Minds and Machines (2024) 34:34
https://doi.org/10.1007/s11023-024-09689-7

A Causal Analysis of Harm

Sander Beckers[1] · Hana Chockler[2] · Joseph Y. Halpern[3]

Received: 3 August 2023 / Accepted: 24 June 2024

Abstract

As autonomous systems rapidly become ubiquitous, there is a growing need for a legal and regulatory framework that addresses when and how such a system harms someone. There have been several attempts within the philosophy literature to define harm, but none of them has proven capable of dealing with the many examples that have been presented, leading some to suggest that the notion of harm should be abandoned and "replaced by more well-behaved notions". As harm is generally something that is caused, most of these definitions have involved causality at some level. Yet surprisingly, none of them makes use of causal models and the definitions of actual causality that they can express. In this paper, which is an expanded version of the conference paper Beckers et al. (Adv Neural Inform Process Syst 35:2365–2376, 2022), we formally define a qualitative notion of harm that uses causal models and is based on a well-known definition of actual causality. The key features of our definition are that it is based on *contrastive* causation and uses a default utility to which the utility of actual outcomes is compared. We show that our definition is able to handle the examples from the literature, and illustrate its importance for reasoning about situations involving autonomous systems.

Keywords Harm · Actual causation · Utility · Counterfactuals

✉ Sander Beckers
srekcebrednas@gmail.com

Hana Chockler
hana.chockler@kcl.ac.uk

Joseph Y. Halpern
halpern@cs.cornell.edu

[1] Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands

[2] Department of Informatics, King's College London, London, UK

[3] Department of Computer Science, Cornell University, Ithaca, USA

Published online: 21 July 2024

1 Introduction

The notion that one should not cause harm is a central tenet in many religions; it is enshrined in the medical profession's Hippocratic Oath, which states explicitly "I will do no harm or injustice to [my patients]" (National Library of Medicine, 2002) it is also a critical element in the law. Not surprisingly, there have been many attempts in the philosophy literature to define harm. Motivated by the observation that we speak of "causing harm", most of these have involved causality at some level. All these attempts have encountered difficulties. Indeed, Bradley (2012) says:

Unfortunately, when we look at attempts to explain the nature of harm, we find a mess. The most widely discussed account, the comparative account, faces counterexamples that seem fatal. But no alternative account has gained any currency. My diagnosis is that the notion of harm is a Frankensteinian jumble ...It should be replaced by other more well-behaved notions.

The situation has not improved much since Bradley's paper (see, e.g., recent accounts like (Carlson et al., 2021; Feit, 2019)). Yet the legal and regulatory aspects of harm are becoming particularly important now, as autonomous systems become increasingly more prevalent. In fact, the new proposal for Europe's AI act (European Commission, 2021) contains over 25 references to "harm" or "harmful", saying such things as "...it is appropriate to classify [AI systems] as high-risk if, in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons ..." [(European Commission, 2021), Proposal preamble, clause (32)]. Moreover, the European Commission recognized that if harm is to play such a crucial role, it must be defined carefully, saying "Stakeholders also highlighted that ...it is important to define ...'harm' [(European Commission, 2021), Part 2, Section 3.1]. Legislative bodies in the UK are also discussing the question of harm and who caused harm in the case of accidents involving autonomous vehicles. The Law Commission of England and Wales and the Scottish Law Commission are recommending that drivers of self-driving cars should not be legally responsible for crashes; rather, the onus should lie with the manufacturer (Law Commission, 2022). In particular, if there is harm then this is caused by the manufacturers. The manufacturers translate this recommendation to a standard according to which the driver does not even have to pay attention while at the wheel. If a complex situation arises on the road requiring the driver's attention, the car will notify the driver, giving them 10 seconds to take control. If the driver does not react in time, the car will flash emergency lights, slow down, and eventually stop (BBC News, 2022). Consider the following example (to which we return later).

Example 1 (Autonomous Car) An autonomous car detects an unexpected stationary car in front of it on a highway. It could alert the driver Bob, who would then have to react within 10 seconds. However, 10 seconds is too long: the car will crash into the stationary car within 8 seconds. The autonomous car's algorithm directs it to crash into the safety fence on the side of the highway, injuring Bob. Bob claims that he was harmed by the car. Moreover, he also claims that, if alerted, he would have been able to find a better solution that would not have resulted in his being injured (e.g., swerving into the incoming traffic then back to his own lane after passing the stationary car). We assume that if the autonomous car had done nothing and collided with the stationary car, both drivers would have been injured much more severely. □

While the causal model depicting this story is fairly straightforward, the decision on whether harm was caused to Bob, and if yes, who or what caused the harm, is far less clear. Indeed, the philosophy literature seems to suggest that trying to determine this systematically is a lost cause. But as this example illustrates, the stakes of having a well-defined notion of harm have become much higher with the advent of automated decision-making. In contrast to human agents, such systems do not have an informal understanding of harm that informs their actions; so we need a formal definition. Situations like that described in Example 1 are bound to arise frequently in the interaction of autonomous systems with human users, in a variety of domains. We briefly outline two of those.

Imagine a UAV used by the military has to decide whether or not it should bomb a suspected enemy encampment. The problem is that the target is not clearly identified, because there are two camps close to each other: one consisting of civilian refugees, another consisting of a rebel group that is about to launch a deadly attack on the refugee camp, killing all of its inhabitants. The UAV's decision is based only on the expected utility of the refugees, and therefore it bombs the camp. Tragically, as it turns out, the camp was that of the refugees. Here we have the intuition that the UAV harmed these refugees, despite the fact that both actions would have led to all the refugees being killed. Examples in which one event (the bombing) preempts another event (the attack) from causing an outcome are known as *Late Preemption* examples in the causality literature; we discuss them later in the paper.

In the healthcare domain, autonomous systems are used for, among other things, classifying MRI brain images suspected of containing a tumor. If an image is classified as having a tumor, the system decides whether to recommend a surgery. While the overall accuracy of the system is superior to that of humans, in some instances the system overlooks an operable tumor. Imagine a patient who has such a tumor and dies from brain cancer as the result of not undergoing surgery, leading to a dispute between the patient's family and the hospital regarding whether the patient was harmed. Even if both parties agree that the patient would probably have been alive if the diagnosis had been performed by a human, the hospital might claim that using the system is the optimal policy, and therefore one should compare the actual outcome only to those that could have occurred under the policy.

Fortunately, the formal tools at our disposal to develop a formal notion of harm have also improved over the past few years; we take full advantage of these developments in this paper, which is an extension of the conference paper (Beckers et al., 2022). Concretely, we provide a formal definition of harm that we believe deals with all the concerns that have been raised, seems to match our intuitions well, and connects closely to work on decision theory and utility. Here we briefly give a high-level overview of the key features of our approach and how they deal with the problems raised in the earlier papers.

There is one set of problems that arise from using counterfactuals that also arise with causality, and can be dealt with using the by-now standard approaches in defining causality. In fact, Carlson et al. (2021) (CJR from now on) raise a number of problems with defining harm causally that can be solved by simply applying the definition of actual causality given by Halpern (2015, 2016). For example, the issue of whether failing to take an action can be viewed as causing harm (e.g., can failing to water a neighbor's plants after promising to do so be viewed as causing harm) can also be dealt with by using the standard definition of causality (which allows lack of an action to be a cause).

We remark that Richens et al. (2022) (RBT from now on) also recently observed that using causality appropriately could deal with some of the problems raised in the harm literature. They also give a formal definition of harm that uses causal models, but it does not make use of a sophisticated definition of actual causality such as the one given by Halpern (2015, 2016). (See Sect. 6 for a comparison of our approach to theirs and more discussion of this issue.) RBT focus on the more quantitative, probabilistic aspects of harm. We also believe that a quantitative account is extremely important; we offer such an account in Beckers et al. (2023). Conceptually though, the qualitative notion comes first: only after establishing whether or not there was harm does it make sense to ask how much harm occurred. Indeed, our quantitative account generalizes the qualitative account we develop here in several ways (see Sect. 7).

In any case, just applying the definition of causality does not deal with all problems. The other key step that we take is to assume that there exists a *default* utility. Roughly speaking, we define an event to cause harm whenever it causes the utility of the outcome to be lower than the default utility. The default may be context-dependent, and there may be disagreement about what the default should be. We view that as a feature of our definition. For example, we can capture the fact that people disagree about whether a doctor euthanizing a patient in great pain causes harm by taking it to be a disagreement about what the appropriate default should be. Likewise, the dispute between the family and the hospital described above can be modeled as a disagreement about the right default. Moreover, by explicitly bringing utility into the picture, we can connect issues that that have been discussed at length regarding utility (e.g., what the appropriate discount factor to apply to the utility of future generations is) to issues of harm.

Lastly, we should point out that despite the progress that has been made in defining actual causality, it is not yet a fully solved problem. In particular, there exist examples which share the same causal structure (they are "structurally isomorphic") and yet our intuitive verdicts about causation differ between them. This is a well-known issue within the causation literature and a common approach to handling it is to invoke additional normative elements into the analysis of causation (see, e.g., Halpern and Hitchcock (2015)).

For example, imagine two soldiers who are awaiting the order of an officer regarding whether or not to shoot a potential enemy. However, one soldier is in fact an enemy spy and will do exactly the opposite of the order. So no matter what the order, one of the two soldiers will shoot and kill the victim. Thus, the binary choice available to the officer functions as a switch between two causal paths that both lead to the victim's death. If the officer orders the soldiers to shoot, the verdict that the officer caused the victim's death sounds unproblematic. But if the officer refrains from giving the order, this verdict sounds far less plausible. (RBT discuss a more complicated variant of this example in their Appendix E to illustrate how this issue affects our approach.) One could appeal to the normative asymmetry between giving an order to shoot and not doing so to separate the two cases, but as this is an issue that is not specific to the use of causation for an analysis of harm, we here set it aside.

2 Causal Models and Actual Causality

We start with a review of causal models (Halpern & Pearl, 2005), since they play a critical role in our definition of harm. The material in this section is largely taken from Halpern (2016). We assume that the world is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The structural equations describe how these values are determined.

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). For simplicity, we assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X (i.e., $F_X = \mathcal{F}(X)$) such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. The structural equations define what happens in the presence of external interventions. Setting the value of some set $\mathbf{X}$ of variables to $\mathbf{x}$ in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{\mathbf{X} \leftarrow \mathbf{x}}$, which is identical to M, except that the equations for $\mathbf{X}$ in $\mathcal{F}$ are replaced by $\mathbf{X} = \mathbf{x}$.

Note that the causal models we consider here are deterministic. In general, one can also consider *probabilistic causal models*. A probabilistic causal model is a tuple $M = (\mathcal{S}, \mathcal{F}, \Pr)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and Pr is a probability on contexts. Deterministic models suffice for offering a qualitative notion of harm, but we use probabilistic causal models for our quantitative generalization (Beckers et al., 2023).

The dependencies between variables in a causal model $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ can be described using a *causal network* (or *causal graph*), whose nodes are labeled by the endogenous and exogenous variables in M, with one node for each variable in $\mathcal{U} \cup \mathcal{V}$. The roots of the graph are (labeled by) the exogenous variables. There is a directed edge from variable X to Y if Y *depends on* X; this is the case if there is some setting of all the variables in $\mathcal{U} \cup \mathcal{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\mathbf{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \mathbf{z}) \neq F_Y(x', \mathbf{z})$. A causal model M is *recursive* (or *acyclic*) if its causal graph is acyclic. It should be clear that if M is an acyclic causal model, then given a *context*, that is, a setting $\mathbf{u}$ for the exogenous variables in $\mathcal{U}$, the values of all the other variables are determined (i.e., there is a unique solution to all the equations). We can determine these values by starting at the top of the graph and working our way down. In this paper, following the literature, we restrict to recursive models.

We call a pair $(M, \mathbf{u})$ consisting of a causal model M and a context $\mathbf{u}$ a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \mathbf{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \mathbf{u})$. The $\models$ relation is defined inductively. $(M, \mathbf{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\mathbf{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\mathbf{u}$). Finally, $(M, \mathbf{u}) \models [\mathbf{Y} \leftarrow \mathbf{y}]\varphi$ if $(M_{\mathbf{Y} \leftarrow \mathbf{y}}, \mathbf{u}) \models \varphi$.

A standard use of causal models is to define *actual causation*: that is, what it means for some particular event that occurred to cause another particular event. There have been a number of definitions of actual causation given for acyclic models (e.g., Beckers (2021); Glymour and Wimberly (2007); Hall (2007); Halpern and Pearl (2005); Halpern (2016); Hitchcock (2001, 2007); Weslake (2015); Woodward (2003)). Although most of what we say in the remainder of the paper applies without change to other definitions of actual causality in causal models, for definiteness, we focus here on what has been called the *modified* Halpern-Pearl definition (Halpern, 2015, 2016), which we briefly review. (See (Halpern, 2016) for more intuition and motivation.)

The events that can be causes are arbitrary conjunctions of primitive events (formulas of the form $X = x$); the events that can be caused are arbitrary Boolean combinations of primitive events. To relate the definition of causality to the (contrastive) definition of harm, we give a contrastive variant of the definition of actual causality; rather than defining what it means for $\mathbf{X} = \mathbf{x}$ to be an (actual) cause of ϕ, we define what it means for $\mathbf{X} = \mathbf{x}$ *rather than* $\mathbf{X} = \mathbf{x}'$ to be a cause of ϕ rather than ϕ'.

Definition 1 $\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}'$ is an *actual cause* of ϕ rather than ϕ' in $(M, \mathbf{u})$ if the following three conditions hold:

AC1. $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge \phi$.

AC2. There is a set $\mathbf{W}$ of variables in $\mathcal{V}$ and a setting $\mathbf{w}$ of the variables in $\mathbf{W}$ such that $(M, \mathbf{u}) \models \mathbf{W} = \mathbf{w}$ and $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}]\phi'$, where $\phi' \Rightarrow \neg\phi$ is valid.

AC3. $\mathbf{X}$ is minimal; there is no strict subset $\mathbf{X}''$ of $\mathbf{X}$ such that there exist values for which the above conditions are satisfied.

AC1 just says that $\mathbf{X} = \mathbf{x}$ cannot be considered a cause of ϕ unless both $\mathbf{X} = \mathbf{x}$ and ϕ actually happen. AC3 is a minimality condition, which says that a cause has no irrelevant conjuncts. AC2 captures the standard but-for condition ($\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}'$ is a cause of ϕ if, had $\mathbf{X}$ beem $\mathbf{x}'$ rather than $\mathbf{x}$, ϕ would not have happened) but allows us to apply it while keeping fixed some variables to the value that they had in the actual setting $(M, \mathbf{u})$. If $\mathbf{X} = \mathbf{x}$ is an actual cause of ϕ and there are two or more conjuncts in $\mathbf{X} = \mathbf{x}$, one of which is $X = x$, then $X = x$ is *part of a cause* of ϕ. In the special case that $\mathbf{W} = \emptyset$, we get the standard but-for definition of causality: if $\mathbf{X} = \mathbf{x}$ had not occurred (because $\mathbf{X}$ was $\mathbf{x}'$ instead) ϕ would not have occurred (because it would have been ϕ').

The reader can easily verify that $\mathbf{X} = \mathbf{x}$ is an actual cause of ϕ according to the standard non-contrastive definition (Halpern, 2016) iff there exist $\mathbf{x}'$ and ϕ' such that $\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}'$ is an actual cause of ϕ rather than ϕ' according to our contrastive definition.

3 Defining Harm

Many definitions of harm have been considered in the literature. The ones most relevant to us are those involving causality and counterfactuals, which have been split into two groups, called the *causal account of harm* and the *counterfactual comparative account account of harm*. CJR discuss many variants of the causal account; they all have the following structure:

An event e harms an agent **ag** *if and only if there is a state of affairs s such that (i) e causes s to obtain, and (ii) s is a harm for* **ag**.

The definitions differ in how they interpret the second clause. We note that although these definitions use the word "cause", it is never defined formally. "Harm" is also not always defined, although in some cases the second clause is replaced by phrases that are intended to be easier to interpret. For example, what (Suits, 2001) calls the *causal-intrinsic badness account* takes *s* to be a harm for **ag** if *s* is "intrinsically bad" for **ag**.

The causal-counterfactual account (see, e.g., Gardner (2015); Northcott (2015); Thomson (2011); Gardner (2015)) has the same structure; the first clause is the same, but now the second clause is replaced by a phrase involving counterfactuals. In its simplest version, this can be formulated as follows: *s* is a harm for **ag** if and only if **ag** would have been better off had *s* not obtained.

Even closer to our account is what has been called the *contrastive causal-counterfactual account*. For example, Bontly (2016) proposed the following[1]:

[1] Although our terminology is, by design, consistent with that of CJR, it is somewhat misleading. Specifically, the terminology suggests that the contrastive causal-counterfactual account is an instance of the causal-counterfactual account. However, it is not; it (or, at least, Bontly's version of it) is a different type of account (with some similarities).

An event e harms a person **ag** *if and only if there is a state of affairs s and a contrast state of affairs s′ such that (i) e rather than a contrast event e′ causes s rather than s′ to obtain, and (ii)* **ag** *is worse off in s than in s′.*

Our formal definition of harm is quite close to Bontly's. We replace "state of affairs" by "outcomes", and associate with each outcome a utility. This is essentially the standard model in decision theory, where actions map states to outcomes, which have associated utilities. Besides allowing us to connect our view to the standard decision-theoretic view (see, e.g., Resnik (1987); Savage (1954)), this choice means that we can benefit from all the work done on utility by decision theorists.

To define harm formally in our framework, we need to both extend and specialize causal models: We specialize causal models by assuming that they include a special endogenous variable O for *outcome*. The various values of the outcome value will be assigned a utility. We often think of an action as affecting many variables, whose values together constitute the outcome. The decision to "package up" all these variables into a single variable O here is deliberate; we do not want to consider the causal impact of some variables that make up the outcome on other variables that make up the outcome (and so do not want to allow interventions on individual variables that make up an outcome; we allow only interventions on complete outcomes). On the other hand, we extend causal models by assigning a utility value to outcomes (i.e., on values of the outcome variable), and by having a default utility.

We thus take a *causal utility model* to be one of the form $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathbf{u}, d)$, where $(\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ is a causal model one of whose endogenous variables is O, $\mathbf{u} : \mathcal{R}(O) \rightarrow [0, 1]$ is a utility function on outcomes (for simplicity, we assume that utilities are normalized so that the best utility is 1 and the worst utility is 0), and $d \in [0, 1]$ is a default utility.[2] As before, we call a pair $(M, \mathbf{u})$, where now M is a causal utility model and $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, a setting.

Just like causality, we define harm relative to a setting. Whether or not an event $\mathbf{X} = \mathbf{x}$ harms an agent in a given setting will depend very much on the choice of utility function and default value. Thus, to justify a particular ascription of harm, we will have to justify both these choices. In the examples we consider, we typically view the utility function to be **ag**'s utility function, but we are not committed to this choice (e.g., when deciding whether harm is caused by a parent not giving a child ice cream, we may use the parent's definition of utility, rather than the child's one). The choice of a default value is more complicated, and will be discussed when we get to examples; for the definition itself, we assume that we are just given the model, including utility function and default value.

The second clause of our definition is a formalization of Bontly's definition, using the definition of causality given in Sect. 2, where the events for us, as in standard causal models, have the form $\mathbf{X} = \mathbf{x}$ and the alternative events have the form $\mathbf{X} = \mathbf{x}'$,

[2] As we said in the introduction, in general, we think of the default utility as being context-dependent, so we really want a function from contexts to default utilities. However, in all the examples we consider in this paper, a single default utility suffices, so for ease of exposition, we make this simplification here.

and they cause outcomes $O = o$ and $O = o'$, respectively. Unlike Bontly's definition (and others), not only do we require that **ag** is worse off in outcome o (the analogue of state of affairs s) than in outcome o' (where "worse off" is formalized by taking the utility to be lower), we also require the utility of o to be lower than the *default utility*. There is also an issue as to whether we consider there to be harm if $\mathbf{X} = \mathbf{x}'$ results in a worse outcome than o. Since intuitions may differ here, we formalize this requirement in a third clause, H3, and use it to distinguish between harm and *strict* harm. We will see the effects of our modifications to Bontly's definition when we consider examples in Sect. 4.

Definition 2 $\mathbf{X} = \mathbf{x}$ *harms* **ag** in $(M, \mathbf{u})$, where $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathbf{u}, d)$, if there exist $o \in \mathcal{R}(O)$ and $\mathbf{x}' \in \mathcal{R}(\mathbf{X})$ such that

H1. $\mathbf{u}(O = o) < d$; and
H2. there exists $o' \in \mathcal{R}(O)$ such that $\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}'$ causes $O = o$ rather than $O = o'$ and $\mathbf{u}(O = o) < \mathbf{u}(O = o')$.

$\mathbf{X} = \mathbf{x}$ *strictly harms* **ag** in $(M, \mathbf{u})$ if, in addition,

H3. $\mathbf{u}(O = o) \leq \mathbf{u}(O = o'')$ for the unique $o'' \in \mathcal{R}(O)$ such that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}'](O = o'')$.

In the special case where Definition 2 is satisfied for some value o' appearing in H2 such that $\mathbf{u}(O = o) < d \leq \mathbf{u}(O = o')$, we say that $\mathbf{X} = \mathbf{x}$ *causes* **ag***'s utility to be lower than the default.*

It is important to point out that it is quite rare for harm and strict harm to come apart. For one, it requires causation and but-for causation to come apart (otherwise $o' = o''$). In addition, it requires O to have at least three values (otherwise again $o' = o''$). Lastly, even if these two conditions are met, we also need that $\mathbf{u}(O = o'') < \mathbf{u}(O = o) < \mathbf{u}(O = o')$. We discuss one example in Sect. 5 in which these conditions are met.

As with most concepts in actual causality, deciding whether harm occurred is intractable. Indeed, it is easy to see that it is at least as hard as causality, which is DP-complete (Halpern, 2015). However, we believe that, for many applications, this will not be a problem. For example, we expect that many of the policies that a policymaker is trying to evaluate with regard to harm can be described with relatively few variables. And in cases where there are many variables, the policymaker may want to *abstract* the model, coming up with a description that involves relatively few variables, since this is much easier to think about, and evaluate harm in this high-level model. (See (Beckers et al., 2019; Beckers & Halpern, 2019) for a formal treatment of abstraction.) To the extent that the relevant causal model can be described using only a few variables, we can decide harm by simply checking all possibilities.

It is useful to compare our definition with the counterfactual comparative account of harm. Here it is, translated into our notation:

Definition 3 $\mathbf{X} = \mathbf{x}$ *counterfactually harms* $\mathbf{ag}$ in $(M, \mathbf{u})$, where $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathbf{u}, d)$ if there exist $o, o' \in \mathcal{R}(O)$ and $\mathbf{x}' \in \mathcal{R}(\mathbf{X})$ such that

C1. $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x} \wedge O = o$;
C2. $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}'](O = o')$;
C3. $\mathbf{u}(O = o) < \mathbf{u}(O = o')$.

That is, $\mathbf{X} = \mathbf{x}$ counterfactually harms $\mathbf{ag}$ if, for some x' and o', $\mathbf{X} = \mathbf{x}$ is what actually happens (C1), $O = o'$ would have happened had $\mathbf{X}$ been set to $\mathbf{x}'$ (C2), and $\mathbf{ag}$ gets higher utility from o' than from o (C3). C1 and C2 together are equivalent to AC1 and AC2 in the special case that $\mathbf{W} = \emptyset$. That is, C1 and C2 essentially amount to but-for causality. C3 differs from our conditions by not taking into account the default value.

Note that Definition 3 has no analogue of AC3, but all the examples focus on cases where $\mathbf{X}$ is actually a singleton, so AC3 is trivially satisfied. The key point from our perspective is that the counterfactual comparative account considers only but-for causality, and does not consider a default value. The examples in the next section show how critical these distinctions are.

As mentioned earlier, RBT recently developed a formal account of harm using causal models. While their account is probabilistic and quantitative, we can consider the special case where everything is deterministic and qualitative. When we do this, their account reduces to a strengthening of Definition 3 that brings it somewhat closer to our account: they also suggest using defaults, but have default actions rather than default utilities. In their version of Definition 3, $\mathbf{X}$ is taken to be the variable representing the action(s) performed and x' is the default action.

In order to deal with the limitations of but-for causality, RBT offer a more general account (see their Appendix A) that uses path-specific causality, instead of actual causation. This makes their account different from ours in some significant respects; see Sect. 6.

4 Examples

We now analyse several examples to illustrate how our definition handles the most prominent issues that have been raised in the literature on harm. Bradley (Bradley, 2012, p. 398) identifies two such issues that strike him "as very serious", namely the problem of preemption, and the problem of distinguishing harm from merely failing to benefit. These problems therefore serve as a good starting point.

4.1 Preemption

To anyone familiar with the literature on actual causation what follows will not come as a surprise. Lewis used examples of preemption to argue that there can be causation without counterfactual dependence (i.e., we need to go beyond but-for causality); this conclusion is now universally accepted. Essentially the same examples show up in the literature on harm: cases of preemption show that an event can cause harm even though the agent's well-being does not counterfactually depend on it. Thus, the counterfactual comparative account of harm fails for the same reason it failed for causality. The good news is that the formal definition of causation (by design) handles problems like preemption well; moreover, the solution carries over directly to our definition of harm. The following vignette is due to Bradley (2012), but issues of preemption show up in many papers on causality (Beckers, 2021; Hall, 2007; Halpern & Pearl, 2005; Halpern, 2016; Hitchcock, 2007; Weslake, 2015); all can be dealt with essentially the same way.

Example 2 (Late preemption) Suppose Batman drops dead of a heart attack. A millisecond after his death, his body is hit by a flaming cannonball. The cannonball would have killed Batman if he had still been alive. So the counterfactual account entails that the heart attack was not harmful to Batman. It didn't make things go worse for him. But intuitively, the heart attack was harmful. The fact that he would have been harmed by the flaming cannonball anyway does not seem relevant to whether the heart attack was actually harmful.

In terms of the formal definition, we take H to represent whether Batman has a heart attack ($H = 0$ if he doesn't; $H = 1$ if he does), C to represent if Batman is hit by a cannonball, and D to represent whether Batman dies. Let $\mathbf{u}$ be the context where $H = 1$. Even without describing the equations, according to the story, $(M, \mathbf{u}) \models H = 1 \wedge D = 1 \wedge [H = 0](D = 1)$: Batman has a heart attack and he dies, but he would have died even if he did not have a heart attack (since he would have been hit by the cannon ball). Thus, C3 does not hold, since $o = o'$; the outcome is the same whether or not Batman has a heart attack.

The standard causal account handles this problem by introducing two new variables: K, for "Batman is killed by the cannonball", and S, for "Batman died of a heart attack", to take into account the temporal asymmetry between death due to a heart attack and death due to a cannonball. (We could also deal with this asymmetry by having "time-stamped" variables that talk about when Batman is alive. For more details on incorporating temporal information by using time-stamped variables, see (Halpern, 2016).) The causal model has the following equations: $D = S \vee K$ (i.e., $D = 1$ if either $S = 1$ or $K = 1$: Batman dies if he has a heart attack or the canonball kills him); $S = H$ (Batman's heart stops if he has a heart attack); and $K = \neg S \wedge C$ (Batman is killed by the canonball if the canonball hits him and his heart is still beating). We now get that Batman's heart attack rather than its absence is a cause of him being alive rather than dead. Clearly $(M, \mathbf{u}) \models H = 1 \wedge D = 1$. If we fix $K = 0$ (its actual value, since the cannonball in fact does not kill Batman; he is already dead by the time the cannonball hits him), then we have that $(M, \mathbf{u}) \models [H = 0, K = 0](D = 0)$, so AC2 holds. Thus, the causal part of H2 holds. (See [(Halpern, 2016), Example 2.3.3] for a detailed discussion of an isomorphic example.)

If we further assume, quite reasonably, that Batman prefers being alive to being dead (so the utility of being alive is higher than that of being dead) and that the default utility is that of him being alive, then H1 and H2 hold. Thus, our definition of harm avoids the counterintuitive conclusion by observing that Batman's heart attack caused his death, thereby causing the utility to be lower than the default. □

Our analysis of preemption is indicative of the more general point that many of the issues plaguing the literature on harm can be resolved by making use of causal models and the definitions of causation that they allow. Causal models allow a more precise and explicit representation of the relevant causal structure, thereby forcing a modeler to make modeling choices that resolve the inherent ambiguity that comes with an informal and underspecified causal scenario. Obviously such modeling choices can be the subject of debate (see (Halpern & Hitchcock, 2010) for a discussion of these modeling choices). The point is not that using causal models by itself determines a unique verdict on whether harm has occurred, but rather that such a debate *cannot even be had* without being explicit about the underlying causal structure.

4.2 Failing to Benefit

One of the central challenges in defining harm is to distinguish it from merely failing to benefit. Although most authors define benefit simply as the symmetric counterpart to harm, we do not believe that this is always appropriate; we return to this issue in Beckers et al. (2023) where we consider more quantitative notions of harm. But for the current discussion, we can set this issue aside: what matters is that merely failing to make someone better off does not in itself suffice to say that there was harm. CJR present the following well-known scenario to illustrate the point.

Example 3 (Golf clubs) Batman contemplates giving a set of golf clubs to Robin, but eventually decides to keep them. If he had not decided to keep them, he would have given the clubs to Robin, which would have made Robin better off.

By keeping the golf clubs, Batman clearly failed to make Robin better off. The counterfactual account considers any such failure to result in harm. Indeed, it is easy to see that C1–C3 hold. If we take GGC to represent whether Batman gives the golf clubs to Robin ($GGC = 1$ if he does; $GGC = 0$ if he doesn't) and the outcome O to represent whether Robin gets the golf clubs ($O = 1$ if he does; $O = 0$ if he doesn't), then $GGC = 0$ is a but-for cause of $GGC = 0$, so C1 and C2 hold. If we further assume that Robin's utility of getting the golf clubs is higher than his utility of not getting them, then C3 holds. Yet it sounds counterintuitive to claim that Batman harmed Robin on this occasion. □

Although H2 holds in our account of harm (for the same reason that C1–C3 hold), we avoid the counterintuitive conclusion by assuming that the default utility is $\mathbf{u}(O = 0)$, so H1 does not hold. This seems to us reasonable; there is nothing in the story that suggests that Robin is entitled to expect golf clubs. On the other hand, if we learn that Batman is a professional golfer, Robin has been his reliable caddy for many years, and that at the start of every past season Batman has purchased a set of golf clubs for Robin, then it sounds quite plausible that the default is for Robin to receive a set of golf clubs. With this default, H1 does hold, and our definition concludes that Robin *has* been harmed. Thus our account can offer different verdicts depending on the choice of default utility. As we said in the introduction, we view this flexibility as a feature of our account. This point is highlighted in the following, arguably more realistic, scenario. (RBT make exactly the same point as we do when they analyze such examples [(Richens et al., 2022), p. 15].)

Example 4 (Tip) Batman contemplates giving a tip to his waiter, but eventually decides to keep the extra money for himself. If he had not decided to keep it, he would have given it to the waiter, which would have made the waiter better off.

To those living in the US, it does not at all sound counterintuitive to claim that Batman harmed the waiter, for his income substantially depends on receiving tips and he almost always does receive a tip. Indeed, if we take the default utility to be that of receiving a tip, then in this example, the waiter is harmed by Batman not giving a tip. By way of contrast, in countries in Europe where a tip would not be expected, it seems to us reasonable to take the default utility to be that of not receiving a tip. In this case, the waiter would not be harmed. □

Examples 3 and 4 are isomorphic as far as the causal structure goes; we can take the utilities to be the same as well. This means that we need additional structure to be able to claim that the agent is harmed in one case and not the other. That additional structure in our framework, which we would argue is quite natural, is the choice of default utility. Note that neither scenario explicitly mentions what the default utility should be. We thus need to rely on further background information to make a case for a particular choice. There can be many factors that go into determining a good default. We therefore do not give a general recipe for doing so. Indeed, as we pointed out in the introduction with the euthanasia example, reasonable people can disagree about the appropriate default (and thus reach different conclusions regarding harm).

4.3 Preventing Worse

There exist situations in which the actual event rather than an alternative event causes a bad outcome rather than a good outcome, but the alternative results in an even worse outcome. Because of the latter, we do not consider these situations to be cases of strict harm, due to condition H3 in Definition 2. From the perspective of the car manufacturer, this is precisely what is going on in our starting Example 1, but Bob might disagree. We now take a closer look at this example to bring out the conflicting perspectives.

Example 5 (Autonomous car) Let O be a three-valued variable capturing the outcome for Bob, with the utility defined as equal to the value of O. $O = 0.5$ stands for the injury resulting from crashing into the safety fence, and a potentially more severe injury resulting from crashing into the stationary car is captured by $O = 0$. Bob not being injured is $O = 1$.

Recall that the system has the built-in standard that the driver's reaction time is 10 seconds, which is too long to avoid colliding into the stationary car. Imagine the

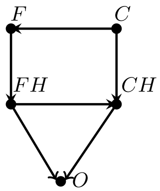

Fig. 1 Causal graph for Ex. 5

manufacturer implemented this standard by restricting the system's actions in such cases to two possibilities: do not intervene ($F = 0$) or drive into the fence ($F = 1$). This means that the causal structure (see Fig. 1) is very similar to our Late Preemption example (Example 2), for hitting the fence preempts the collision with the stationary car. We therefore add a variable to capture the asymmetry between hitting the fence and hitting the stationary car: *FH* and *CH* respectively. The equation for O is then such that $O = 1$ if $FH = CH = 0$, $O = 0.5$ if $FH = 1$, and $O = 0$ if $CH = 1$ and $FH = 0$.

As the autonomous car drives towards the fence only because there is a stationary car, the equation for F is $F = C$ (where C represents the presence of the car). The fact that hitting the fence prevents hitting the car is captured in the equation for *CH*: $CH = C \wedge \neg FH$. Lastly, we have $FH = F$. The context is such that $F = 1$ and $C = 1$, and thus $FH = 1, O = 0.5$, and $CH = 0$.

Did the system harm Bob? CJR answer this in the negative for an example that is essentially the same as this one (see their "Many Threats" example), and use this verdict to argue against the causal-counterfactual account, which reaches the opposite verdict. They also claim that Bontly's contrastive causal-counterfactual account reaches the correct verdict because there is no contrastive causation here. As they do not give a definition of causation, it is impossible to reconstruct how they arrive at this verdict. We disagree with CJR: we believe that there is contrastive causation here. Indeed, the car's hitting the fence rather than not hitting it is a cause of Bob being mildly injured rather than not being injured at all. To see why, observe that taking $\mathbf{W}$ to be *CH*, we get that $F = 1$ rather than $F = 0$ causes $O = 0.5$ rather than $O = 1$: $(M, \mathbf{u}) \models [F \leftarrow 0, CH \leftarrow 0]O = 1$.

Therefore, if we assume that the default utility is that of Bob not being injured, conditions H1 and H2 are satisfied and there is harm. Notice though that $F = 1$ rather than $F = 0$ is also a but-for cause of $O = 0.5$ rather than $O = 0$, that is, Bob's being mildly injured rather than severely injured counterfactually depends on the system's action. This is where condition H3 kicks in: it ensures that we do not consider there to be strict harm caused if the alternative would have resulted in an even worse outcome.

Thus, the car manufacturer could make the case that although their policy harmed Bob, it was justified in doing so.

More generally, it is an easy consequence of our definitions that in cases where $\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}'$ causes harm but not strict harm, the alternative event *would* have resulted in strict harm, i.e., $\mathbf{X} = \mathbf{x}'$ rather than $\mathbf{X} = \mathbf{x}$ would have caused strict harm. As a result, it is sensible in such cases for someone to argue that they were justified in causing harm, as the alternative would have been worse.

Bob, on the other hand, believes he has been strictly harmed, because he claims that he could have prevented the collision if he had been alerted. This disagreement can be captured formally by stating that Bob is using a three-valued variable F instead of a binary one, where the third option ($F = 2$) corresponds to alerting Bob. Incorporating this variable into the model (and assuming that Bob is correct regarding his driving skills) we would again get that $F = 1$ rather than $F = 2$ causes $O = 0.5$ rather than $O = 1$, but with the important distinction that H3 *is* satisfied for these contrast values and thus the system's action does strictly harm Bob. Our analysis does not resolve the conflict (and it is not meant to do so), instead it allows for a precise formulation of the source of the disagreement. □

4.4 Multiple Contrasts

The previous cases all involved a binary outcome; there were only two relevant events that could occur. CJR discuss cases that involve more than two possible events in order to argue against existing counterfactual accounts. The following example gives one instance of their argument.

Example 6 (Tear gas) The Joker sprays tear gas in exactly one of Batman's eyes. If he had not done that, he would have sprayed tear gas in both of Batman's eyes, which would have made Batman even worse off. One of the alternatives available to the Joker, however, was to simply leave Batman alone.

Intuitively here Joker harms Batman when he sprays him. To argue that the "incorrect" answer is obtained by the definition of harm they focus on, CJR consider a specific alternative event, namely, that Joker sprays tear gas in both of Batman's eyes, while observing that other alternatives (like leaving Batman alone) are also available. Rather than existentially quantifying over $\mathbf{x}'$, as we have done, (both in Definition 2 and the gloss of the counterfactual harm definition given in Definition 3), they take a version of counterfactual harm where $\mathbf{X} = \mathbf{x}'$ is taken to be the closest alternative to $\mathbf{X} = \mathbf{x}$ (according to some implicit, but unspecified, notion of closeness). Both our definition of harm and our gloss of the counterfactual definition (with the obvious assumptions about utility, and taking the default utility to be that of Batman being unharmed for our definition) agree that Joker did harm Batman in this case, as we would expect.

In this example, there are three events of interest (Joker sprays tear gas in one eye; Joker sprays tear gas in both eyes; Joker doesn't spray tear gas at all). We can model this using a variable TG that takes on three possible values (say, 0, 1, and 2). According to Definition 3, as long as one of them leads to a better utility than what actually happened, there was harm. But as the golf clubs example shows, this conclusion is not always justified; in general, we need to take defaults into account. □

Now we present an example, due to Shiffrin (1999), that illustrates the role of both the choice of the range of variables in the causal model and the choice of default.

Example 7 Betty is drowning in a fast-moving river. Veronica rescues her by grabbing her arm and pulling her out, accidentally fracturing Betty's humerus.

Did Veronica's rescue harm Betty? Shiffrin claims it does because Veronica could have pulled her out without breaking her arm. Indeed, Klocksiem (2012), in his analysis, points out that "it seems possible to rescue someone from drowning without breaking her arm". The first step in our analysis is to decide whether we should allow this possibility. That is, suppose that we have a variable P that describes how and whether Veronica pulls out Betty. We can take $P = 0$ if Veronica does not pull out Betty, $P = 1$ if she pulls her out by grabbing (and breaking) her arm. The modeler must then decide whether to allow P to take a value, say 2, where $P = 2$ if Veronica rescues Betty in such a way that Betty's arm is not broken. Reasonable people might disagree whether such an event is possible. First suppose we decide that P can take only values 0 and 1. Then the possible outcomes are that Betty drowns ($O = 0$) or Betty is saved ($O = 1$). In this model, any utility function that makes the utility of drowning worse than that of being saved would result in Veronica's rescue not harming Betty.

Now suppose that we allow $P = 2$. Then we would take $O = 1$ to represent Betty being saved but her arm being broken, and $O = 2$ to represent Betty being saved without her arm being broken. In that case, whether Veronica harms Betty depends on the default. If we take the default utility to be $\mathbf{u}(O = 2)$ then Veronica does cause Betty harm, while if we take it to be $\mathbf{u}(O = 0)$, she does not. Note that the latter choice is quite defensible. Given Betty's situation, making it out alive in whatever way possible would presumably be all that matters to her. □

This example clearly shows that to apply our framework in practice, it is important to have some guidelines on what count as a reasonable choice, both in the choice of variables and values and the choice of default value. As we mentioned in the introduction, Halpern and Hitchcock (2010) discuss this issue in the context of causal models; to the best of our knowledge, this issue has not been discussed in the context of default values. While this issue is beyond the scope of the current paper, we should make clear that we would not, in general, expect there to be a unique "correct" model. As we have said repeatedly, reasonable people can disagree about these choices.

There is one final issue we would like to address: why we consider a contrastive definition rather than just giving a definition in the spirit of the causal-counterfactual account. Definition 2 explicitly invokes a contrastive outcome o' whose utility is better than that of the actual outcome. We could have instead just defined harm as the result of causing an outcome whose utility is worse than the default.

One reason why we did not do so is that the default utility is not always achievable, and it would be counterintuitive to say that the agent was harmed if the outcome has a utility lower than the default, even though it is the best possible outcome. For example, there are diseases for which a surgery can only provide a temporary cure; in this case, a successful surgery gives the patient a temporary relief, and an unsuccessful surgery results in the patient's death. While the default utility for the patient, as for all people, is to be alive and healthy, saying that a successful surgery harmed the patient seems wrong. In fact, defining harm as the result of causing an outcome with the utility worse than the default provides counterintuitive results even when the default utility is achievable, as the following example demonstrates.

Example 8 (Pills) Consider the following vignette, again taken from CJR (where it is presented as a problem for both the causal-counterfactual and contrastive causal-counterfactual accounts):

> Barney suffers from a painful disease. On Monday, he can either take Pill A or not. On Tuesday, he will have another choice, between taking Pill B or not. Barney believes that he will be completely cured just in case he takes only Pill A, and partially cured just in case he takes both pills. Accordingly, he takes Pill A on Monday and does not take Pill B on Tuesday ...He is, however, misinformed about the effects of the pills. Taking only Pill A causes his disease to be merely partially cured. If he had taken both pills, he would have been completely cured. Had he not taken Pill A on Monday, on the other hand, nothing he could have done later would have produced even a partial cure.

To capture this in our framework, let O be a three-valued variable that captures Barney's health: $O = 2$ if he is fully cured, $O = 1$ if he is partially cured, and $O = 0$ if he is not cured at all. A and B capture whether or not Barney takes pills A and B respectively. The equation for O is then such that $O = 2$ if $A = B = 1$, $O = 1$ if $A = 1$ and $B = 0$, and $O = 0$ otherwise. As Barney considers taking pill B only if he fails to take pill A, the equation for B is $B = \neg A$. The context is such that $A = 1$; therefore, $B = 0$ and $O = 1$.

CJR claim that taking the pill does not harm Barney; we agree. Yet it easy to see that $A = 1$ rather than $A = 0$ causes $O = 1$ rather than $O = 0$. Indeed, $A = 1$ is a but-for cause of $O = 1$: had Barney not taken the pill, O would have been 0. It is easy to see why this is a problem for the causal-counterfactual account: Barney would have been better off if $O = 1$ had not obtained; specifically, he would be better off if O had been 2 (although this is not the outcome that results when changing A to 0 and therefore is not a problem for the counterfactual comparative account). CJR also view it as a problem for the contrastive causal-counterfactual account, because in applying it, they compare $O = 1$ to the outcome $O = 2$ (which, again, is not the outcome that obtains by switching A to 0), since they take the closest world to the one where Barney takes just one pill to be the world where he takes both pills. Since, unlike CJR, we do not use a similarity-based account of counterfactuals, we do not need to consider the "closest" state of affairs, so we avoid this problem. We simply compare $O = 1$ to the outcome $O = 0$ caused by switching to $A = 0$. $O = 0$ has utility worse than that of the outcome obtained from $A = 1$, so there is no harm according to our definition, for what we view as the “right” reasons. Assuming that the default utility is $\mathbf{u}(O = 2)$, $A = 1$ does cause an outcome whose utility is worse than the default and therefore a non-contrastive version of our definition would not have given the desired result.

We remark that the reliance on a similarity-based account of counterfactuals (and the modeling choices made for what the closest world is) are at the root of a number of other examples raised by CJR that they view as problematic (e.g., their “Stone” example). □

We conclude this section with one more example taken from CJR.

Example 9 (More tear gas) The Joker, who is very determined to hurt Batman, sprays tear gas in exactly one of Batman’s eyes. He does not have enough tear gas to spray it in both of Batman’s eyes. Riddler, equipped with his own can of tear gas, is tempted to follow the Joker’s noxious example, but eventually decides that enough is enough. Hence, if Batman had not had tear gas in exactly one eye, that would have been because Riddler would have sprayed additional tear gas in Batman’s other eye (whereby Batman would have had tear gas in both eyes). If the Joker had not sprayed tear gas Riddler would have left Batman alone (whereby Batman would not have had tear gas in any eye).

CJR point out that other accounts would take Joker’s action to be harmless, because without it, Batman would have been worse off (due to having tears sprayed in both eyes by Riddler). We completely agree with them that Joker’s action caused harm, and our account obtains this result. By our account, $J = 1$ rather than $J = 0$ causes $O = 1$ rather than $O = 0$, where the outcome O counts the number of eyes in which Batman has tear gas and J represents Joker’s spraying or not. To see this, simply take the fact that Riddler did not spray as the witness $\mathbf{W}$: holding this fact fixed, then clearly if we set J to 0, we would get that O=0. □

5 Discussion of H3

As we mentioned, H3 is intended to capture the intuition that there is no strict harm if the alternative would have resulted in an even worse outcome. For example, following the reasoning of the car manufacturer in Example 5, the system’s decision to drive into the fence rather than doing nothing is not strictly harmful because Bob would have suffered even worse injuries had the system done nothing. Since H1 and H2 are satisfied for this particular contrastive event, our definition would reach the opposite verdict if it weren’t for H3. Note that the counterfactual comparative account (Definition 3) also says that there is no harm: the alternative event under consideration would have given a worse outcome, so that C3 is not satisfied, and therefore there is no harm. Considering H3 gives more insight into the differences between the counterfactual account and ours.

Suppose that we consider some contrastive event $\mathbf{X} = \mathbf{x}'$ such that $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x} \land O = o$ and $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}'](O = o'')$, so C1 and C2 hold, and the first half of H2 holds if $o'' \neq o$: $\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}'$ causes $O = o$ rather than $O = o''$. H3 plays no role if H1 is not satisfied, so for simplicity, suppose that H1 also holds. Then it is easy to see that whenever $\mathbf{u}(O = o) \neq \mathbf{u}(O = o'')$, our definition

of strict harm gives the same verdict as the counterfactual comparative definition for this particular contrast (i.e., for this choice of $\mathbf{x}'$): if $\mathbf{u}(O = o) < \mathbf{u}(O = o'')$, then $o'' \neq o$, so H2 holds, as do H3 and C3; it follows that according to both definitions $\mathbf{X} = \mathbf{x}$ harms the agent. On the other hand, if $\mathbf{u}(O = o) > \mathbf{u}(O = o'')$, then neither C3 nor H3 hold (for this choice of $\mathbf{x}'$).

What happens if $\mathbf{u}(O = o) = \mathbf{u}(O = o'')$? This can happen for two reasons:

1. there is no but-for causation, that is, $o = o''$;
2. there is but-for causation but the counterfactual outcome $O = o''$ happens to have utility identical to the actual outcome.

Thus, roughly speaking (and ignoring the key role played by the default utility), our definition differs from the counterfactual comparative account only if $\mathbf{X} = \mathbf{x}$ rather than $\mathbf{X} = \mathbf{x}''$ is not a but-for cause of the actual utility: changing $\mathbf{x}$ into $\mathbf{x}''$ does not change the agent's utility.

Examples in which the first reason is relevant are widespread and crucial to our analysis, for those are precisely the examples in which actual causation (Definition 1) and but-for causation come apart. Our Late Preemption example (Example 2) offers one illustration, the literature on actual causation contains many more. An example where the second reason is relevant involves a more subtle way in which but-for causation comes apart from actual causation. Consider a "Sophie's choice" like setting: An agent must choose whether $X = 1$ or $X = 2$. There are two children, who will either live or die depending on the choice: if $X = i$ is chosen, then child i lives ($L_i = 1$) and child $3 - i$ dies ($L_{3-i} = 0$). The possible outcomes are that both children live (o_{11}), just child 1 lives (o_{10}), just child 2 lives (o_{01}), and neither child lives (o_{00}), where $d = \mathbf{u}(O = o_{11}) > \mathbf{u}(O = o_{10}) = \mathbf{u}(O = o_{01}) > \mathbf{u}(O = o_{00})$. In fact, $X = 1$ is chosen, so we get but-for causality, but switching from $X = 1$ to $X = 2$ gives an outcome of equal utility. However, if we hold $L_1 = 1$ fixed (which we can do in our framework to show causality) and switch to $X = 2$, then we get the outcome $O = o_{11}$. Thus, in our framework $X = 1$ strictly harms the agent; in the causal counterfactual framework, it does not.

This emphasizes the point we (and RBT) made that one set of problems that occur in defining harm is identical to the type of problems that occur in defining causation, and can be solved in the same way.

6 Comparison to RBT

In this section, we do a more careful comparison of our approach and that of RBT. RBT focus on choices made by an agent, where these are choices of what action to take, and assume that there is a default action, to which they compare the choice made by the agent. It follows easily from RBT's definition that if the agent performs the default action, there is no harm. (See Example 10 below for an illustration.) Yet there are many instances in which performing what seems like a perfectly reasonable default action does cause harm, albeit accidentally. Simply imagine a doctor prescribing a standard and very reliable medication to a patient, and the patient unfortunately suffering a very rare allergic reaction to the medication, where the reaction is far worse than the initial condition that the patient had. Then clearly the doctor harmed the patient. The most obvious choice of default action here is the actual action (and, in fact, RBT themselves mention following "clinical guidelines" as an example of a default action in their Appendix D). But this means that according to RBT's definition there would not be harm here. Although we use a default utility, there is no need for this default utility to be the utility of a default action, so we do not have this problem.

RBT are aware of this problem and discuss it in their Appendix D. They state that a "harm query" requires the specification of the default action. Thus, a single example can allow for different harm queries, each with its own default. This is how RBT attempt to avoid the conclusion that the doctor did not cause harm in the example above. But then what determines whether some action may legitimately be considered as the default for some harm query? (There is little purpose to the notion of default if any action can be taken as the default.) RBT do not offer a systematic answer, and in fact later suggest that each context determines a unique default action after all (saying that questions about harm "are asked within a context which implies which baseline comparison [to the default] should be made").

Furthermore, although RBT explicitly start out with the question of when an event causes harm (see their Question 1), the fact that their harm queries are relative to a particular choice of default means that they have failed to answer their question, a point that they acknowledge in Appendix D. To make matters worse, RBT's harm queries are also relative to a choice point of an entirely different kind, namely the choice of causal paths to consider.

This brings us to another significant difference between our approach and that of RBT: although RBT use causal models, unlike us, they do not use a sophisticated definition of actual causality such as the one given by Halpern (2015, 2016). In their Definition 3, RBT consider but-for causality. Not surprisingly, this will not suffice to deal with problematic examples where the more general notion of causality is needed (see, e.g., Example 2). They address this issue in their Appendix A by presenting Definition 9, which generalizes Definition 3 to allow for path-dependent causality. As was the case for the choice of default action, RBT do not offer a systematic explanation of how they choose which paths to consider, but in both Example 2 and the corresponding example of late preemption they consider, by choosing the appropriate paths, they can simulate the effects of AC2. (Specifically, they can simulate the effect of choosing a set **W** of variables and fixing them to their actual values.) As a consequence, they get the same results as those obtained by Halpern's definition of actual causality. It is not clear whether this will always be the case. More importantly, the ability to determine harm relative to some choice of paths gives the modeler a significant extra degree of freedom to tailor the results obtained. We believe that if paths are going to be used, there needs to be a more principled analysis of how to go about choosing them.

Lastly, RBT's focus on actions as the only events that can cause harm illustrates a more general underlying difference between their approach and ours: in many of their examples, they seem to conflate intuitions about explicitly moral judgments involving blame and responsibility with judgments about harm. Although we agree that harm may often be invaluable in forming judgments about blame and responsibility, we want to stress that we take judging an event to cause harm as morally neutral. On our analysis, which is supported by common usage of the term, natural events (like forest fires) and human actions alike can cause harm. Given that judgments about the former causing harm are morally neutral, we have to either conclude that the latter are as well, or come up with separate analyses of harm for natural events and human actions. As a result, although RBT's definition can trivially be generalized to include natural events (a point they make as well), the choice of what makes for an appropriate default value is not so easily generalized within their framework.

We present one of RBT's examples that they use to criticize our account to illustrate this issue. (We changed the names of the protagonists to Batman and Robin for ease of comparison.)

Example 10 (Omission problem) Batman can choose to give Robin his golf clubs or not. He has no obligation to do so. Unbeknownst to him, Eve is planning to rob Robin, but if Robin is holding a golf club she will not dare rob him. Batman decides not to give Robin his golf clubs and Robin is robbed by Eve. By choosing not to gift his clubs, did Batman harm Robin?

RBT point out (correctly) that under the natural choice of default utility – Robin is not robbed and does not have golf clubs—our account would answer this in the affirmative: Batman not giving the golf clubs rather than giving them causes Robin to get robbed rather than not, and thus harms Robin. RBT take this to be an undesirable outcome, for according to them this example is simply another case of failing to benefit, just as when Batman chose not to give his clubs without Robin getting robbed (Ex. 3).

We disagree. Causing Robin not to have golf clubs is significantly different from causing Robin to get robbed, and this difference is captured by the choice of default utility being the same in both examples, while observing that the actual utility is lower only in the latter. This does not in any way imply that Robin's action is blameworthy, but that is not the issue under consideration.

Yet, in many of RBT's examples, they make it sound as if it is the issue under consideration. For example, when discussing another alleged counterexample to our account in which Alice steals from Bob but Bob ends up getting reimbursed and is none the wiser, so that our account would not judge Bob to have been harmed, they state that "stealing from someone is harmful, regardless of whether or not someone else responds by reimbursing the victim". This statement only makes sense when interpreting harm as a moral notion, something closely related to blame. Similarly, they explicitly invoke intuitions about morality to defend their judgment about harm elsewhere (see their Case 2 of Preventing Worse). As mentioned, we disagree with this interpretation.

This disagreement is in fact intertwined with the different uses of default between our accounts, for example, the fact that RBT consider defaults with respect to causes as opposed to defaults with respect to (the utility of) effects. According to RBT, that Batman does not harm Robin in the example without robbing (Example 3) "relies on the judgement that Batman [Alice] does not have an ethical obligation to provide Robin [Bob] with golf clubs, therefore his choice does not constitute harm to Robin. In our definition of harm, this implies the obvious default action to be that of Batman not giving Robin clubs." Given that Batman's ethical obligations are identical in both examples, RBT have no choice but to conclude that even when Batman caused Robin to get robbed, he merely failed to benefit him as opposed to causing him harm. □

7 Conclusion

We have defined a qualitative notion of harm, and shown that it deals well with the many problematic examples in the philosophy literature. We believe that our definition will be widely applicable in the regulatory frameworks that we expect to be designed soon in order to deal with autonomous systems.

Of course, having a qualitative notion of harm is only a first step. For practical applications, we often need to quantify harm; for example, we may want to choose the least harmful of a set of possible interventions. As we said, we develop a quantitative notion of harm in Beckers et al. (2023). While one could just define a quantitive notion that considers the difference between the utility of the actual outcome and the default utility (this is essentially what RBT do), we believe that the actual problem is more nuanced. For example, even if we can agree on the degree of harm to an individual, if there are many people involved and there is a probability of each one being harmed, should we just sum the individual harms, weighted by the probability? We argue that this is not always appropriate, and discuss alternatives, drawing on work from the decision-theory literature.

Acknowledgements The authors would like to thank two reviewers of the journal for their detailed comments on the current article, and Jonathan Richens for a fruitful discussion of a preliminary version of the earlier conference paper. Hana Chockler was supported in part by the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1), the UKRI Strategic Priorities Fund to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1), and CHAI - EPSRC AI Hub for Causality in Healthcare AI with Real Data (EP/Y028856/1). Joe Halpern was supported in part by NSF grant IIS-1703846, ARO grant W911NF-22-1-0061, MURI grant W911NF-19-1-0217, and a grant from the Cooperative AI Foundation.

References

BBC News (2022). "Self-driving" cars to be allowed on UK roads this year. https://www.bbc.co.uk/news/technology-56906145.

Beckers, S. (2021). Causal sufficiency and actual causation. *Journal of Philosophical Logic, 50*, 1341–1374.

Beckers, S., Chockler, H., & Halpern, J. Y. (2022). A causal analysis of harm. *Advances in Neural Information Processing Systems, 35*, 2365–2376.

Beckers, S., Chockler, H. & Halpern, J.Y. (2023). Quantitative harm. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2023), pp. 363–371.

Beckers, S., Eberhardt, F. & Halpern, J.Y. (2019). Approximate causal models. In: Proc. 35th Conference on Uncertainty in Artificial Intelligence (UAI 2019). The full version appears at arxiv.org/abs/1906.11583.

Beckers, S. & Halpern, J.Y. (2019). Abstracting causal models. In: Proceeding Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). The full version appears at arxiv.org/abs/1812.03789.

Bontly, T. (2016). Causes, contrasts, and the non-identity problem. *Philosophical Studies, 173*, 1233–1251.

Bradley, B. (2012). Doing away with harm. *Philosophy and Phenomenological Research, 85*, 390–412.

Carlson, E., & Johansson, J. (2021). Causal accounts of harming. *Pacific Philosophical Quarterly, 103*(2), 420–445.

European Commission. (2021). Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://artificialintelligenceact.eu/the-act/. Accessed 8 Aug 8 2021.

Feit, N. (2019). Harming by failing to benefit. *Ethical Theory and Moral Practice, 22*, 809–823.

Gardner, M. (2015). A harm-based solution to the non-identity problem. *Ergo: An Open Access Journal of Philosophy, 2*, 427–444.

Glymour, C., & Wimberly, F. (2007). Actual causes and thought experiments. In J. Campbell, M. O'Rourke, & H. Silverstein (Eds.), *Causation and explanation* (pp. 43–67). MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies, 132*, 109–136.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Halpern, J., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science, 66*(2), 413–457.

Halpern, J. Y., & Hitchcock, C. (2010). Actual causation and the art of modeling. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Causality, probability, and heuristics: A tribute to Judea pearl* (pp. 383–406). College Publications.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: a structural-model approach. Part I: Causes. *British Journal for Philosophy of Science, 56*(4), 843–887.

Halpern, J.Y. (2015). A modification of the Halpern-Pearl definition of causality. In: Proceeding of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 3022–3033.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy XCVII, I*(6), 273–299.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review, 116*, 495–532.

Klocksiem, J. (2012). A defense of the counterfactual comparative account of harm. *American Philosophical Quarterly, 49*(4), 285–300.
Law Commission. (2022). Legal reforms to allow safe introduction of automated vehicles announced. https://www.lawcom.gov.uk/legal-reforms-to-allow-safe-introduction-of-automated-vehicles-announced/.
National Library of Medicine. (2002). Hippocratic Oath. https://www.nlm.nih.gov/hmd/greek/greek_oath.html.
Northcott, R. (2015). Harm and causation. *Utilitas, 2*, 147–164.
Resnik, M. D. (1987). *Choices: An introduction to decision theory*. University of Minnesota Press.
Richens, J. G., Beard, R. & Thompson, D. H. (2022). Counterfactual harm. In *Proceeding of the Advances in Neural Information Processing Systems 35 (NeurIPS '22)*.
Savage, L. J. (1954). *Foundations of statistics*. Wiley.
Shiffrin, S. (1999). Wrongful life, procreative responsibility, and the significance of harm. *Legal Theory, 5*, 397–413.
Suits, D. B. (2001). Why death is not bad for the one who died. *American Philosophical Quarterly, 38*, 69–84.
Thomson, J. J. (2011). More on the metaphysics of harm. *Philosophy and Phenomenological Research, 82*, 436–458.
Weslake, B. (2015). A partial theory of actual causation. British Journal for the Philosophy of Science.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.