematic for counterfactual theories because our causal judgements in them, and hence derivative judgements of harm, deviate from the counterfactual dependence criterion (see below). Taking causation as a primitive neatly sidesteps the difficulty.

However, as usual in philosophy, going primitive brings with it costs as well as benefits. If the maneuver does too much theoretical work, then correspondingly less is

being done by the analysis. Here, I think the very motivation for counterfactualism in the first place is its connection to causation, and in turn causation's to harm. If we take causation as a primitive, why in addition explicitly appeal to counterfactuals at all? And if counterfactuals are relevant to our judgements of harm (as they surely are), making the connection between counterfactuals and causation serves to *explain* that relevance; taking causation as a primitive, on the other hand, leaves it a mystery.

Moreover, a counterfactual approach is enhanced in other ways too by explicit embedding in the causation literature. For example, the distinction between harm's effect term and a bad, discussed in section 1, is otherwise quite lost. Thomson's definition, meanwhile, in effect does incorporate contrasts on the effect side via its relativization to s*. I endorse this feature, supported as it is by contrastive considerations about causation generally. But these same considerations also suggest including contrasts on the cause side too – something missing from Thomson's definition. For instance, recycling a standard causation case, it seems wrong to say that Socrates sipping *rather than guzzling* hemlock caused him harm, whereas it seems right that Socrates sipping hemlock *rather than wine* did so.

One line of defense might be that Thomson's definition implicitly incorporates such contrast sensitivity in the cause slot already, via its causation clause. After all, part of her definition is 'if Y causes A', so if causation is understood contrastively, does that not already imply appropriate sensitivity to cause-contrasts? But if that is true, then why incorporate effect-contrasts explicitly either? By exactly the same argument, causation understood contrastively should make redundant the explicit mention of s*. It seems hard to justify a half-way house; contrasts should be mentioned either in both slots or in none.

Perhaps, moving away from Thomson's own view, one might take causation as primitive and then just define harm as causing a bad, with no further elaboration. This would again escape the problems raised by pre-emption cases. Nevertheless, in keeping with the remarks above, I think it is a theoretical improvement to analyze causation explicitly. There is a price for that though, namely that we thereby incur the obligation of dealing with pre-emption cases. I turn to those now.

6) PRE-EMPTION CASES

A baseball batter hits a pitch. The ball starts off travelling in the direction of the moon. If it reached the moon, it would harm two astronauts on its surface, perhaps by damaging their spaceship. A fielder catches the ball. If the ball had not been caught by the fielder, it would have fallen to the ground shortly afterwards. Claim: The fielder's catch prevented the ball from harming the astronauts. The usual reaction to this claim is to disagree strongly.[11]

A second example: Imagine you are staying at a hotel that gives out a coupon to all guests, entitling them to a free drink at the hotel bar. A freshly arriving guest has checked in and so is now entitled to a coupon. Upon leaving the hotel, you haven't used your coupon, so you return it to reception. The receptionist then hands this particular coupon to the new guest, who is on his way to the hotel bar. He uses this coupon to get himself a free drink. The receptionist has an ample supply of coupons: if you had not returned your coupon, the receptionist would have given the guest one of the other coupons. Unfortunately, the guest's drink harms him by causing him to fall ill. Here is the

question: did your returning your coupon to reception harm the new guest? The typical answer people give is 'no'.[12]

Finally, consider a third example:

> Suppose that a criminal wants to steal from S's store. Since the burglary will go more smoothly if S is not present, the criminal hires some thugs to break S's legs the day before the proposed crime. When the thugs arrive at S's home, however, they find that the local loan shark is already there breaking S's legs. (Hanser 2008, 434)

As Hanser says, people typically agree here that the loan shark is harming S.

What to make of these cases? Begin by noting that all three have exactly the same structure: a pre-empting harmer (or preventer); a pre-empted harmer (or preventer); and the harm itself. But despite this structural identity, our judgements *vary*. In particular, the pre-emptor is not judged to cause (or prevent) harm in the first two cases, but is so judged in the third. The lesson is that our judgement of pre-emptive harm is unstable, being strongly influenced by framing effects.[13] Several responses are possible. One is to define harm simply to track our various judgements here, and leave the variation between cases a mystery. (In effect, this is Thomson's approach.) Another is to engage explicitly with the psychology. In particular, perhaps we might formulate and test an error theory, which would serve to explain away some judgements as explicable cases of framing effects misleading us. In this way, for instance, a counterfactual theory might be insulated from the awkward (for it) judgements that undoubtedly arise in – some – pre-emption cases. (Northcott Manuscript discusses this in more detail, with a focus on causation itself rather than harm.)

Pre-emption cases have long proven a bugbear for counterfactual theories. Norcross (2005) discusses the shortcomings of Parfit's various attempts to handle them, for instance, while Feldman (1992) and McMahan (2002) discuss various elaborate scenarios of causal overdetermination more generally. The point now is not to claim that counterfactual theories are out of the woods – much further work would be required before we could claim that. It is merely to suggest that the implications of such cases are rather muddier than often claimed. In particular, I think it is premature to use them to write off a counterfactual account of harm, especially given that account's many other virtues. In any case, no one seriously doubts the close connection between causation and counterfactuals; pre-emption cases merely cast doubt on whether that connection is one of definition.

7) BENEFITS OF GOING CAUSAL

At the heart of this paper's approach is that judgements of harm are sensitive to contrasts in both the cause and effect slots. One advantage of being explicit about this is that it makes us realize the significance of just how context determines which contrasts are salient.[14] Much empirical work, for instance, has revealed the role of *norms* in causal selection (Hitchcock and Knobe 2009). What we deem a cause rather than mere background condition depends among other things on what is statistically rare rather than common (e.g. the dropped match rather than oxygen in the atmosphere), what is morally notable rather than neutral (the drunk driver rather than difficult weather conditions), or what is dysfunctional rather than functional (the short circuit rather than the electricity being turned on). Because of the connection between harm and causation, what counts as

harmful will therefore also be sensitive to these norms in the same way. And such causal – and therefore harm – selection criteria are naturally represented via selection of contrasts.

To finish, consider two last examples, both taken from the causation literature. Each further illustrates how sensitivity to contrasts is necessary to any analysis of harm. The first concerns some favorite plants of mine: I leave town for a couple of weeks and entrust their care to you. Alas, you neglect to water them, and so they die. It seems clear that you have thereby done me harm. Now consider – it is also true that the Queen of England did not water my plants. Has she also done me harm? Moreover, the local builder failed to build unannounced a special gutter from my roof that would have brought water to my plants. This failure too meant that my plants died; has the builder too done me harm? Presumably, the usual view would be that the Queen and the builder have not harmed me, even though my friend has.

Yet this distinction between the friend case and the Queen and builder ones is invisible to all previous accounts of harm. Thus, the plants (and so my feelings) end in the same actual state in each case, and moreover were also in identical previous states too. Therefore a non-comparative account is stymied, and so is a temporal-comparative one. Nor does Hanser's event-based theory have any apparent means for making the requisite distinction. A simple counterfactual-comparative account also struggles here, for in all cases there is exactly the same pattern of counterfactual dependence. Finally, Thomson's variant of the counterfactual view characteristically rests everything on taking 'cause' as a primitive. We do indeed deem my friend's neglect, but not the Queen's or builder's, the cause of my plants dying. As before, this means that Thomson's account

can accommodate the different harm intuitions, but only at the cost of giving up on any attempt to explain them.

A contrastive view, however, handles the example straightforwardly (Schaffer 2005). As the discussion above of norms would suggest, the salient contrast is that my friend did water the plants, not that the Queen or builder (or aliens or anyone else) did. Accordingly, a contrastive account correctly judges that only the friend here has harmed me.

The second example features positive rather than absence causation. Two assassins, Captain and Assistant, are on a mission to kill Victim (Hitchcock 2003; the example is originally due to Michael McDermott). Upon spotting Victim, Captain yells 'fire!', and Assistant fires. Overhearing the order, Victim ducks and survives unscathed. Did Captain's yelling 'fire!' benefit Victim by causing her to survive? The answer is unclear. On one hand, the yell alerted Victim and so indeed enabled her to survive since if left unalerted she would surely (it is stipulated) have died. On the other hand, if Captain had not yelled then Assistant would never have fired in the first place and so Victim never been endangered, so the yell can hardly be held to have *caused* Victim's survival. Captain's yell both initiates the threat to Victim (i.e. Assistant's shot) and also the mechanism protecting her (her overhearing and consequently ducking).

It turns out that this unclarity can be resolved by appeal to contrasts. Consider two, more detailed versions (Northcott 2008, 112-113):

> First version: a Captain is training an Assistant in assassination. Only the latter has a gun. They are stalking a Victim in a crowded market place when there occurs a great surge of people that threatens to carry Victim off to safety. Captain and Assistant become

separated and lose visual contact. Assistant will not shoot without authorization from Captain. Therefore, in order still to have any chance of killing Victim before she gets away, Captain as an emergency measure yells to Assistant to 'fire!' As a result, Assistant indeed fires. However, the noise of the yell is also heard by Victim who consequently ducks and as a result survives the shot before indeed escaping with the crowd. The question is: was Captain's yell the cause of Victim's survival? Intuitively, the answer seems to be 'no' since even if Captain had not yelled still Victim would have got away in any case. Thus Victim would likely be thanking the fortuitous surge of the crowd for enabling her to escape rather than thanking Captain for doing his best, under the circumstances, to *prevent* that escape.

Second version: suppose instead that there was no surge in the crowd and that in fact Captain and Assistant were standing at leisure on a balcony overlooking Victim, with plenty of time to select the moment to fire. Assume there is incentive not to let Victim get away without taking at least one shot at her. Captain could communicate to the eager Assistant at any moment by the prearranged signal of raising a finger. However, just as he is indeed about to raise his finger, Captain impulsively yells out loud 'fire!', which alerts Victim who consequently ducks and as a result survives the shot and escapes. Now again the question is: was Captain's yell the cause of Victim's survival? This time the answer seems to be 'yes', since this time the yell made all the difference. If only he had signaled silently as arranged, rather than bursting out into a yell, Captain could have ensured a successful assassination. That is, if only Captain had not yelled, the mission would not have failed and Victim would not have survived.

The crucial matter is the salient contrast to Captain's actual action of yelling 'fire!' If the contrast is that the Captain gave no signal at all, i.e. continued quietly as before, then the first version is instantiated, his yell did not cause Victim's survival and therefore did not benefit her. But if the salient contrast, as in the second version, is that Captain did give a signal, but a silent one, then now his yell did not benefit Victim. Once again, no other account of harm can explain the difference. To analyze harm successfully, we must go contrastive-counterfactual.

r.northcott@bbk.ac.uk

References

-- Ben Bradley, 'When is Death Bad for the One Who Dies?', *Nous* 38 (2004), pp. 1-28.
-- Fred Feldman, *Confrontations with the Reaper: A Philosophical Study of the Nature and Value of Death* (Oxford, 1992).
-- Matthew Hanser, 'The Metaphysics of Harm', *Philosophy and Phenomenological Research* 77 (2008), pp. 421-450.
-- Matthew Hanser, 'Still More on the Metaphysics of Harm', *Philosophy and Phenomenological Research* 82 (2011), pp. 459-469.
-- H.L.A. Hart and Tony Honoré, *Causation in the Law*, 2nd edn. (Oxford, 1985).
-- Christopher Hitchcock, 'The Role of Contrast in Causal and Explanatory Claims', *Synthese* 107 (1996), pp. 395-419.
-- Christopher Hitchcock, 'Of Humean Bondage', *British Journal for the Philosophy of Science* 54 (2003), pp. 1-25.
-- Christopher Hitchcock and Joshua Knobe, 'Cause and Norm', *Journal of Philosophy* 106 (2009), pp. 587-612.
-- David Lewis, 'Causation', *Journal of Philosophy* 70 (1973), pp. 556–567.
-- David Lewis, 'Causation as Influence', *Causation and Counterfactuals*, ed. John Collins, Ned Hall, and L.A. Paul (Cambridge Massachusetts, 2004), pp. 75-106.
-- Cei Maslen, 'Causes, Contrasts and the Nontransitivity of Causation', *Causation and Counterfactuals*, ed. John Collins, Ned Hall, and L.A. Paul (Cambridge Massachusetts, 2004), pp. 341-358.
-- Jeff McMahan, *The Ethics of Killing: Problems at the Margins of Life* (Oxford, 2002).
-- Thomas Nagel, 'Death', *Mortal Questions*, Thomas Nagel (Cambridge, 1979), pp. 1-10.
-- Alastair Norcross, 'Harming in Context', *Philosophical Studies* 123 (2005), pp. 149-173.
-- Robert Northcott, 'Causation and Contrast Classes', *Philosophical Studies* 139 (2008), pp. 111-123.
-- Robert Northcott, 'Partial Explanations in Social Science', *Oxford Handbook of Philosophy of Social Science*, ed. Harold Kincaid (Oxford, 2012), pp. 130-153.
-- Robert Northcott, Manuscript.
-- Judea Pearl, *Causality* (New York, 2000).
-- Jonathan Schaffer, 'Causes Need Not be Physically Connected to their Effects: The Case for Negative Causation', *Contemporary Debates in Philosophy of Science*, ed. Christopher Hitchcock (Oxford, 2004), pp. 197-216.
-- Jonathan Schaffer, 'Contrastive Causation', *Philosophical Review* 114 (2005), pp. 297-328.
-- Seana Shiffrin, 'Wrongful Life, Procreative Responsibility, and the Significance of Harm', *Legal Theory* 5 (1999), pp. 117-148.
-- Peter Spirtes, Clark Glymour and Richard Scheines, *Causation, Prediction, and Search*, 2nd edn. (Cambridge Massachusetts, 2000).
-- Judith Jarvis Thomson, 'More on the Metaphysics of Harm', *Philosophy and Phenomenological Research* 82 (2011), pp. 436-458.
-- Bas Van Fraassen, *The Scientific Image* (Oxford, 1980).
-- James Woodward, *Making Things Happen: A Theory of Causal Explanation* (Oxford, 2003).

[1] Complicating matters further, badness so understood is unnecessary for harm, because harm requires only a decrease in well-being and thus is consistent with well-being still remaining high in absolute terms.

Hanser gives the example of an injury that harms a Nobel prize winner by reducing their cognitive capacity from exceptional to merely very good.

[2] To repeat, on the contrastive view that this paper will endorse, strictly speaking what is caused is not just a level of well-being but instead this level *rather than* some counterfactual alternative. But causal talk often takes a surface binary form (something which a contrastive view is able to accommodate – Northcott 2008).

[3] Admittedly, it is unclear what if any role causation plays in some areas of fundamental physics. But: first, these doubts apply equally to any notion of causation, not just the difference-making one; and second, whatever our view of causation's ultimate metaphysical status, causation as difference-making is undoubtedly typical of everyday 'special-science' situations such as those that feature in discussions of harm (Woodward 2003).

[4] The phrase 'e leaves A in a … state' is intended to be shorthand for the fact that we can consistently say that causation is ultimately a relation between events, while it simultaneously being legitimate to talk of a person being caused to be in a particular state (see also Thomson 2011, 458).

[5] There is much detail to be added here about a contrastive definition of causation, such as how choice of c and e is constrained, what determines which c* and e* are salient, and technical wrinkles arising from the fact that c* and e* are in general *sets* of contrast events (Northcott 2008, Schaffer 2005, Norcross 2005, Maslen 2004, Van Fraassen 1980).

[6] In fact, as Hanser acknowledges, the first objection follows Shiffrin (1999), who presents a number of similar examples.

[7] Following the literature, I take 'beneficial' to be the opposite of harmful, i.e. to involve an increase in well-being.

[8] I am assuming here that it makes sense to understand well-being in a time-indexed way. For most currently popular measures, it does.

[9] Other kinds of counterexample have been given besides (e.g. Norcross 2005, 149-150).

[10] Hanser thinks our intuitions assign greater moral weight to cases of positive causation than to those of mere prevention, and then objects that a counterfactual approach treats the two cases symmetrically (2008, 428). In reply: first, it is questionable whether our intuitions really do follow that pattern always and everywhere (Schaffer 2004). But when they do, might this be explained away as the illicit seeping of type

considerations into intuitions about a token case? If, as a matter of statistical fact, most morally blameworthy harmings we come across are cases of positive causation rather than of prevention, then our intuitions might have become more responsive to the former than the latter. (Norcross (2005, 161) makes a similar point, albeit in a different context, when saying that our intuitions in such circumstances are 'the result of the all too common confusion of judgements of actions with judgements of character.' Thomson (2011, 440) also endorses this kind of explaining away, albeit again in a different context.) Hanser's own formal scheme, it is true, does treat the two cases differently. On the other hand, it offers no explanation for why this formal difference should imply a moral difference.

[11] Experiment confirms this for a purely causal version of the same scenario, in which people are asked whether the fielder's catch prevented the ball reaching the moon (Northcott Manuscript).

[12] Or at least, experiment shows that this is the typical answer to the analogous question framed in causal rather than harm terms. More particularly, subjects typically disagree with the claim that you returning your unused coupon enabled the guest to get a free drink (Northcott Manuscript).

[13] Notice though that in all three cases judgements of harm do track judgements of causation, again endorsing a causalist approach.

[14] Norcross (2005) notes that more than salience may be relevant.