The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility

Joseph Y. Halpern
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
halpern@cs.cornell.edu

Max Kleiman-Weiner
Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
maxkw@mit.edu

Abstract

We provide formal definitions of *degree of blameworthiness* and *intention* relative to an *epistemic state* (a probability over causal models and a utility function on outcomes). These, together with a definition of actual causality, provide the key ingredients for moral responsibility judgments. We show that these definitions give insight into commonsense intuitions in a variety of puzzling cases from the literature.

1 Introduction

The need for judging *moral responsibility* arises both in ethics and in law. In an era of autonomous vehicles and, more generally, autonomous AI agents that interact with or on behalf of people, the issue has now become relevant to AI as well. We will clearly need to imbue AI agents with some means for evaluating moral responsibility. There is general agreement that a definition of moral responsibility will require integrating causality, some notion of *blameworthiness*, and *intention* (Cushman 2015; Malle, Guglielmo, and Monroe 2014; Weiner 1995). Previous work has provided formal accounts of causality (Halpern 2016); in this paper, we provide formal definitions of blameworthiness and intention in the same vein.

These notions are notoriously difficult to define carefully. The well-known *trolley problem* (Thomson 1985) illustrates some of them: Suppose that a runaway trolley is headed towards five people who will not be able to get out of the train's path in time. If the trolley continues, it will kill all five of them. An agent $\mathbf{ag}$ is near a switchboard, and while $\mathbf{ag}$ cannot stop the trolley, he can pull a lever which will divert the trolley to a side track. Unfortunately, there is a single man on the side track who will be killed if $\mathbf{ag}$ pulls the lever.

Most people agree that it is reasonable for $\mathbf{ag}$ to pull the lever. But now consider a variant of the trolley problem known as *loop* (Thomson 1985), where instead of the side track going off in a different direction altogether, it rejoins the main track before where the five people are tied up. Again, there is someone on the side track, but this time $\mathbf{ag}$ knows that hitting the man on the loop will stop the train before it hits the five people on the main track. How morally responsible is $\mathbf{ag}$ for the death of the man on the side track if he pulls the lever? Should the answer be different in the loop version of the problem? Pulling the lever in the loop condition is typically judged as less morally permissible than in the condition without a loop (Mikhail 2007).

The definitions given here take as their starting point the *structural-equations* framework used by Halpern and Pearl (2005) (HP from now on) in defining causality. This framework allows us to model counterfactual statements like "agent $\mathbf{ag}$ would have still performed action a even if outcome φ had not occurred". Evaluating such statements is the key to defining intention and moral responsibility, just as it is for defining actual causation. To evaluate such counterfactual statements requires a model of what $\mathbf{ag}$ would have done in that new situation. We assume that $\mathbf{ag}$ is an expected utility maximizer with beliefs expressed as probabilities over causal models and preferences expressed as a utility function over outcomes. This makes it possible to define $\mathbf{ag}$'s *degree of blameworthiness*; rather than $\mathbf{ag}$ either being blameworthy or not for an outcome, he is only blameworthy to some degree (a number in [0,1]). Using the same framework, we can also define *intention*. Roughly speaking, an agent who performs action a intends outcome φ if he would not have done a if a had no impact on whether φ occurred.

2 Structural equations and HP causality

The HP approach assumes that the world is described in terms of variables and their values. Some variables have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the random 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. We assume that there is a special endogenous variable A called the *action variable*; the possible values of A are the actions that the agent can choose among.[1]

[1]In a more general setting with multiple agents, each performing actions, we might have a variable $A_{\mathbf{ag}}$ for each agent $\mathbf{ag}$. We might also consider situations over time, where agents perform sequences of actions, determined by a strategy, rather than just a single action. Allowing this extra level of generality has no impact on the framework presented here.

For example, in the trolley problem, we can assume that A has two possible values: $A = 0$ if the lever was not pulled and $A = 1$ if it was. Which action is taken is determined by an exogenous variable. The two possible outcomes in the trolley problem are described by two other endogenous variables: O_1, which is 1 if the five people on the main track die, and 0 if they don't, and O_2, which is 1 if the person on the sidetrack dies, and 0 otherwise.

A *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, that 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*), and $\mathcal{F}$ is a set of *modifiable structural equations*, relating the values of the variables. Formally, $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. In the trolley problem as modeled above, there are two equations: $O_1 = 1 - A$ (the five people die if the agent does nothing) and $O_2 = A$ (the one person on the side track dies if the agent pulls the lever).

Just as HP do, we restrict attention to *acyclic* causal models, where there is a total ordering $\prec$ of the endogenous variables (the ones in $\mathcal{V}$) such that if $X \prec Y$, then X is independent of Y, that is, $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$. If $X \prec Y$, then the value of X may affect the value of Y, but the value of Y cannot affect the value of X. 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 $\mathcal{U}$, there is a unique solution for all the equations.

Given a causal model $M = (\mathcal{S}, \mathcal{F})$, a vector $\vec{X}$ of distinct variables in $\mathcal{V}$, and a vector $\vec{x}$ of values for the variables in $\vec{X}$, the causal model $M_{\vec{X} \leftarrow \vec{x}}$ is identical to M, except that the equation for the variables $\vec{X}$ in $\mathcal{F}$ is replaced by $\vec{X} = \vec{x}$. Intuitively, this is the causal model that results when the variables in $\vec{X}$ are set to $\vec{x}$ by some external action that affects only the variables in $\vec{X}$ (and overrides the effects of the causal equations).

To define causality carefully, it is useful to have a language to reason about causality. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, 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 $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A pair $(M, \vec{u})$ consisting of a causal model and a context is called a *causal setting*. A causal formula ψ is true or false in a causal setting. As in HP, $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the causal 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 $\vec{u}$ (i.e., the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models \varphi$.

In the full paper,[2] the HP definition of causality is given. The details are not necessary for understanding the remaining definitions. Indeed, the HP definition could be replaced by another definition of causality (e.g., (Glymour and Wimberly 2007; Hall 2007; Halpern and Pearl 2005; Hitchcock 2001; 2007; Woodward 2003; Wright 1988)).

3 Degree of blameworthiness

We now apply this formal language to study blameworthiness. For agent $\mathbf{ag}$ to be morally responsible for an outcome φ, he must be viewed as deserving of blame for φ. Among other things, for $\mathbf{ag}$ to be deserving of blame, he must have placed some likelihood (before acting) on the possibility that a would cause φ. If $\mathbf{ag}$ did not believe it was possible for a to cause φ, then in general we do not want to blame $\mathbf{ag}$ for φ (assuming that $\mathbf{ag}$'s beliefs are reasonable; see below).

In general, an agent has uncertainty regarding the structural equations that characterize a causal model and about the context. This uncertainty is characterized by a probability distribution Pr on a set $\mathcal{K}$ of causal settings.[3] Let $\mathcal{K}$ consist of causal settings $(M, \vec{u})$, and let Pr be a probability measure on $\mathcal{K}$. Pr should be thought of as describing the probability *before* the action is performed. For ease of exposition, we assume that all the models in $\mathcal{K}$ have the same signature (set of endogenous and exogenous variables). We assume that an agent's preferences are characterized by a utility function $\mathbf{u}$ on *worlds*, where a *world* is a complete assignment to the endogenous variables. Thus, an *epistemic state* for an agent $\mathbf{ag}$ consists of a tuple $\mathcal{E} = (\Pr, \mathcal{K}, \mathbf{u})$.

Given an epistemic state for an agent $\mathbf{ag}$, we can determine the extent to which $\mathbf{ag}$ performing action a affected, or made a difference, to an outcome φ (where φ can be an arbitrary Boolean combination of primitive events). Formally, we compare a to all other actions a' that $\mathbf{ag}$ could have performed. Let $[\![\varphi]\!]_{\mathcal{K}} = \{(M, \vec{u}) \in \mathcal{K} : (M, \vec{u}) \models \varphi\}$; that is, $[\![\varphi]\!]_{\mathcal{K}}$ consists of all causal settings in $\mathcal{K}$ where φ is true. Thus, $\Pr([\![[A = a]\varphi]\!]_{\mathcal{K}})$ is the probability that performing action a results in φ. Let

$$\delta_{a,a',\varphi} = \max(0, \Pr([\![[A = a]\varphi)]\!]_{\mathcal{K}} - \Pr([\![[A = a']\varphi]\!]_{\mathcal{K}})),$$

so that $\delta_{a,a',\varphi}$ measures how much more likely it is that φ will result from performing a than from performing a' (except that if performing a' is more likely to result in φ than performing a, we just take $\delta_{a,a',0}$ to be 0).

The difference $\delta_{a,a,\varphi'}$ is clearly an important component of measuring the blameworthiness of a relative to a'. But there is another component, which we can think of as the cost of doing a. Suppose that Bob could have given up his life to save Tom. Bob decided to do nothing, so Tom died. The difference between the probability of Tom dying if Bob does nothing and if Bob gives up his life is 1 (the maximum possible), but we do not typically blame Bob for not

[2] Available at www.cs.cornell.edu/home/halpern/moralresp.pdf.

[3] Chockler and Halpern (2004) also used such a probability to define a notion of *degree of blame*.

giving up his life. What this points out is that blame is also concerned with the *cost* of an action. The cost might be cognitive effort, time required to perform the action, emotional cost, or (as in the example above) death.

We assume that the cost is captured by some outcome variables. The cost of an action a is then the impact of performing a on these variables. We call the variables that we consider the *action-cost* variables. Intuitively, these are variables that talk about features of an action: Is the action difficult? Is it dangerous? Does it involve emotional upheaval? Which variables count as action-cost variables depends in part on the modeler, but we do assume that the action-cost variables satisfy some minimal properties. To make these properties precise, we need some definitions.

Given a causal setting $(M, \vec{u})$ and endogenous variables $\vec{X}$ in M, let $w_{M,\vec{X}\leftarrow\vec{x},\vec{u}}$ be the unique world determined by setting $\vec{X}$ to $\vec{x}$ in $(M, \vec{u})$. Thus, for each endogenous variable V, the value of V in world $w_{M,\vec{X}\leftarrow\vec{x},\vec{u}}$ is v iff $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}](V = v)$. Given an action a and outcome variables $\vec{O}$, let $\vec{o}_{M,A\leftarrow a,\vec{u}}$ be the value of $\vec{o}$ when we set A to a in the setting $(M, \vec{u})$; that is, $(M, \vec{u}) \models [A \leftarrow a](\vec{O} = \vec{o}_{M,A\leftarrow a,\vec{u}})$. Thus, $w_{M,\vec{O}\leftarrow\vec{o}_{M,A\leftarrow a,\vec{u}},\vec{u}}$ is the world that results when $\vec{O}$ is set to the value that it would have if action a is performed in causal setting $(M, \vec{u})$. To simplify the notation, we omit the M and $\vec{u}$ in the subscript of $\vec{o}$ (since they already appear in the subscript of w), and just write $w_{M,\vec{O}\leftarrow\vec{o}_{A\leftarrow a},\vec{u}}$. The world $w_{M,\vec{O}\leftarrow\vec{o}_{A\leftarrow a},\vec{u}}$ isolates the effects of a on the variables in $\vec{O}$.

With this background, we can state the properties that we expect the set $\vec{O}_c$ of action-cost variables to have:

- for all causal settings $(M, \vec{u})$ and all actions a, we have
$$\mathbf{u}(w_{M,\vec{O}_c\leftarrow\vec{o}_{A\leftarrow a},\vec{u}}) \leq 0$$
(so performing a is actually costly, as far as the variables in $\vec{O}_c$ go);
- for all causal settings $(M, \vec{u})$, all actions a, and all subsets $\vec{O}'$ of $\vec{O}_c$, we have
$$\mathbf{u}(w_{M,\vec{O}_c\leftarrow\vec{o}_{A\leftarrow a},\vec{u}}) \leq \mathbf{u}(w_{M,\vec{O}'\leftarrow\vec{o}'_{A\leftarrow a},\vec{u}})$$
(so all variables in $\vec{O}_c$ are costly—by not considering some of them, the cost is lowered).

Definition 3.1: The *cost* of action a (with respect to $\vec{O}_c$), denoted $c(a)$, is $|\mathbf{u}(w_{M,\vec{O}_c\leftarrow\vec{o}_{A\leftarrow a},\vec{u}})|$. ∎

If we think of φ as a bad outcome of performing a, then the blameworthiness of a for φ relative to a' is a combination of the likelihood to which the outcome could have been improved by performing a' and cost of a relative to the cost of a'. Thus, if $c(a) = c(a')$, the blameworthiness of a for φ relative to a' is just $\delta_{a,a',\varphi}$. But if performing a' is quite costly relative to performing a, this should lead to a decrease in blameworthiness. How much of a decrease is somewhat subjective. To capture this, choose $N > \max_{a'} c(a')$ (in general, we expect N to be situation-dependent). The size of N is a measure of how important we judge cost to be in determining blameworthiness; the larger N is, the less we weight the cost.

Definition 3.2: The *degree of blameworthiness of a for φ relative to a' (given c and N)*, denoted $db_N(a, a', \varphi)$, is $\delta_{a,a',\varphi}\frac{N-\max(c(a')-c(a),0)}{N}$. The degree of blameworthiness of a for φ, denoted $db_N(a, \varphi)$ is $\max_{a'} db_N(a, a', \varphi)$. ∎

Intuitively, we view the cost as a mitigating factor when computing the degree of blameworthiness of a for φ. We can think of $\frac{N-\max(c(a')-c(a),0)}{N}$ as the mitigation factor. No mitigation is needed when comparing a to a' if the cost of a is greater than that of a'. And, indeed, because of the $\max(c(a) - c(a'), 0)$ term, if $c(a) \geq c(a')$ then the mitigation factor is 1, and $db(a, a', \varphi) = \delta_{a,a',\varphi}$. In general, $\frac{N+\max(c(a)-c(a'),0)}{N} \leq 1$. Moreover, $\lim_{N\to\infty} db_N(a, a', \varphi) = \delta_{a,act',\varphi}$. Thus, for large values of N, we essentially ignore the costliness of the act. On the other hand, if N and $c(a')$ are both close to $\max_{a''} c(a'')$ and $c(a) = 0$, then $db_N(a, a', \varphi)$ is close to 0. Thus, in the example with Bob and Tom above, if we take costs seriously, then we would not find Bob particularly blameworthy for Tom's death if the only way to save Tom is for Bob to give up his own life.

The need to consider alternatives when determining blameworthiness is certainly not new, as can be seen from the essays in (Widerker and McKenna 2006). What seems to be new here is the emphasis on blameworthiness with respect an outcome and taking cost into account. The following example shows the impact of the former point.

Example 3.3: Suppose that agent $\mathbf{ag}$ is faced with the following dilemma: if $\mathbf{ag}$ doesn't pull the lever, six anonymous people die; if $\mathbf{ag}$ does pull the lever, the first five people will still die, but the sixth will be killed with only probability 0.2. If $\mathbf{ag}$ does not pull the lever, $\mathbf{ag}$ is not blameworthy for the five deaths (no matter what he did, the five people would have died), but has some degree of blameworthiness for the sixth. The point here is that just the existence of a better action a' is not enough. To affect $\mathbf{ag}$'s blameworthiness for outcome φ, action a' must be better in a way that affects φ. ∎

Defining the degree of blameworthiness of an action for a particular outcome, as done here, seems to be consistent with the legal view. A prosecutor considering what to charge a defendant with is typically considering which outcomes that defendant is blameworthy for.

Blameworthiness is defined relative to a probability distribution. We do not necessarily want to use the agent's subjective probability. For example, suppose that the agent had several bottles of beer, goes for a drive, and runs over a pedestrian. The agent may well have believed that the probability that his driving would cause an accident was low, but we clearly don't want to use his subjective probability that he will cause an accident in determining blameworthiness. Similarly, suppose that a doctor honestly believes that a certain medication will have no harmful side effects for a patient. One of his patients who had a heart condition takes the medication and dies as a result. If the literature distributed to the doctor included specific warning about dire side-effects for patients with this heart condition but the doctor was lazy and didn't read it, again, it does not seem reasonable to use the doctor's probability distribution. Rather, we want to use the probability distribution that he should have had, had he read the relevant literature. We can use whatever probability distribution we consider most appropriate in the definition.

In using the term "blameworthiness", we have implicitly been thinking of φ as a bad outcome. If φ is a good outcome, it seems more reasonable to use the term "praiseworthiness". However, defining praiseworthiness raises some significant new issues. We mention a few of them here:

- Suppose that all actions are costless, Bob does nothing and, as a result, Tom lives. Bob could have shot Tom, so according to the definition Bob's degree of blameworthiness for Tom living is 1. Since living is a good outcome, we may want to talk about praiseworthiness rather than blameworthiness, but it still seems strange to praise Tom for doing the obvious thing. This suggests that for praiseworthiness, we should compare the action to the "standard" or "expected" thing to do. To deal with this, we assume that there is a *default action* a_0, which we can typically think of as "doing nothing" (as in the example above), but does not have to be. Similarly, we typically assume that the default action has low cost, but we do not require this. The praiseworthiness of an act is typically compared just to the default action, rather than to all actions. Thus, we consider just $\delta_{a,a_0,\varphi}$, not $\delta_{a,a',\varphi}$ for arbitrary a'.
- It does not seem that there should be a lessening of praise if the cost of a is even lower than that of the default. On the other hand, it seems that there should be an increase in praise the more costly a is. For example, we view an action as particularly praiseworthy if someone is risking his life to perform it. This suggests that the degree of praiseworthiness of a should be $\delta_{a,a_0,\varphi}$ if $c(a) \le c(a_0)$, and $\delta_{a,a_0,\varphi}\frac{M-c(a_0)+c(a)}{M}$ if $c(a) > c(a_0)$. But this has the problem that the degree of praiseworthiness might be greater than 1. If the agent put a lot of effort into a (i.e., $c(a) - c(a_0)$ is large) because his main focus was some other outcome $\varphi' \neq \varphi$, and there is another action a' that would achieve φ at much lower cost, then it seems unreasonable to give the agent quite so much praise for his efforts in achieving φ.
- We typically do not praise someone for an outcome that was not intended (although we might well find someone blameworthy for an unintended outcome).

In the full paper, we give a definition of praiseworthiness that takes these concerns into account.

The focus of these definitions has been on the blame (or praise) due to a single individual. Things get more complicated once we consider groups. Consider how these definitions play out in the context of the *Tragedy of the Commons* (Hardin 1968), where there are many agents, each of which can perform an action (like fishing, or letting his sheep graze on the commons) which increases his individual utility, but if all agents perform the action, they all ultimately suffer (fish stocks are depleted; the commons is overgrazed).

Example 3.4: Consider a collective of fishermen. Suppose that if more than a couple of agents fail to limit their fishing, the fish stocks will collapse and there will be no fishing allowed the following year. The fisherman in fact all do fish, so the fish stocks collapse.

Each agent is clearly part of the cause of the outcome. To determine a single agent's degree of blameworthiness, we must consider that agent's uncertainty about how many of the other fisherman will limit their fishing. If the agent believes (perhaps justifiably) that, with high probability, very few of them will limit their fishing, then his blameworthiness will be quite low. As we would expect, under minimal assumptions about the probability measure $\Pr$, the more fisherman there are and the larger the gap between the expected number of fish taken and the number that will result in overfishing limitations, the lower the degree of blameworthiness. Moreover, a fisherman who catches less fish has less blameworthiness. In all these cases, it is less likely that changing his action will lead to a change in outcome. ∎

The way that blameworthiness is assigned to an individual fisherman in Example 3.4 essentially takes the actions of all the other fisherman as given. But it is somewhat disconcerting that if each of N fisherman justifiably believed that all the other fisherman would overfish, then each might have degree of blameworthiness significantly less than the $1/N$ that we might intuitively give them if they all caught roughly the same number of fish.

One way to deal with this is to consider the degree of blame we would assign to all the fisherman, viewed as a collective (i.e., as a single agent). The collective can clearly perform a different action that would lead to the desired outcome. Thus, viewed as a collective, the fishermen have degree of blameworthiness close to 1.

How should we allocate this "group moral blameworthiness" to the individual agents? We believe that Chockler and Halpern's (2004) notion of responsibility and blame can be helpful in this regard, because they are intended to measure how responsibility and blame are diffused in a group. It seems that when ascribing moral responsibility in group settings, people consider both an agent as an individual and as a member of a group. Further research is needed to clarify this issue.

4 Intention

The definition of degree of blameworthiness does not take intention into account. In the trolley problem, an agent who pulls the lever so that only one person dies is fully blameworthy for that death. However, it is clear that the agent's intent was to save five people, not kill one; the death was an unintended side-effect. Usually, agents are not held responsible for accidents and the moral permissibility of an action does not take into account unintended side-effects.

Two types of intention have been considered in the literature (see, e.g., (Cohen and Levesque 1990)): (1) whether agent $\mathbf{ag}$ intended to perform action a (perhaps it was an accident) and (2) did $\mathbf{ag}$ (when performing a) intend outcome φ (perhaps φ was an unintended side-effect of a, which was actually performed to bring about outcome o'). Intuitively, an agent intended to perform a (i.e., a was not accidental) if his expected utility from a is at least as high as his expected utility from other actions. The following definition formalizes this intuition.

Definition 4.1: Action a is *intended in* $(M,\vec{u})$ *given epistemic state* $\mathcal{E} = (\Pr, \mathcal{K}, \mathbf{u})$ if $(M,\vec{u}) \models A = a$ (a was actually performed in causal setting $(M,\vec{u})$), $|\mathcal{R}(A)| \geq 2$ (a is not the only possible action), and for all $a' \in \mathcal{R}(A)$,

$$\sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})(\mathbf{u}(w_{M,A\leftarrow a,\vec{u}}) - \mathbf{u}(w_{M,A\leftarrow a',\vec{u}})) \geq 0.$$

■

The assumption that $|\mathcal{R}(A)| \geq 2$ captures the intuition that we do not say that a was intended if a was the only action that the agent could perform. We would not say that someone who is an epileptic intended to have a seizure, since they could not have done otherwise. What about someone who performed an action because there was a gun held to his head? In this case, it depends on how we model the set A of possible actions. If we take the only feasible action to be the act a that was performed (so we view the agent as having no real choice in the matter), then the action was not intentional. But if we allow for the possibility of the agent choosing whether or not to sacrifice his life, then we would view whatever was imposed as intended.

The intuition for the agent intending outcome $\vec{O} = \vec{o}$ is that, had a been unable to affect $\vec{O}$, $\mathbf{ag}$ would not have performed a. But this is not quite right for several reasons, as the following examples show.

Example 4.2: Suppose that a patient has malignant lung cancer. The only thing that the doctor believes that he can do to save the patient is to remove part of the lung. But this operation is dangerous and may lead to the patient's death. In fact, the patient does die. Certainly the doctor's operation is the cause of death, and the doctor intended to perform the operation. However, if the variable O represents the possible outcomes of the operation, with $O = 0$ denoting that the patient dies and $O = 1$ denoting that the patient is cured, while the doctor intended to affect the variable O, he certainly did not intend the actual outcome $O = 0$. ■

Example 4.3: Suppose that Louis plants a bomb at a table where his cousin Rufus, who is standing in the way of him getting an inheritance, is going to have lunch with Sibella. Louis get 100 units of utility if Rufus dies, 0 if he doesn't die, and -200 units if he goes to jail. His total utility is the sum of the utilities of the relevant outcomes. He would not have planted the bomb if doing so would not have affected whether Rufus dies. On the other hand, Louis would still have planted the bomb even if doing so had no impact on Sibella. Thus, we can conclude that Louis intended to kill Rufus but did not intend to kill Sibella.

Now suppose that Louis has a different utility function, and prefers that both Rufus and Sibella die. Specifically, Louis get 50 units of utility if Louis dies and 50 units of utility if Sibella dies. Again, he gets -200 if he goes to jail, and his total utility is the sum of the utilities of the relevant outcomes. With these utilities, intuitively, Louis intends both Rufus and Sibella to die. Even if he knew that planting the bomb had no impact on whether Rufus lives (perhaps because Rufus will die of a heart attack, or because Rufus is wearing a bomb-proof vest), Louis would still plant the bomb (since he would get significant utility from Sibella dying). Similarly, he would plant the bomb even if it had no impact on Sibella. Thus, according to the naive definition, Louis did not intend to kill either Rufus or Sibella. ■

Our definition will deal with both of these problems. We actually give our definition of intent in two steps. First, we define what it means for agent $\mathbf{ag}$ to intend to affect the variables in $\vec{O}$ by performing action a.

To understand the way we formalize this intuition better, suppose first that a is deterministic. Then $w_{M,A\leftarrow a,\vec{u}}$ is the world that results when action a is performed in the causal setting $(M,\vec{u})$ and $w_{M,(A\leftarrow a',\vec{O}\leftarrow\vec{o}_{A\leftarrow a}),\vec{u}}$ is the world that results when act a' is performed, except that the variables in $\vec{O}$ are set to the values that they would have had if a were performed rather than a'. If $\mathbf{u}(w_{M,A\leftarrow a,\vec{u}}) < \mathbf{u}(w_{M,(A\leftarrow a',\vec{O}\leftarrow\vec{o}_{A\leftarrow a}),\vec{u}})$, that means that if the variables in $\vec{O}$ are fixed to have the values they would have if a were performed, then the agent would prefer to do a' rather than a. Similarly, $\mathbf{u}(w_{M,(\vec{O}\leftarrow\vec{o}_{A\leftarrow a}),\vec{u}}) > \mathbf{u}(w_{M,(\vec{O}\leftarrow\vec{o}_{A\leftarrow a'}),\vec{u}})$ says that the agent prefers how a affects the variables in $\vec{O}$ to how a' affects these variables. Intuitively, it will be these two conditions that suggest that the agent intends to affect the values of the variables in $\vec{O}$ by performing a; once their values are set, the agent would prefer a' to a.

The actual definition of the agent intending to affect the variables in $\vec{O}$ is slightly more complicated than this in several respects. First, if the outcome of a is probabilistic, we need to consider each of the possible outcomes of performing a and weight them by their probability of occurrence. To do this, for each causal setting $(M,\vec{u})$ that the agent considers possible, we consider the effect of performing a in $(M,\vec{u})$ and weight it by the probability that the agent assigns to $(M,\vec{u})$. Second, we must deal with the situation discussed in Example 4.3 where Louis intends both Rufus and Sibella to die. Let D_R and D_S be variables describing whether Rufus and Sibella, respectively, die. While Louis certainly intends to affect D_R, he will not plant the bomb only if both Rufus and Sibella die without the bomb (i.e., only if both D_R and D_S are set to 0). Thus, to show that the agent intends to affect the variable D_R, we must consider a superset of D_R (namely, $\{D_R, D_S\}$). Third, we need a minimality condition. If Louis intended to kill only Rufus, and Sibella dying was an unfortunate byproduct, we do not want to say that he intended to affect $\{D_R, D_S\}$, although he would not have planted the bomb if both D_R and D_S were set to 0. There is a final subtlety: when considering whether $\mathbf{ag}$ intended to perform a, what alternative actions should we compare a to? The obvious answer is "all other actions in A". Indeed, this is exactly what was done by Kleiman-Weiner et al. (2016) (who use an approach otherwise similar in spirit to the one proposed here, but based on influence diagrams rather than causal models). We instead generalize to allow a *reference set* $REF(a)$ of actions which, as the notation suggests, can depend on a, and compare a only to actions in $REF(a)$.

As we shall see, we need this generalization to avoid some problems. We discuss $REF(a)$ in more detail below.

Definition 4.4: An agent $\mathbf{ag}$ *intends to affect* $\vec{O}$ *by doing action* a *given epistemic state* $\mathcal{E} = (\Pr, \mathcal{K}, \mathbf{u})$ *and reference set* $REF(a) \subseteq A$ if and only if there exists a superset $\vec{O}'$ of $\vec{O}$ such that (a) $\sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})\mathbf{u}(w_{M,A\leftarrow a,\vec{u}}) < \max_{a'\in REF(a)} \sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})\mathbf{u}(w_{M,(A\leftarrow a',\vec{O}'\leftarrow\vec{o}'_{A\leftarrow a}),\vec{u}})$, and (b) $\vec{O}'$ is minimal; that is, for all strict subsets $\vec{O}^*$ of $\vec{O}'$, we have $\sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})\mathbf{u}(w_{M,A\leftarrow a,\vec{u}}) \geq \max_{a'\in REF(a)} \sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})\mathbf{u}(w_{M,(A\leftarrow a',\vec{O}^*\leftarrow\vec{o}'_{A\leftarrow a}),\vec{u}})$. ∎

Part (a) says that if the variables in $\vec{O}'$ were given the value they would get if a were performed, then some act a' becomes better than a. Part (b) says that $\vec{O}'$ is the minimal set of outcomes with this property.

What should $REF(a)$ be? If there are only two actions in A, then $REF(a)$ should clearly be the other act. A natural generalization is to take $REF(a) = A - \{a\}$. The following example shows why this will not always work.

Example 4.5: Suppose that Daniel is a philanthropist who is choosing a program to support. He wants to choose among programs that support schools and health clinics and he cares about schools and health clinics equally. If he chooses program 1 he will support 5 schools and 4 clinics. If he chooses program 2 he will support 2 schools and 5 clinics. Assume he gets 1 unit of utility for each school or clinic supported. The total utility of a program is the sum of the utility he gets for the schools and clinics minus 1 for the overhead of both programs. We can think of this overhead as the cost of implementing a program versus not implementing any of the programs. By default he can also do nothing which has utility 0 since it avoids any overhead and doesn't support any schools or clinics.

Clearly his overall utility is maximized by choosing program 1. Intuitively, by doing so, he intends to affect both the schools and clinics. Indeed, if he could support 5 schools and 4 clinics without the overhead of implementing a program, he would do that. However, if we consider all alternatives, then the minimality condition fails. If he could support 5 schools he would switch to program 2, but if he could support 4 clinics he would still choose program 1. This gives the problematic result that Daniel intends to support only schools. The problem disappears if we take the reference set to consist of just the default action: doing nothing. Then we get the desired result that Daniel intends to both support the 5 schools *and* the 4 clinics. ∎

It might seem that by allowing $REF(a)$ to be a parameter of the definition we have allowed too much flexibility, leaving room for rather *ad hoc* choices. There are principled reasons for restricting $REF(a)$ and not taking it to be all acts other than a in general. For one thing, the set of acts can be large, so there are computational reasons to consider fewer acts. If there is a natural default action (as in Example 4.5), this is often a natural choice for $REF(a)$. However, we cannot take $REF(a)$ to be just the default action if a is itself the default action (since then the first part of Definition 4.4 would not hold for any set $\vec{O}$ of outcomes). We discuss the choice of reference set in more detail in the full paper.

Given the variables that the agent intends to affect, we can determine the outcomes that the agent intends.

Definition 4.6: Agent $\mathbf{ag}$ *intends to bring about* $\vec{O} = \vec{o}$ *in* $(M, \vec{u})$ *by doing action* a *given epistemic state* $\mathcal{E} = (\Pr, \mathcal{K}, \mathbf{u})$ *and reference set* $REF(a)$ if and only if (a) $\mathbf{ag}$ intended to affect $\vec{O}$ by doing action a in epistemic state $\mathcal{E}$ given $REF(a)$, (b) there exists a setting $(M', \vec{u}')$ such that $\Pr(M', \vec{u}') > 0$ and $(M', \vec{u}') \models [A \leftarrow a](\vec{O} = \vec{o})$, (c) for all values $\vec{o}^*$ of $\vec{O}$ such that there is a setting $(M', \vec{u}')$ with $\Pr(M', \vec{u}') > 0$ and $(M', \vec{u}') \models [A \leftarrow a](\vec{O} = \vec{o}^*)$, we have $\sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})\mathbf{u}(w_{M,\vec{O}\leftarrow\vec{o},\vec{u}}) \geq \sum_{(M,\vec{u})\in\mathcal{K}} \Pr(M,\vec{u})\mathbf{u}(w_{M,\vec{O}\leftarrow\vec{o}^*,\vec{u}})$. ∎

Part (b) of this definition says that $\mathbf{ag}$ considers $\vec{O} = \vec{o}$ a possible outcome of performing a (even if it doesn't happen in the actual situation $(M, \vec{u})$). Part (c) says that, among all possible values of $\vec{O}$ that $\mathbf{ag}$ considers possible, $\vec{o}$ gives the highest expected utility.

This definition seems to capture natural language usage of the word "intends" if a is deterministic or close to deterministic. But if a is probabilistic, then we typically use the word "hopes" rather than intends. It seems strange to say that $\mathbf{ag}$ intends to win \$5,000,000 when he buys a lottery ticket (if \$5,000,000 is the highest payoff); it seems more reasonable to say that he hopes to win \$5,000,000. Similarly, if a doctor performs an operation on a patient who has cancer that he believes has only a 30% chance of complete remission, it seems strange to say that he "intends" to cure the patient, although he certainly hopes to cure the patient by performing the operation. In addition, once we think in terms of "hopes" rather than "intends", it may make sense to consider not just the best outcome, but all reasonably good outcomes. For example, the agent who buys a lottery ticket might be happy to win any prize that gives over \$10,000, and the doctor might also be happy if the patient gets a remission for 10 years.

5 Complexity considerations

Since $w_{M,\vec{X}=\vec{x},\vec{u}}$ can be computed in time polynomial in the size of M, it easily follows that, given an epistemic state $\mathcal{E} = (\Pr, \mathcal{K}, \mathbf{u})$, $\delta_{a,a',\varphi}$ can be computed in time polynomial in $|\mathcal{K}|$. Thus, the degree of blameworthiness of an action a for outcome φ can be computed in time polynomial in $|\mathcal{K}|$ and the cardinality of the range of A. Similarly, whether a is (un)intended in $(M, \vec{u})$ given $\mathcal{E}$ can be computed in time polynomial in $|\mathcal{K}|$ and the cardinality of the range of A.

The complexity of determining whether $A = a$ is part of a cause of φ in $(M, \vec{u})$ is Σ_2^p-complete, that is, it is at the second level of the polynomial hierarchy (Sipser 2012). This complexity is due to the "there exists–for all" structure of the problem (there exist sets $\vec{X}$ and $\vec{W}$ of variables such that for all strict subsets of $\vec{X}$ ...). The problem of determining if $\mathbf{ag}$ intended to bring about $\vec{O} = \vec{o}$ has a similar "there exists–for all" structure; we conjecture that it is also Σ_2^P-complete. While this makes the general problem quite intractable, in practice, things may not be so bad. Recall that $\mathbf{ag}$ intends to bring about $\vec{O} = \vec{o}$ if there exists a superset $\vec{O}'$ of $\vec{O}$ (intuitively, all the outcomes that $\mathbf{ag}$ intends to affect) with the appropriate properties. In practice, there are not that many outcomes that determine an agent's utility. If we assume that $|\vec{O}'| \leq k$ for some fixed k, then the problem becomes polynomial in the number of variables in the model and the number of actions; moreover, the polynomial has degree k. In practical applications, it seems reasonable to assume that there exists a (relatively small) k, making the problem tractable.

6 Related work

Amazon lists over 50 books in Philosophy, Law, and Psychology with the term "Moral Responsibility" in the title, all of which address the types of issues discussed in this paper. There are dozens of other books on intention. Moreover, there are AI systems that try to build in notions of moral responsibility (see, e.g., (Dehghani et al. 2008; Mao and Gratch 2012; Scheutz, Malle, and Briggs 2015)). Nevertheless, there has been surprisingly little work on providing a formal definition of moral responsibility of the type discussed here. We now briefly discuss some of the work most relevant to this project, without attempting to do a comprehensive survey of the relevant literature. We go into more detail in the full paper.

As mentioned in the introduction, Chockler and Halpern (2004) define a notion of responsibility that tries to capture the diffusion of responsibility when multiple agents contribute to an outcome but no agent is a *but-for* cause of that outcome, that is, no agent can change the outcome by just switching to a different action. For example, the degree of responsibility of a voter for the outcome $1/(1+k)$, where k is the number of changes needed to make the vote critical (i.e., a but-for cause). Chockler and Halpern also use epistemic states (although without the utility component): they define a notion of *degree of blame* given an epistemic state $\mathcal{E}$, which is the expected degree of responsibility with respect $\mathcal{E}$. These notions of blame and responsibility do not take utility into account, nor do they consider potential alternative actions or intention.

Cohen and Levesque (1990) initiated a great deal of work in AI on reasoning about an agent goals and intentions. They define a modal logic that includes operators for goals and beliefs, and define formulas $INTEND_1(\mathbf{ag}, a)$—agent $\mathbf{ag}$ intends action a—and $INTEND_2(\mathbf{ag}, p)$—agent $\mathbf{ag}$ intends goal p. $INTEND_1(\mathbf{ag}, a)$ is the analogue of Definition 4.1, while $INTEND_2(\mathbf{ag}, p)$ is the analogue of Definition 4.6; a goal for Cohen and Levesque is essentially an outcome. Roughly speaking, agent $\mathbf{ag}$ intends to bring about φ if $\mathbf{ag}$ has a plan that he believes will bring about φ (belief is captured using a modal operator, but we can think of it as corresponding to "with high probability"), is justified in believing so, and did not intend to bring out $\neg\varphi$ prior to executing the plan. Their framework does not allow us to model an agent's utility, nor can they express counterfactuals.

Kleiman-Weiner et al. (2015) give a definition of intention in the spirit of that given here. Specifically, it involves counterfactual reasoning and takes expected utility into account. It gets the same results for intention in the standard examples as the definition given here, for essentially the same reasons. However, rather than using causal models, they use influence diagrams. The agent's intention when performing a is then a minimal set of nodes whose fixation in the influence diagram would result in some action a' having expected utility at least as high as that of a. They use their definition of intention along with a model for the inference of the agent's epistemic state and utility function to model human judgments in many moral dilemmas.

Perhaps closest to this paper is the work of Braham and van Hees (2012). They say that an agent $\mathbf{ag}$ is morally responsible for an outcome φ if (a) his action a was a cause of φ, (b) $\mathbf{ag}$ intended to perform a, and (c) $\mathbf{ag}$ had no eligible action a' with a higher *avoidance potential*. They do not give a formal definition of intentionality, instead assuming that in situations of interest to them, it is always satisfied. Roughly speaking, the avoidance potential of a with respect to φ is the probability that a does not result in φ. Thus, the notion of the avoidance potential of an action a for φ being greater than that of a' is somewhat related to having $\delta_{a,a',\varphi} > 0$, although the technical details are quite different.

In the full paper, other recent work (Berreby, Bourgne, and Ganascia 2015; Gaudou et al. 2013; Lorini, Longin, and Mayor 2014; Poel, Royakkers, and Zwart 2015; Vallentyne 2008) is also discussed.

7 Conclusion

People's ascriptions of moral responsibility seem to involve three components that we have called here causality, degree of blameworthiness, and intention. We have given formal definitions of the latter two. Because it is not clear exactly how intention and degree of blame should be combined, we have left them here as separate components of moral responsibility.[4] Considerations of moral responsibility have become more pressing as we develop driverless cars, robots that will help in nursing homes, and software assistants. The framework presented here should help in that regard.

Our definitions of blameworthiness and intention were given relative to an epistemic state that included a probability measure and a utility function. This means that actions could be compared in terms of expected utility; this played a key role in the definitions. But there are some obvious concerns: first, agents do not "have" complete probability measures and utility functions. Constructing them requires nontrivial computational effort. Things get even worse if we try to consider what the probability and utility of a "reasonable" person should be; there will clearly be far from complete agreement about what these should be. And even if we could agree on a probability and utility, it is not clear that

[4] In his influential work, Weiner (1995) distinguishes causality, responsibility, and blame. Responsibility corresponds roughly to what we have called blameworthiness, while blame roughly corresponds to blameworthiness together with intention.

maximizing expected utility is the "right" decision rule. One direction for further research is to consider how the definitions given here play out if we use, for example, a set of probability measures rather than a single one, and/or use decision rules other than expected utility maximization (e.g., maximin). Another issue that deserves further investigation is responsibility as a member of the group vs. responsibility as an individual (see the brief discussion after Example 3.4).

Acknowledgments: Halpern was supported in part by NSF grants IIS-1703846 and IIS-1718108, AFOSR grant FA9550-12-1-0040, ARO grant W911NF-17-1-0592, and a grant from the Open Philanthropy project. Kleiman-Weiner was supported in part by a Hertz Foundation Fellowship. We thank Sander Beckers, Tobias Gerstenberg and Jonathan Phillips for interesting discussions and many useful comments on the topics of this paper.

References

Berreby, F.; Bourgne, G.; and Ganascia, J.-G. 2015. Modelling moral reasoning and ethical responsibility with logic programming. In *Proc. 20th Int. Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 2015)*, 532–548.

Braham, M., and van Hees, M. 2012. An anatomy of moral responsibility. *Mind* 121(483):601–634.

Chockler, H., and Halpern, J. Y. 2004. Responsibility and blame: A structural-model approach. *Journal of A.I. Research* 20:93–115.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42(2–3):213–261.

Cushman, F. 2015. Deconstructing intent to reconstruct morality. *Current Opinion in Psychology* 6:97–103.

Dehghani, M.; Tomai, E.; Forbus, K.; and Klenk, M. 2008. An integrated reasoning approach to moral decision-making. In *Proc. Twenty-Third National Conference on Artificial Intelligence (AAAI '08)*, 1280–1286.

Gaudou, B.; Lorini, E.; ; and Mayor, E. 2013. Moral guilt: an agent-based model analysis. In *9th Conference of the European Social Simulation Association (ESSA 2013)*, 95–106.

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

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

Halpern, J. Y., and 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. 2016. *Actual Causality*. Cambridge, MA: MIT Press.

Hardin, G. 1968. The tragedy of the commons. *Science* 162:1243–1248.

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

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

Kleiman-Weiner, M.; Gerstenberg, T.; Levine, S.; and Tenenbaum, J. B. 2015. Inference of intention and permissibility in moral decision making. In *Proc. 37th Annual Conference of the Cognitive Science Society (CogSci 2015)*.

Kleiman-Weiner, M.; Gerstenberg, T.; Levin, S.; and Tenenbaum, J. B. 2016. Inference of intention and permissibility in moral decision making. In preparation.

Lorini, E.; Longin, D.; and Mayor, E. 2014. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation* 24(6):1313–1339.

Malle, B. F.; Guglielmo, S.; and Monroe, A. E. 2014. A theory of blame. *Psychological Inquiry* 25(2):147–186.

Mao, W., and Gratch, J. 2012. Modeling social causaility and responsibility judgment in multi-agent interactions. *Journal of A.I. Research* 44:223–273.

Mikhail, J. 2007. Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences* 11(4):143–152.

Poel, I. v. d.; Royakkers, L.; and Zwart, S. D. 2015. *Moral Responsibility and the Problem of Many Hands*. New York: Routledge.

Scheutz, M.; Malle, B.; and Briggs, G. 2015. Towards morally sensitive action selection for autonomous social robots. In *Proc. International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Sipser, M. 2012. *Introduction to Theory of Computation*. Boston: Thomson Course Technology, third edition.

Thomson, J. J. 1985. The trolley problem. *Yale Law Journal* 94:1395–1415.

Vallentyne, P. 2008. Brute luck and responsibility. *Politics, Philosophy and Economics* 7:57–80.

Weiner, B. 1995. *Judgments of Responsibility*. New York: The Guildford Press.

Widerker, D., and McKenna, M. 2006. *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. Ashgate.

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

Wright, R. W. 1988. Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review* 73:1001–1077.