Quantifying Harm

Sander Beckers[1] , **Hana Chockler**[2] and **Joseph Y. Halpern**[3]
[1]Institute for Logic, Language, and Computation, University of Amsterdam
[2]Department of Informatics, King's College London
[3]Computer Science Department, Cornell University
srekcebrednas@gmail.com, hana.chockler@kcl.ac.uk, halpern@cs.cornell.edu

Abstract

In [Beckers *et al.*, 2022a], we defined a qualitative notion of harm: either harm is caused, or it is not. 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. We first present a quantitative definition of harm in a deterministic context involving a single individual, then we consider the issues involved in dealing with uncertainty regarding the context and going from a notion of harm for a single individual to a notion of "societal harm", which involves aggregating the harm to individuals. We show that the "obvious" way of doing this (just taking the expected harm for an individual and then summing the expected harm over all individuals) can lead to counterintuitive or inappropriate answers, and discuss alternatives, drawing on work from the decision-theory literature.

1 Introduction

AI systems are playing an ever-expanding role in making decisions, in applications ranging from hiring and interviewing to healthcare to autonomous vehicles. Perhaps not surprisingly, this is leading to increasing scrutiny of the harm and benefit caused by (the decisions made by) such systems. To take just one example, the new proposal for Europe's AI act [European Commission, 2021] contains over 29 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, taking into account both the severity of the possible harm and its probability of occurrence …" [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]. Unfortunately, defining harm appropriately has proved difficult. 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. …My diagnosis is that the notion of harm is a Frankensteinian jumble …It should be replaced by other more well-behaved notions.

In [Beckers *et al.*, 2022a], we defined a qualitative notion of harm (was there harm or wasn't there) in deterministic settings with no uncertainty and only a single agent, which dealt well with all the difficulties raised in the philosophy literature (which also focused on qualitative harm in deterministic settings; see [Carlson *et al.*, 2021] for an extensive overview). The key features of our definition are that it is based on causal models and the definition of causality given by Halpern [2015; 2016], assumes that there is a *default* utility, and takes harm to be caused only if the outcome has utility lower than the default.

While getting such a definition is an important first step, it does not address the more quantitative aspects of harm, which will clearly be critical in comparing, for example, the harm caused by various options, and for taking into account "both the severity of the possible harm and its probability of occurrence", as suggested in the European AI Act proposal. In this paper, we extend our earlier definition so as to provide a quantitative notion of harm.

The first step is relatively straightforward: we define a quantitative notion of harm in a deterministic setting. Roughly speaking, we take the amount of harm to be the difference between the actual utility and the default utility. Once we have this, we need to be able to aggregate harm across different settings. There are two forms of aggregation that we must consider. The first involves dealing with uncertainty regarding the outcome. Here we confront issues that are well known from the decision-theory literature. There have been many rules proposed for making decisions in the presence of uncertainty: maximizing expected utility, if uncertainty is characterized probabilistically; *maximin* (maximizing the worst-case utility) [Wald, 1950] or *minimax regret* [Niehans, 1948; Savage, 1951] if there is no quantitative characterization of uncertainty; *maximin expected utility* if uncertainty is described using a set of probability measures [Gärdenfors and Sahlin, 1982; Gilboa and Schmeidler, 1989]. We consider one other approach—*probability weighting*—shortly. All of these approaches can be applied to harm.

Another issue that has received extensive attention in the decision-theory literature and applies equally well to harm is that of combining utilities or harms of different people. This issue arises when we must determine the harm caused to society of, say, a vaccination treatment, where perhaps some people will react badly to the vaccine. Even assuming that we can compute the harm caused to each individual, we must consider the total harm caused to all individuals. An obvious approach would be to just sum up the harm caused to each individual, but this assumes that individuals are somehow commensurate, that is, that one person's harm of 1 should be treated identically to another person's harm of 1. Even if we are willing to accept this, there is another issue to consider: fairness. Suppose that we have two policies, each of which cause 1 unit of harm to 1,000 people in a population of 100,000 (and cause no harm to anyone else). We would feel quite differently about a policy if the 1,000 people to whom harm was caused all came from a particular identifiable population (say, poor African-Americans) than if the 1,000 people were effectively chosen at random.

Finally, when different policies result in different probabilities of people being harmed, additional subtleties arise. Heidari et al. [2021] (HBKL from now on) consider a number of examples of government policies that may cause harm to each member of a population of n individuals. (For simplicity, they assume that if harm is caused, there is 1 unit of harm.) Suppose that the harm caused by a policy P is characterized by the tuple $(p_1, \ldots, p_n)$, where p_i is the probability that individual i suffers 1 unit of harm. Thus, the total expected harm of policy P is $p_1 + \cdots + p_n$. As HBKL point out and is underscored by Example 1, we may feel very differently about two policies, even if they cause the same amount of expected harm. For example, we feel differently about a policy that necessarily harms individual 1 and does not harm anyone else compared to a policy that gives each individual a probability $1/n$ of being harmed. Indeed, there is a long line of work in psychology [Jenni and Loewenstein, 1997] that suggests that we find it particularly troubling to single out one victim and concentrate all the risk of harm on him. (This is clearly related to the issue of unfairness to subpopulations.) HBKL suggest getting around these issues by aggregating harm using an approach familiar from the decision theory literature: *probability weighting*. The idea is to apply a weight function w to the probability and to compute the weighted expected harm. Under the simplifying assumption used above that, if harm is caused, it is always 1 unit of harm, the weighted expected harm would be $w(p_1) + \cdots + w(p_n)$ (so we get back the standard expression for expected harm by taking the weighting function w to be the identity (cf. [Prelec, 1998; Quiggin, 1993]).

As HBKL point out, the policies that are often adopted in practice seem to be the ones that optimize weighted expected harm if we use the probability weighting functions that empirical work has shown that people use. HBKL take the probability function to be one that overweights small probabilities and underweights larger probabilities. While this works well for their examples, the situation is actually more nuanced. To quote [Kahneman and Tversky, 1979, p. 283] (who were the first to raise the issue):

> Because people are limited in their ability to comprehend and evaluate extreme probabilities, highly unlikely events are either neglected or overweighted, and the difference between high probability and certainty is either neglected or exaggerated. Thus, small probabilities generate unpredictable behavior. Indeed, we observe two opposite reactions to small probabilities.

Indeed, as we shall see, there are examples best explained by assuming people essentially ignore small probabilities, effectively treating them as 0, and others that are best explained by people overweighting small probabilities.

Richens, Beard, and Thompson [2022] (RBT from now on) also proposed a quantitative and causality-based definition of harm. We already discussed what we take to be problems in their approach in our paper on qualitative harm; they carry over to the quantitative setting as well. Consider the following example that they use to motivate their approach:

Example 1. *Consider two treatments for a disease which, when left untreated, has a 50% mortality rate. Treatment 1 has a 60% chance of curing a patient, and a 40% chance of having no effect, in which case the disease progresses as if untreated (so that there is a 50% mortality rate). Treatment 2 has an 80% chance of curing a patient and a 20% chance of killing them. Treatments 1 and 2 have identical recovery rates, yet doctors systematically favor Treatment 1.*

We agree with RBT that the explanation for this lies in the fact that Treatment 1 causes less harm than Treatment 2. However, we offer a different analysis that results in differences in the degree of harm. Specifically, for RBT, Treatment 1 never causes harm whereas Treatment 2 harms 10% of all patients (namely those patients who would have recovered had they not been given Treatment 2). On our analysis, Treatment 1 harms 16% of all patients, compared to 20% for Treatment 2. These quantitative differences arise due to our different views on qualitative harm; we leave a detailed discussion of this example to the full paper [Beckers *et al.*, 2022b], and return to a discussion of RBT in Section 7.

The rest of the paper is organized as follows. In Section 2 we briefly review causal models and the definition of actual causality, since these form the basis of our definition. In Section 3 we provide the definition of quantitative harm in a single context for a single agent; in Sections 4 and 5, we discuss how to extend this basic definition to situations where there is uncertainty about the context and there are many individuals, each of which may potentially suffer harm. In Section 6, we briefly discuss analogous definitions for benefits. In Section 7 and in the full paper [Beckers *et al.*, 2022b], we compare our work to that of RBT.

2 Causal models and actual causality

We start with a review of causal models and actual causation, 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 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 $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$. 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 $\vec{u}$ 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). In this paper, following the literature, we restrict to recursive models. We call a pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ a *(causal) setting*.

A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$, where ϕ is a Boolean combination of primitive events, $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\phi$. The special case where $k = 0$ is abbreviated as ϕ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$ says that ϕ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ 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 $\vec{u}$ (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 $\vec{u}$). Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$, where $M_{\vec{Y} \leftarrow \vec{y}}$ is the causal model that is identical to M, except that the equations for variables in $\vec{Y}$ in $\mathcal{F}$ are replaced by $Y = y$ for each $Y \in \vec{Y}$ and its corresponding value $y \in \vec{y}$.

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; Hitchcock, 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 [Halpern, 2016] calls the *modified* Halpern-Pearl definition, 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 find it useful to give a contrastive variant of the definition of actual causality; moreover, we are interested only in whether $\vec{X} = \vec{x}$ causes an outcome $O = o$. Thus, rather than defining what it means for $\vec{X} = \vec{x}$ to be an (actual) cause of an arbitrary formula ϕ, we restrict ourselves to defining what it means for $\vec{X} = \vec{x}$ *rather than* $\vec{X} = \vec{x}'$ to be a cause of $O = o$ rather than $O = o'$.

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

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge O = o.$

AC2. *There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{w}$ of the variables in $\vec{W}$ such that $(M, \vec{u}) \models \vec{W} = \vec{w}$ and $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]O = o'$, where $o \neq o'$.*

AC3. *$\vec{X}$ is minimal; there is no strict subset $\vec{X}''$ of $\vec{X}$ such that $\vec{X}'' = \vec{x}''$ can replace $\vec{X} = \vec{x}'$ in AC2, where $\vec{x}''$ is the restriction of $\vec{x}$ to the variables in $\vec{X}''$.*

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of $O = o$ unless both $\vec{X} = \vec{x}$ and $O = o$ actually happen. AC3 is a minimality condition, which says that a cause has no irrelevant conjuncts. AC2 captures the standard but-for condition ($\vec{X} = \vec{x}$ rather than $\vec{X} = \vec{x}'$ is a cause of $O = o$ if, had $\vec{X}$ been $\vec{x}'$ rather than $\vec{x}$, $O = o$ 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 $(M, \vec{u})$. In the special case that $\vec{W} = \emptyset$, we get the standard but-for definition of causality: if $\vec{X} = \vec{x}$ had not occurred (because $\vec{X}$ was $\vec{x}'$ instead) $O = o$ would not have occurred (because it would have been $O = o'$).

3 Quantitative harm in a single context for a single agent

In this section, we extend the qualitative notion of harm in a given context introduced in our previous work [Beckers *et al.*, 2022a] to a quantitative notion. Both the qualitative and the quantitative notions are defined relative to a particular context in a causal utility model, which is just like a causal model, except that there is a default utility d, and it is assumed that there is a a special endogenous variable O (for outcome), whose value determines the utility. The fact that harm is defined relative to a given context (just like causality) means that, implicitly, there is no uncertainty about the context.

Formally, a *causal utility model* is a tuple $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathbf{u}, d)$, where $((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ is a causal model one of whose endogenous variables is O, $\mathbf{u} : \mathcal{R}(O) \rightarrow I\!R$ is a utility function on outcomes, and $d \in I\!R$ is a default utility. Like causation, harm is assessed relative to a setting $(M, \vec{u})$.

Definition 2. *If $\vec{X} = \vec{x}$ rather than $\vec{X} = \vec{x}'$ causes $O = o$ rather than $O = o'$ in $(M, \vec{u})$, where $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathbf{u}, d)$, then the* (quantitative) harm *to agent* **ag** *relative to $(\vec{X} = \vec{x}', O = o')$, denoted $QH(M, \vec{u}, \vec{X} = \vec{x}', O = o')$, is $\max(0, \min(d_h, \mathbf{u}(o')) - \mathbf{u}(o)))$. The* quantitative harm *to agent* **ag** *caused by $\vec{X}$ in $(M, \vec{u})$, denoted $QH(M, \vec{u}, \vec{X})$, is $\max_{\vec{x}', o'} QH(M, \vec{u}, \vec{X} = \vec{x}', O = o')$ if there is some $\vec{x}'$ and o' such that $\vec{X} = \vec{x}$ rather than $\vec{X} = \vec{x}'$ causes $O = o$ rather than $O = o'$; if there is no such $\vec{x}'$ and o', then the quantitative harm is taken to be 0. (Note that the values $\vec{x}$ and o are uniquely determined by $(M, \vec{u})$, which is why they do not need to appear in the parametrization.)*

In other words, the quantitative harm caused by $\vec{X} = \vec{x}$ is the maximum difference between the default utility or the utility of the contrastive outcome, whichever is lower, and the utility of the actual outcome (except that we take the harm to be 0 if this difference is negative or if $\vec{X} = \vec{x}$ did not cause the actual outcome). Definition 2 is a generalization of our definition of qualitative harm. Quantitative harm as we have defined it here is positive iff there is qualitative harm.[1]

As mentioned in the introduction, decision theory often focuses on maximizing (expected) utility. In many cases this corresponds to minimizing the quantitative harm, but as the following example illustrates, the two approaches can come apart even if we restrict to a single context and a single agent.

Example 2. *Alice has a meal in a restaurant. The bill comes to* $\$100$. *Let $O = o$ be the variable representing the tip, and let the utility be $o/100$. That is, $\mathbf{u}(\$100) = 1$, and $\mathbf{u}(\$20) = 0.2$. It is customary to give a 20% tip, hence it seems reasonable to take the default utility to be 0.2. However, Alice only has* $\$5$ *in her wallet, the restaurant accepts only cash tips, and there is no ATM nearby. The outcome that maximizes the waiter's utility is thus for Alice to tip \$5, corresponding to Alice giving the waiter all the cash she has. By Definition 2, if Alice gives* $\$5$*, the waiter is not harmed. If, on the other hand, Alice gives only* $\$1$*, the waiter is harmed, and the harm is $0.05 - 0.01 = 0.04$, which is the difference between the maximum utility and the actual utility.*

Now suppose that Alice in fact has $\$30$ *in her wallet. Then Alice would maximize the waiter's utility with a tip of \$30. Yet if our goal is to minimize harm, then any tip of \$20 or more results in a harm of 0.*

[1] In [Beckers *et al.*, 2022a] we make a distinction between harm and *strict* harm: strict harm adds a further requirement, denoted H3, to the definition of harm. We argued in our companion paper that H3 rarely plays a role, so we have chosen to ignore it here for ease of exposition. However, we could define quantitative strict harm as being identical to quantitative harm, except that we take the quantitative strict harm to be 0 whenever H3 is not satisfied. (See the full paper for the formal definition of both qualitative harm and strict harm, as well as results on the complexity of computing harm.)

4 Quantitative harm when there is uncertainty about contexts

In general, there may be uncertainty both about the causal model (i.e., how the world works, as described by the equations) and the context (what is true in the world). In decision theory, this uncertainty is usually taken into account by computing the expected utility, where the expectation is taken with respect to a known probability distribution. Analogously, we could define the notion of expected quantitative harm by simply computing the product of the quantitative harm in each causal setting and the probability of that setting. The next example illustrates that even using this straightforward generalization of harm already results in some interesting differences with expected utility.

Example 3. *Suppose that a doctor has a choice of either prescribing medication ($X = 1$) or performing surgery ($X = 0$) on a patient. The medication keeps the patient stable, but does not completely cure the patient. Call this outcome $O = 1$, and assume that it has utility $.5$. On the other hand, the surgery cures the patient completely with probability $1-p$ ($O = 0$, with utility 1), but has a small probability p of the patient dying ($O = 2$, with utility 0), due to factors such as the patient's tolerance of anesthesia and the surgeon's skill.*

The expected utility of $X = 1$ is 0.5, while the expected utility of $X = 0$ is $1 - p$. Assuming that $p < 0.5$, $X = 0$ is the choice that maximizes expected utility. If we take the default utility to be 1, which is reasonable if the patient views any deviation from their normal health as unacceptable, then the harm caused by $X = 1$ is 0.5, while the expected harm caused by $X = 0$ is p, so minimizing expected harm would again lead to choosing $X = 0$. However, suppose that the patient has been taking the medication for some time, and has gotten used to the treatment. In this case, 0.5 seems like a reasonable choice for the default. With this choice, $X = 1$ has expected harm 0, while $X = 1$ has expected harm $0.5p$, so the choice that minimizes expected harm is $X = 1$. Intuitively, by taking an appropriate choice of default utility, a harm-based approach allows us to capture the idea that one should not risk obtaining a bad outcome when there exists an alternative that is guaranteed to result in an outcome that is good enough.

While taking expectation is very natural, it sometimes leads to unreasonable conclusions.

Example 4. *Research has shown that the probability of a fatal accident when driving at the speed limit is 1 in a million,* *and that driving at* 80% *of the speed limit results in* 50% *fewer fatal accidents than driving at the speed limit, so the probability of a fatal accident when driving at 80% of the speed limit is 1 in 2,000,000. However, research has also shown that the majority of people do drive at the speed limit, and would prefer buying a driverless car that does so as well. Furthermore, this preference remains even after people have been informed of these numbers. Based on this research, the manufacturer of a driverless car company needs to implement a policy regarding the typical driving speed of their cars. Either cars drive at the maximum speed allowed by the speed limit,* $X = 1$*, or cars drive at* 80% *of the speed limit,* $X = 0$*. (Obviously a more realistic model would here use a continuous variable* X*.)*

For any given trip, there are three outcomes: $O = 2$ *if the driver arrives safely at its destination in the quickest way (legally) possible,* $O = 1$ *if the driver arrives safely at its destination but takes a bit more time, or* $O = 0$ *if the car crashes and the driver dies. For each driver, the utilities are* $\mathbf{u}(O = 2) = 1$*,* $\mathbf{u}(O = 1) = 0.9$*,* $\mathbf{u}(O = 0) = -1,000,000$*.*

Maximizing expected utility results in a preference for $X = 0$*, but this does not match how people react. Taking the default utility to be 1, which seems reasonable, and minimizing expected harm leads to the same preference. Nor does it help to take the default to be* 0.9*.*

We can deal with this problem by using the idea of probability weighting from the decision-theory literature. We assume that agents use a *probability weighting function* w, where $w : [0, 1] \rightarrow [0, 1]$. In order to make use of this function, from now on we take a causal utility model to also include a probability $\Pr$ over the exogenous settings $\vec{u} \in \mathcal{R}(\mathcal{U})$. (Note that, in general, there might also be uncertainty regarding the causal equations, but for ease of exposition, we ignore this complication here.)

Definition 3. *The* weighted quantitative harm (WQH) *to agent* **ag** *caused by* $\vec{X}$ *relative to model* M *and weighting function* w *is*

$$WQH(M, \vec{X}, w) = \sum_{\vec{u} \in \mathcal{R}(\mathcal{U})} w(\Pr(\vec{u})) QH(M, \vec{u}, \vec{X}).$$

Applying Definition 3 to our example, taking M to be a causal model of the driving situation, we can assume for simplicity that there are three contexts of interest: in u_0, the agent does not have a fatal accident if either $X = 0$ or $X = 1$, in u_1 he has a fatal accident if $X = 1$ but not if $X = 0$, and in u_2 he has a fatal accident if either $X = 0$ or $X = 1$. We can then take the probabilities to be $999,999/1,000,000$ for u_0 and $1/2,000,000$ for both u_1 and u_2. Deciding on a policy then amounts to determining the equation for X: either we choose $X = 1$, or we choose $X = 0$. (Of course, in general, more complicated policies can be considered.)

In practice, people tend to discount the probability of fatal accidents; they treat it as being essentially 0. We can capture this by taking, for example, $w(1/2,000,000) = 0$ and $w(999,999/1,000,000) = 1$. Sure enough, for this choice of w, the weighted harm of $X = 1$ is lower than that of $X = 0$.

In this case, w underweights the low probabilities. But in other cases, people overweight probabilities. As Gigerenzer [2006] observed, after the terrorist attack on September 11, 2001, a lot of Americans decided to reduce their air travel and drive more, presumably because they were overweighting the likelihood of another terrorist attack. (As Gigerenzer points out, the net effect was a significant increase in the number of deaths.) As mentioned in the introduction, HBKL give other examples where overweighting gives answers that seem to match how people feel about issues.

Perhaps the most common explanation for this effect is that people underweight probabilities when they make "decisions from experience" (i.e., based on their past experience with the events of interest), although this can flip due to recent bad experiences (as in the case of a terrorist attack) and overweight probabilities when they make "decisions from description" (i.e., the type of situation studied in a lab, where a situation is described in words) [Hertwig *et al.*, 2004]. In our example, agents' experience is that people never have fatal accidents, so they underweight the probability. On the other hand, if the agent recently had a death in the family due to a fatal accident, it is likely that he would use a w that overweights these probabilities.

As we shall see in the next section, combining probability weighting with harm also lets us deal with other apparently paradoxical observations due to Norcross [1998].

5 Aggregating harm for different individuals

Up to now we have considered harm for a single individual. Further issues arise when we try to aggregate harm across many individuals, as we will certainly need to do when we consider societal policies. The most straightforward approach to determining "societal harm" when there are a number of individuals involved is to sum the harm done to each individual. By using defaults appropriately, this straightforward approach already lets us avoid some obvious problems with maximizing expected utility.

Example 5 (Forced organ donation)**.** *Suppose that Billy is a healthy person, strolling by a hospital. In the hospital, there are* 5 *patients in need of a heart, liver, kidney, lung, and pancreas transplant, respectively. Suppose for simplicity that these patients will die without the transplant, and it is not available elsewhere, while Billy will die if these organs are harvested from him. Expected utility maximization would suggest that saving five lives is better than saving one, so the hospital should kidnap Billy. On the other hand, if we take the default that Billy and each of the patients continue in their current state of health, then harvesting Billy's organs clearly harms Billy, while not harvesting Billy's organs harms no one, and is thus the action that minimizes harm.*

Combining probability weighting with harm also lets us deal with other apparently paradoxical observations due to Norcross [1998] that we mentioned in the previous section. Norcross considers three events, where it seems that A results in more harm than B which results in more harm than C which results in more harm than A, leading to an inconsistent cycle. A is the event that one person dies a premature death; B is the event of 5,000,000 people suffering moderate headaches, and C is the event that 5,000,000 people each incur a one in a million risk of dying. Norcross claims that most people would take A to involve greater harm than B and would continue to do so if we replaced the "5,000,000" in B by any other number. Yet clearly, if we just add up harms, then as long as the harm of a moderate headache is positive, there must be some number N of people such that N people suffering moderate headaches results in greater harm than a single premature death. Norcross further offers a scenario to argue that most people view C as involving greater harm than A. Finally, Norcross provides a scenario where B seems to involve greater harm than C.

It is instructive to look more carefully at the precise scenarios that Norcross considers. To argue that B involves greater harm than C, Norcross considers a scenario where 5,000,000 people each drive out to get headache medication (under the assumption that driving results in a one in a million risk of dying). On the other hand, to argue that C involves greater harm than A, Norcross considers a scenario where 5,000,000 people in a city running some small risk of dying from some poisonous gas. While both scenarios involve 5,000,000 people incurring a small risk of dying, the nature of the stories is quite different. In one case, the risk involves something that people do every day—driving—which their personal experience tells them involves a negligible risk of death. On the other hand, the second scenario involves a scary new risk (which presumably people read about, rather than having personal experience with). The former scenario is one where people are likely to underweight the probability of death (essentially treating it as 0), while the second scenario is one where people overweight the small probability. Thus, although the actual probabilities are the same, the weighted probabilities are quite different, and hence the weighted expected harm is quite different. There is actually no cycle here. Rather, there are two quite different instances of C, for which people compute the harm very differently.

But there is a whole set of other issues that arise when dealing with societal harms: there is a concern that we may disproportionately affect certain identifiable groups. For example, a policy requiring certain people to work during a pandemic may have a disproportionate impact on certain groups. These groups may be the gender-based or ethnicity-based groups traditionally considered in the fairness literature, but in general, they need not be. For example, a new freeway may lead to a disproportionate harm to people living in a certain completely integrated middle-class neighborhood. The "group" might be just an individual. Indeed, we can see Norcross's scenario A as an instance of this phenomenon, where the group is the individual that suffers the premature death, which is intuitively more harmful than any number of people suffering a moderate headache.

We now briefly sketch a more formal approach to computing harm that takes this type of fairness into account. We assume that the groups that cannot be disproportionately harmed by a policy must be identified in advance. A model would include a list of all such groups.

Definition 4. *A* collective utility model *is a tuple* $((\mathcal{U},\mathcal{V},\mathcal{R}),\mathcal{F},\Pr,A,\mathbf{u}_A,d_A,\mathcal{G},\alpha,\beta)$*, where* $((\mathcal{U},\mathcal{V},\mathcal{R}),\mathcal{F})$ *is a causal model,* $\Pr$ *is a probability on contexts,* $\mathbf{u}_A$ *and* d_A *consist of the utility functions and default utilities for each agent in* A*,* $\mathcal{G}$ *is a set of* identifiable *subsets of* A*, and* α *and* β *are two additional real-valued parameters, to be explained shortly. Given* $\vec{X}$*, we can compute the (weighted) harm caused by* $\vec{X}$ *to each agent* $a \in A$*, according to Definition 3. We then sum the harms caused to each agent, but add a penalty* α *if some group in* $G \in \mathcal{G}$ *is disproportionately harmed, where* G *is* disproportionately harmed *if the average harm caused to the agents in* G *is* β *greater than the average harm caused to the agents in* A*.*

Intuitively, $\mathcal{G}$ consists of sets of agents that should not be disproportionately harmed. $\mathcal{G}$ could include, for example, all "small" sets of agents (say, all sets of size at most 5), since people may consider it unfair that a small group of agents should suffer disproportionately. Note that we can take α large so that if there is a policy that does not harm any identifiable group, it is guaranteed to be preferred to one that does harm an identifiable group. On the other hand, if every policy harms some identifiable group, then we are back to comparing policies by just summing the harm caused to individuals.

6 Harm vs. benefit

In many situations, we need to trade off benefits and harms. Although many authors view benefit as the opposite of harm [Carlson *et al.*, 2021; Richens *et al.*, 2022], we here suggest a definition for which this is not necessarily the case, while still allowing for the aggregation of benefits and harms. We replace the default value d by an *interval* $D = [d_h, d_b]$, where utility lower than d_h is a harm, utility higher than d_b is a benefit, and all values within D are neither harm nor benefit. To motivate choosing an interval rather than a single value, we can go back to the tipping example. We can imagine that there is an acceptable range $[d_h, d_b]$ of tips. Tips below d_h are unacceptable, and viewed as harms; tips above d_b are particularly generous, and viewed as benefits. We would not expect $d_h = d_b$, in general. That said, when doing a cost-benefit analysis, we quite often do take there to be a baseline, where anything below the baseline is a cost, and anything above it is a benefit (which amounts to taking $d_h = d_b$).

7 Comparison to RBT

As mentioned earlier, RBT also proposed a quantitative and causality-based definition of harm. Our previous paper outlined several objections to their qualitative definition; here we focus instead on the differences in the quantitative definition.

While both we and RBT distinguish causing harm from causing a decrease in utility, and we both have a notion of default, there are several significant differences between our approach and theirs. Rather than having a default utility, RBT assume a default action; in a fixed context (i.e., what we consider in Section 3), we could choose to take the default utility to be the utility of the default action. However, when computing harm, RBT only do a pairwise comparison of the harm of a given action to the harm of the default action, and use only but-for causation rather than the more general definition of causation given by, say, [Halpern, 2016]. As the analysis of Example 1 in the supplementary material shows, this can lead to a significant difference in the calculation of harm. Moreover, assuming a default action (in this case, that of providing no treatment) seems to lead to an inappropriate conclusion that is avoided by using a default utility instead.

Going on, rather than minimizing expected harm (and possibly applying a weighting function to the probability, as we do), when deciding which action a to perform, RBT maximize a different utility function, namely $U[a] - \lambda h[a]$, where U is the expected utility of a, λ is a user-dependent *harm aversion* coefficient, and $h[a]$ is the harm as calculated above, that is, the decrease in utility caused (in the "but-for" sense) by doing a rather than the default action. Not surprisingly, this leads to quite different harms than we would calculate.

Furthermore, RBT do not pay special attention to the issues that arise when aggregating harm, but instead simply compute societal harm by summing individual harms. As discussed in Section 5, this approach can lead to preferences that do not match those of most people.

Finally, RBT view benefit as the opposite of harm (i.e., in the notation of Section 6, they take $d_h = d_b$). As we pointed out, in general it seems more appropriate not to treat benefit and harm symmetrically, and allow for a default interval.

8 Discussion and conclusion

We have given a formal definition of quantitative harm, based on our earlier definition of qualitative harm. While the definition of quantitative harm for a single individual in a fixed context, where there is no uncertainty, is fairly straightforward given our definition of qualitative harm, as we have pointed out, there are subtleties that arise when we add probability and when we need to take into account fairness issues. We have suggested an approach for dealing with fairness issues, but clearly work needs to be done to understand the extent to which it captures how people actually deal with these issues. For people to be comfortable with policies enacted by, for example, government agencies (such as the European AI Act), the formal approach will have to be reasonably close to their heuristics. The situation with probability overweighting and underweighting is even more subtle. Research has shown that people do both overweight and underweight low-probability events (see, e.g., [Hertwig *et al.*, 2004; Zielonka and Tyszka, 2017]). We suspect that the underweighting that occurs when people make decisions from experience could itself reflect a normative preference. Perhaps there are actions (and their consequences) with which we have experience precisely because we consider them to be part of our normal lives. As a result, we are prepared to accept higher risks resulting from such actions than from actions (or events) which are considered abnormal or neutral. This seems to fit well with the distinction between scenarios B and C from the Norcross example: people consider the flexibility of being able to drive to the pharmacy whenever they so choose to be part of a normal life, whereas presumably they do not particularly value the ability to live near a factory that produces poisonous gas. In any case, while we do have some understanding of when overweighting and underweighting occurs, a policy-maker will have to weigh normative and descriptive considerations in deciding how to compute societal harm; assuming that people always overweight low-probability events, as HKBL do, is clearly not appropriate (although it may well be appropriate for the applications considered by HKBL). Although we have focused on probabilistic representations of uncertainty, another direction worth exploring is non-probabilistic representations of uncertainty.

In addition to the weighting of probabilities, quantitative harm is influenced by the default utility: what matters is the difference between the default utility and the utility of the outcome (rather than just the utility of the outcome). Although we have here argued for this view simply by showing that it accords well with intuition for the examples discussed, recent empirical research shows that people do seem to take into account a context-dependent default in precisely this manner. Indeed, Rigoli et al. [2016] have shown that people make different choices when confronted with two cases that have identical causal structure and identical probability distributions over the possible monetary outcomes, but where the first (second) case is stipulated to be a low-value (resp., high-value) context. Crucially, this behaviour is observed despite the fact that the subjects are informed that the context *does not* influence the probabilities of the outcomes. Rigoli et al. explain their results by assuming that the context changes the utility function itself, but their experiments can just as easily be explained by assuming that the context changes the default utility instead: in a low-value context, people take the default utility to be lower than in a high-value context, and therefore the same utility results in different amounts of quantitative harm for each context. It would be interesting to construct experiments that can distinguish between the two proposals.

Finally, it is worth mentioning complexity considerations. As with most concepts involving actual causality, deciding whether harm occurred is intractable even in the single-agent qualitative case where there is no uncertainty. Indeed, it is easy to see that it is at least as hard as causality, which is proved in [Halpern, 2015] to be DP-complete. Adding quantitative considerations does not make it any easier, although we suspect it will not be any harder either. That said, we do not believe that in practice, complexity considerations will be a major impediment to applying these definitions. In many cases of interest, the set of variables and their possible values is small. Exhaustive search is polynomial in the set of combinations of possible values of the variables, so the problem will be polynomial time in this case. Furthermore, if we consider but-for causality (i.e., take $\vec{W} = \emptyset$ in AC2), which often suffices, then the problem becomes polynomial time in the number of combinations of possible values of $\vec{X}$.

This paper (and our previous one) constitute only a first step towards providing a formal approach for determining harm in practice. Clearly more work needs to be done, ranging from investigating whether other elements need to be added to our framework; doing both empirical and philosophical studies on the concrete factors that determine the default utility, the weighting function, and the fairness parameters; and investigating complexity issues more carefully.

Moreover, there is obviously a close connection between harm and blame, in that one is usually blameworthy for an outcome only if that outcome constitutes a harm. Yet, unlike blame, harm does not always contain a moral dimension, since also natural events can cause harm. Therefore it is worthwhile to develop an account that integrates both harm and blame into a full theory of moral responsibility. We believe that this paper already provides a rich and useful framework, one that will be critical for dealing with the ethical and regulatory issues of deploying AI systems.

Acknowledgements

The authors would like to thank the IJCAI reviewers for their detailed comments and Jonathan Richens for a fruitful discussion of a preliminary version of this paper. Sander Beckers was supported by the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. Hana Chockler was supported in part by the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1) and the UKRI Strategic Priorities Fund to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1). Joe Halpern was supported in part by NSF grant IIS-1703846, ARO grant W911NF-22-1-0061, and MURI grant W911NF-19-1-0217.

References

[Beckers *et al.*, 2022a] S. Beckers, H. Chockler, and J. Y. Halpern. A causal analysis of harm. In *Proc. Advances in Nueral Information Processing Systems 35 (NeurIPS '22)*, 2022.

[Beckers *et al.*, 2022b] S. Beckers, H. Chockler, and J. Y. Halpern. A quantitative account of harm. Available at https://arxiv.org/abs/2209.15111, 2022.

[Beckers, 2021] S. Beckers. Causal sufficiency and actual causation. *Journal of Philosophical Logic*, 50:1341–1374, 2021.

[Bradley, 2012] B. Bradley. Doing away with harm. *Philosophy and Phenomenological Research*, 85:390–412, 2012.

[Carlson *et al.*, 2021] E. Carlson, J. Johansson, and O. Risberg. Causal accounts of harming. *Pacific Philosophical Quarterly*, 2021.

[European Commission, 2021] European Commission. 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, 2021. https://artificialintelligenceact.eu/the-act/; accessed Aug. 8, 2021.

[Gärdenfors and Sahlin, 1982] P. Gärdenfors and N. Sahlin. Unreliable probabilities, risk taking, and decision making. *Synthese*, 53:361–386, 1982.

[Gigerenzer, 2006] G. Gigerenzer. Out of the frying pan into the fire: behavioral reactions to terrorist attacks. *Risk Analysis*, 26(2):347–351, 2006.

[Gilboa and Schmeidler, 1989] I. Gilboa and D. Schmeidler. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

[Glymour and Wimberly, 2007] C. Glymour and F. Wimberly. Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein, editors, *Causation and Explanation*, pages 43–67. MIT Press, Cambridge, MA, 2007.

[Hall, 2007] N. Hall. Structural equations and causation. *Philosophical Studies*, 132:109–136, 2007.

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

[Halpern, 2015] J. Y. Halpern. A modification of the Halpern-Pearl definition of causality. In *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3022–3033, 2015.

[Halpern, 2016] J. Y. Halpern. *Actual Causality*. MIT Press, Cambridge, MA, 2016.

[Heidari *et al.*, 2021] H. Heidari, S. Barocas, J. M. Kleinberg, and K. Levy. On modeling human perceptions of allocation policies with uncertain outcomes. In *EC '21: The 22nd ACM Conference on Economics and Computation*, pages 589–609. ACM, 2021.

[Hertwig *et al.*, 2004] R. Hertwig, G. Barron, E.U. Weber, and I. Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539, 2004.

[Hitchcock, 2001] C. Hitchcock. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, XCVIII(6):273–299, 2001.

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

[Jenni and Loewenstein, 1997] K. Jenni and G. Loewenstein. Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, 14(3):235–257, 1997.

[Kahneman and Tversky, 1979] D. Kahneman and A. Tversky. Prospect theory, an analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[Niehans, 1948] J. Niehans. Zur preisbildung bei ungewissen erwartungen. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, 84(5):433–456, 1948.

[Norcross, 1998] A. Norcross. Great harms from small benefits grow: how death can be outweighed by headaches. *Analysis*, 58(2):152–158, 1998.

[Prelec, 1998] D. Prelec. The probability weighting function. *Econometrica*, 66(3):497–527, 1998.

[Quiggin, 1993] J. Quiggin. *Generalized Expected Utility Theory: The Rank-Dependent Expected Utility Model*. Kluwer, Boston, 1993.

[Richens *et al.*, 2022] J. G. Richens, R. Beard, and D. H. Thompson. Counterfactual harm. In *Proc. Advances in Nueral Information Processing Systems 35 (NeurIPS '22)*, 2022.

[Rigoli *et al.*, 2016] F. Rigoli, R. B. Rutledge, P. Dayan, and R. J. Dolan. The influence of contextual reward statistics on risk preference. *Neuroimage*, 128:74–84, 2016.

[Savage, 1951] L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951.

[Wald, 1950] A. Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

[Weslake, 2015] B. Weslake. A partial theory of actual causation. *British Journal for the Philosophy of Science*, 2015. To appear.

[Woodward, 2003] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, U.K., 2003.

[Zielonka and Tyszka, 2017] P. Zielonka and T. Tyszka, editors. *Large Risks with Low Probabilities*. IWA Publishing, 2017.