Responsibility and Blame: A Structural-Model Approach

Hana Chockler
School of Engineering and Computer Science
Hebrew University
Jerusalem 91904, Israel.
Email: hanac@cs.huji.ac.il

Joseph Y. Halpern*
Department of Computer Science
Cornell University
Ithaca, NY 14853, U.S.A.
Email: halpern@cs.cornell.edu

Abstract

Causality is typically treated an all-or-nothing concept; either *A* is a cause of *B* or it is not. We extend the definition of causality introduced by Halpern and Pearl 2001a to take into account the *degree of responsibility of A* for B. For example, if someone wins an election 11-0, then each person who votes for him is less responsible for the victory than if he had won 6-5. We then define a notion of *degree of blame,* which takes into account an agent's epistemic state. Roughly speaking, the degree of blame of *A* for *D* is the expected degree of responsibility of *A* for *B,* taken over the epistemic state of an agent.

1 Introduction

There have been many attempts to define *causality* going back to Hume 1739, and continuing to the present (see, for example, [Collins *et al,* 2003; Pearl, 2000] for some recent work). While many definitions of causality have been proposed, all of them treat causality is treated as an all-or-nothing concept. That is, *A* is either a cause of B or it is not. As a consequence, thinking only in terms of causality docs not at times allow us to make distinctions that we may want to make. For example, suppose that Mr. B wins an election against Mr. G by a vote of 11-0. Each of the people who voted for Mr. B is a cause of him winning. However, it seems that their degree of responsibility should not be as great as in the case when Mr. B wins 6-5.

In this paper, we present a definition of responsibility that takes this distinction into account. The definition is an extension of a definition of causality introduced by Halpern and Pearl 2001a. Like many other definitions of causality going back to Hume 1739, this definition is based on counterfactual dependence. Roughly speaking, *A* is a cause of *B* if, had *A* not happened (this is the counterfactual condition, since *A* did in fact happen) then *B* would not have happened. As is well known, this naive definition does not capture all the subtleties involved with causality. In the case of the 6-5 vote, it is clear that, according to this definition, each of the voters for Mr. B is a cause of him winning, since if they had voted against Mr. B, he would have lost. On the other hand, in the case of the 11-0 vote, there are no causes according to the naive counterfactual definition. A change of one vote does not makes no difference. Indeed, in this case, we do say in natural language that the cause is somewhat "diffuse".

While in this case the standard counterfactual definition may not seem quite so problematic, the following example (taken from [Hall, 2003]) shows that things can be even more subtle. Suppose that Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had Suzy not thrown. Thus, according to the naive counterfactual definition, Suzy's throw is not a cause of the bottle shattering. This certainly seems counter to intuition.

Both problems are dealt with the same way in [Halpern and Pearl, 2001a]. Roughly speaking, the idea is that *A* is a cause of *B* if *B* counterfactually depends on *C under some contingency.* For example, voter 1 is a cause of Mr. B winning even if the vote is 11-0 because, under the contingency that 5 of the other voters had voted for Mr. G instead, voter 1 's vote would have become critical; if he had then changed his vote, Mr. B would not have won. Similarly, Suzy's throw is the cause of the bottle shattering because the bottle shattering counterfactually depends on Suzy's throw, under the contingency that Billy doesn't throw. (There are further subtleties in the definition that guarantee that, if things are modeled appropriately, Billy's throw is not a cause. These are discussed in Section 2.)

It is precisely this consideration of contingencies that lets us define degree of responsibility. We take the degree of responsibility of *A* for *B* to be *1/(N* + 1), where *N* is the minimal number of changes that have to be made to obtain a contingency where *B* counterfactually depends on *A.* (If *A* is not a cause of B, then the degree of responsibility is 0.) In particular, this means that in the case of the 11-0 vote, the degree of responsibility of any voter for the victory is 1/6, since 5 changes have to be made before a vote is critical. If the vote were 1001-0, the degree of responsibility of any voter would be 1/501. On the other hand, if the vote is 5-4, then the degree of responsibility of each voter for Mr. B for Mr. B's victory is 1; each voter is critical. As we would expect, those

* Supported in part by NSF under grant CTC-0208535 and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by ONR under grant N00014-01-1-0795.

voters who voted for Mr. G have degree of responsibility 0 for Mr. B's victory, since they are not causes of the victory. Finally, in the case of Suzy and Billy, even though Suzy is the only cause of the bottle shattering, Suzy's degree of responsibility is 1/2, while Billy's is 0. Thus, the degree of responsibility measures to some extent whether or not there are other potential causes.

When determining responsibility, it is assumed that everything relevant about the facts of the world and how the world works (which we characterize in terms of what are called *structural equations)* is known. For example, when saying that voter 1 has degree of responsibility 1/6 for Mr. B's win when the vote is 11-0, we assume that the vote and the procedure for determining a winner (majority wins) is known. There is no uncertainty about this. Just as with causality, there is no difficulty in talking about the probability that someone has a certain degree of responsibility by putting a probability distribution on the way the world could be and how it works. But this misses out on important component of determining what we call here *blame:* the epistemic state. Consider a doctor who treats a patient with a particular drug resulting in the patient's death. The doctor's treatment is a cause of the patient's death; indeed, the doctor may well bear degree of responsibility 1 for the death. However, if the doctor had no idea that the treatment had adverse side effects for people with high blood pressure, he should perhaps not be held to blame for the death. Actually, in legal arguments, it may not be so relevant what the doctor actually did or did not know, but what he *should have known.* Thus, rather than considering the doctor's actual epistemic state, it may be more important to consider what his epistemic state should have been. But, in any case, if we are trying to determine whether the doctor is to blame for the patient's death, we must take into account the doctor's epistemic state.

We present a definition of blame that considers whether agent *a* performing action b is to blame for an outcome φ. The definition is relative to an epistemic state for a, which is taken, roughly speaking, to be a set of situations before action *b* is performed, together with a probability on them. The degree of blame is then essentially the expected degree of responsibility of action *b* for φ (except that we ignore situations where φ was already true or *b* was already performed). To understand the difference between responsibility and blame, suppose that there is a firing squad consisting of ten excellent marksmen. Only one of them has live bullets in his rifle; the rest have blanks. The marksmen do not know which of them has the live bullets. The marksmen shoot at the prisoner and he dies. The only marksman that is the cause of the prisoner's death is the one with the live bullets. That marksman has degree of responsibility 1 for the death; all the rest have degree of responsibility 0. However, each of the marksmen has degree of blame 1/10.[1]

While we believe that our definitions of responsibility and blame are reasonable, they certainly do not capture all the connotations of these words as used in the literature. In the philosophy literature, papers on responsibility typically are concerned with *moral responsibility* (see, for example, [Zimmerman, 1988]). Our definitions, by design, do not take into account intentions or possible alternative actions, both of which seem necessary in dealing with moral issues. For example, there is no question that Truman was in part responsible and to blame for the deaths resulting from dropping the atom bombs on Hiroshima and Nagasaki. However, to decide whether this is a morally reprehensible act, it is also necessary to consider the alternative actions he could have performed, and their possible outcomes. While our definitions do not directly address these moral issues, we believe that they may be helpful in elucidating them.

The rest of this paper is organized as follows. In Section 2 we review the basic definitions of causal models based on structural equations, which are the basis for our definitions of responsibility and blame. In Section 3, we review the definition of causality from [Halpern and Pearl, 2001a], and show how it can be modified to give a definition of responsibility. In Section 3.3, we give our definition of blame. In Section 4, we discuss the complexity of computing responsibility and blame. Proofs of the theorems can be found in the full paper, available at http://www.cs.cornell.edu/home/halpern/papers/blame.ps.

2 Causal Models: A Review

In this section, we review the details of the definitions of causal models from [Halpern and Pearl, 2001a].

A *signature* is a tuple $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$, where $\mathcal{U}$ is a finite set of *exogenous* variables, $\mathcal{V}$ is a set of *endogenous* variables, and the function $\mathcal{R} : \mathcal{U} \cup \mathcal{V} \rightarrow \mathcal{D}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y from the range $\mathcal{D}$. Intuitively, the exogenous variables are ones whose values are determined by factors outside the model, while the endogenous variables are ones whose values are ultimately determined by the exogenous variables. A *causal model* over signature $\mathcal{S}$ is a tuple $M = \langle \mathcal{S}, \mathcal{F} \rangle$, where $\mathcal{F}$ associates with every endogenous variable $X \in \mathcal{V}$ a function F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U) \times (\times_{Y \in \mathcal{V} \setminus \{X\}} \mathcal{R}(Y))) \rightarrow \mathcal{R}(X)$. That is, F_X describes how the value of the endogenous variable X is determined by the values of all other variables in $\mathcal{U} \cup \mathcal{V}$. If the range $\mathcal{D}$ contains only two values, we say that M is a *binary causal model*.

We can describe (some salient features of) a causal model M using a *causal network*. This is a graph with nodes corresponding to the random variables in $\mathcal{V}$ and an edge from a node labeled X to one labeled Y if F_Y depends on the value of X. Intuitively, variables can have a causal effect only on their descendants in the causal network; if Y is not a descendant of X, then a change in the value of X has no affect on the value of Y. For ease of exposition, we restrict attention to what are called *recursive* models. These are ones whose associated causal network is a directed acyclic graph (that is, a graph that has no cycle of edges). Actually, it suffices for our purposes that, for each setting $\vec{u}$ for the variables in U, there is no cycle among the edges of causal network. We call a setting $\vec{u}$ for the variables in U a *context*. It should be clear that if M is a recursive causal model, then there is always a unique solution to the equations in M, given a *context*.

[1] We thank Tim Williamson for this example.

The equations determined by $\{F_X : X \in \mathcal{V}\}$ can be thought of as representing processes (or mechanisms) by which values are assigned to variables. For example, if $F_X(Y, Z, U) = Y + U$ (which we usually write as $X = Y + U$), then if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set. This equation also gives counterfactual information. It says that, in the context $U = 4$, if Y were 4, then X would be $u + 4$, regardless of what value X, Y, and Z actually take in the real world.

While the equations for a given problem are typically obvious, the choice of variables may not be. For example, consider the rock-throwing example from the introduction. In this case, a naive model might have an exogenous variable *U* that encapsulates whatever background factors cause Suzy and Billy to decide to throw the rock (the details of U do not matter, since we are interested only in the context where *U's* value is such that both Suzy and Billy throw), a variable *ST* for Suzy throws *(ST* = 1 if Suzy throws, and *ST* = 0 if she doesn't), a variable *DT* for Billy throws, and a variable *BS* for bottle shatters. In the naive model, whose graph is given in Figure 1, *BS* is 1 if one of *ST* and BT is 1. (Note that the graph omits the exogenous variable U, since it plays no role. In the graph, there is an arrow from variable *X* to variable *Y* if the value of *Y* depends on the value of *X.)*

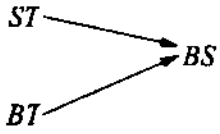

Figure 1: A naive model for the rock-throwing example.

This causal model does not distinguish between Suzy and Billy's rocks hitting the bottle simultaneously and Suzy's rock hitting first. A more sophisticated model might also include variables SH and BH, for Suzy's rock hits the bottle and Billy's rock hits the bottle. Clearly BS is 1 iff one of BH and BT is 1. However, now, SH is 1 if ST is 1, and $BH = 1$ if $BH = 1$ and $SH = 0$. Thus, Billy's throw hits if Billy throws *and* Suzy's rock doesn't hit. This model is described by the following graph, where we implicitly assume a context where Suzy throws first, so there is an edge from SH to BH, but not one in the other direction.

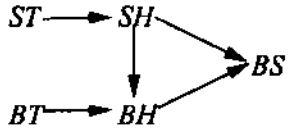

Figure 2: A better model for the rock-throwing example.

Given a causal model $M = (\mathcal{S}, \mathcal{F})$, a (possibly empty) vector $\vec{X}$ of variables in $\mathcal{V}$, and vectors $\vec{x}$ and $\vec{u}$ of values for the variables in $\vec{X}$ and $\mathcal{U}$, respectively, we can define a new causal model denoted $M_{\vec{X}\leftarrow\vec{x}}$ over the signature $\mathcal{S}_{\vec{X}} = (\mathcal{U}, \mathcal{V} - \vec{X}, \mathcal{R}|_{\mathcal{V}-\vec{X}})$. Formally, $M_{\vec{X}\leftarrow\vec{x}} = (\mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}\leftarrow\vec{x}})$, where $F_Y^{\vec{X}\leftarrow\vec{x}}$ is obtained from F_Y by setting the values of the variables in $\vec{X}$ to $\vec{x}$. 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 $\vec{X}$; we do not model the action or its causes explicitly. For example, if M is the more sophisticated model for the rock-throwing example, then $M_{ST\leftarrow 0}$ is the model where Suzy doesn't throw.

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a formula of the form $X = x$, for $X \in V$ and $x \in \mathcal{R}(X)$, is called a *primitive event*. A *basic causal formula* 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 holds in the counterfactual world that would arise if Y_i is set to y_i, $i = 1, \ldots, k$. A *causal formula* is a Boolean combination of basic causal formulas.

A causal formula φ is true or false in a causal model, given a *context*. We write $(M, \vec{u}) \models \varphi$ if φ is true in causal model M given context $\vec{u}$. $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ if the variable X has value x in the unique (since we are dealing with recursive models) solution to the equations in $M_{\vec{Y}\leftarrow\vec{y}}$ in context $\vec{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations $F_Z^{\vec{Y}\leftarrow\vec{y}}$, $Z \in \mathcal{V} - \vec{Y}$, with the variables in $\mathcal{U}$ set to $\vec{u}$). We extend the definition to arbitrary causal formulas in the obvious way.

3 Causality and Responsibility

3.1 Causality

We start with the definition of cause from [Halpern and Pearl, 2001a].

Definition 3.1 *We say that $\vec{X} = \vec{x}$ is a* cause *of φ in $(M, \vec{u})$ if the following three conditions hold:*

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

AC2. *There exist a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}) \models Z = z^*$ for $Z \in \vec{Z}$, then*

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\varphi$. *That is, changing $(\vec{X}, \vec{W})$ from $(\vec{x}, \vec{w})$ to $(\vec{x}', \vec{w}')$ changes φ from* **true** *to* **false**.

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ *for all subsets $\vec{Z}'$ of $\vec{Z}$. That is, setting $\vec{W}$ to $\vec{w}'$ should have no effect on φ as long as $\vec{X}$ has the value $\vec{x}$, even if all the variables in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.*

AC3. *$(\vec{X} = \vec{x})$ is minimal, that is, no subset of $\vec{X}$ satisfies*

AC1 just says that *A* cannot be a cause of *B* unless both *A* and *B* are true, while AC3 is a minimality condition to prevent, for example, Suzy throwing the rock and sneezing from

being a cause of the bottle shattering. Eiter and Lukasiewicz 2002b showed that one consequence of AC3 is that causes can always be taken to be single conjuncts. Thus, from here on in, we talk about $X = x$ being the cause of φ, rather than $\vec{X} = \vec{x}$ being the cause. The core of this definition lies in AC2. Informally, the variables in $\vec{Z}$ should be thought of as describing the "active causal process" from X to φ. These are the variables that mediate between X and φ. AC2(a) is reminiscent of the traditional counterfactual criterion, according to which $X = x$ is a cause of φ if change the value of X results in φ being false. However, AC2(a) is more permissive than the traditional criterion; it allows the dependence of φ on X to be tested under special *structural contingencies*, in which the variables $\vec{W}$ are held constant at some setting $\vec{w}'$. AC2(b) is an attempt to counteract the "permissiveness" of AC2(a) with regard to structural contingencies. Essentially, it ensures that X alone suffices to bring about the change from φ to $\neg\varphi$; setting $\vec{W}$ to $\vec{w}'$ merely eliminates spurious side effects that tend to mask the action of X.

To understand the role of AC2(b), consider the rock-throwing example again. In the model in Figure 1, it is easy to see that both Suzy and Billy are causes of the bottle shattering. Taking $\vec{Z} = \{ST, BS\}$, consider the structural contingency where Billy doesn't throw ($BT = 0$). Clearly $[ST \leftarrow 0, BT \leftarrow 0]BS = 0$ and $[ST \leftarrow 1, BT \leftarrow 0]BS = 1$ both hold, so Suzy is a cause of the bottle shattering. A symmetric argument shows that Billy is also the cause.

But now consider the model described in Figure 2. It is still the case that Suzy is a cause in this model. We can take $\vec{Z} = \{ST, SH, BS\}$ and again consider the contingency where Billy doesn't throw. However, Billy is *not* a cause of the bottle shattering. For suppose that we now take $\vec{Z} = \{BT, BH, BS\}$ and consider the contingency where Suzy doesn't throw. Clearly AC2(a) holds, since if Billy doesn't throw (under this contingency), then the bottle doesn't shatter. However, AC2(b) does not hold. Since $BH \in \vec{Z}$, if we set BH to 0 (it's original value), then AC2(b) requires that $[BT \leftarrow 1, ST \leftarrow 0, BH \leftarrow 0](BS = 1)$ hold, but it does not. Similar arguments show that no other choice of $(\vec{Z}, \vec{W})$ makes Billy's throw a cause.

3.2 Responsibility

The definition of responsibility in causal models extends the definition of causality.

Definition 3.2 *The* degree of responsibility of $X = x$ for φ in $(M, \vec{u})$, *denoted* $\mathbf{dr}((M, \vec{u}), (X = x), \varphi)$, *is* 0 *if* $X = x$ *is not a cause of* φ *in* $(M, \vec{u})$; *it is* $1/(k+1)$ *if* $X = x$ *is a cause of* φ *in* $(M, \vec{u})$ *and there exists a partition* $(\vec{Z}, \vec{W})$ *and setting* $(x', \vec{w}')$ *for which AC2 holds such that (a)* k *variables in* $\vec{W}$ *have different values in* $\vec{w}'$ *than they do in the context* $\vec{u}$ *and (b) there is no partition* $(\vec{Z}', \vec{W}')$ *and setting* $(x'', \vec{w}'')$ *satisfying AC2 such that only* $k' < k$ *variables have different values in* $\vec{w}''$ *than they do the context* $\vec{u}$.

Intuitively, $dr((M, \vec{u}), (X = x), \varphi)$ measures the minimal number of changes that have to be made in $\vec{u}$ in order to make φ counterfactually depend on X. If no partition of $\mathcal{V}$ to $(\vec{Z}, \vec{W})$ makes φ counterfactually depend on $(X = x)$, then the minimal number of changes in $\vec{u}$ in Definition 3.2 is taken to have cardinality ∞, and thus the degree of responsibility of $X = x$ is 0. If φ counterfactually depends on $X = x$, that is, changing the value of X alone falsifies φ in $(M, \vec{u})$, then the degree of responsibility of $X = x$ in φ is 1. In other cases the degree of responsibility is strictly between 0 and 1. Note that $X = x$ is a *cause* of φ iff the degree of responsibility of $X = x$ for φ is greater than 0.

Example 3.3 Consider the voting example from the introduction. Suppose there are 11 voters. Voter i is represented by a variable X_i, $i = 1, \ldots, 11$; the outcome is represented by the variable O, which is 1 if Mr. B wins and 0 if Mr. B wins. In the context where Mr. B wins 11–0, it is easy to check that each voter is a cause of the victory (that is $X_i = 1$ is a cause of $O = 1$, for $i = 1, \ldots, 11$). However, the degree of responsibility of $X_i = 1$ for is $O = 1$ is just $1/6$, since at least five other voters must change their votes before changing X_i to 0 results in $O = 0$. But now consider the context where Mr. B wins 6–5. Again, each voter who votes for Mr. B is a cause of him winning. However, now each of these voters have degree of responsibility 1. That is, if $X_i = 1$, changing X_i to 0 is already enough to make $O = 0$; no other variables need to change.

Example 3.4 It is easy to see that Suzy's throw has degree of responsibility 1/2 for the bottle shattering both in the naive model described in Figure 1 and the "better" model of Figure 2 in the context where both Suzy and Billy throw. In both cases, we must consider the contingency where Billy docs not throw. Although Suzy is the only cause of the bottle shattering in the latter model, her degree of responsibility is still only 1/2, since the bottle would have shattered even if she hadn't thrown. This is a subtlety not captured by the notion of causality.

Interestingly, in a companion paper [Chockler *et al*, 2003] we apply our notion of responsibility to program verification. The idea is to determine the degree of responsibility of the setting of each state for the satisfaction of a specification in a given system. For example, given a specification of the form Op (eventually *p* is true), if/; is true in only one state of the verified system, then that state has degree of responsibility 1 for the specification. On the other hand, if *p* is true in three states, each state only has degree of responsibility 1/3. Experience has shown that if there are many states with low degree of responsibility for a specification, then either the specification is incomplete (perhaps *p* really did have to happen three times, in which case the specification should have said so), or there is a problem with the system generated by the program, since it has redundant states.

The degree of responsibility can also be used to provide a measure of the degree of fault-tolerance in a system. If a component is critical to an outcome, it will have degree of responsibility 1. To ensure fault tolerance, we need to make sure that no component has a high degree of responsibility for an outcome. Going back to the example of O P,, the degree of responsibility of 1/3 for a state means that the system is robust to the simultaneous failures of at most two states.

3.3 Blame

The definitions of both causality and responsibility assume that the context and the structural equations are given; there is no uncertainty. We are often interested in assigning a degree of *blame* to an action. This assignment depends on the epistemic state of the agent before the action was performed. Intuitively, if the agent had no reason to believe that his action would result in a certain outcome, then he is not to blame for the outcome (even if in fact his action caused the outcome).

To deal with the fact that we are considering two points in time—before the action was performed and after—we add superscripts to variables. We use a superscript 0 to denote the value of the random variable before the action was performed and the superscript 1 to denote the value of the random variable after. Thus, $Y^0 = 1$ denotes that the random variable Y has value 1 before the action is performed, while $Y^1 = 2$ denotes that it had value 1 afterwards. If ψ is a Boolean combination of (unsuperscripted) random variables, we use ψ^0 and ψ^1 to denote the value of ψ before and after the action is performed, respectively.

There arc two significant sources of uncertainty for an agent who is contemplating performing an action:

- what the true situation is; for example, a doctor may be uncertain about whether a patient has high blood pressure.
- how the world works; for example, a doctor may be uncertain about the side effects of a given medication.

In our framework, the "true situation" is determined by the context and "how the world works" is determined by the structural equations. Thus, we model an agent's uncertainty by a pair (AC, Pr), where AC is a set of pairs of the form (A/, a), where M is a causal model and u is a context, and Pr is a probability distribution over AC. Following [Hlalpern and Pearl, 2001b], who used such epistemic states in the definition *of explanation,* we call a pair (A/, u) a *situation.*

Roughly speaking, the degree of blame that setting X to x has for ψ is the expected degree of responsibility of $X = x$ for ψ, taken over the situations $(M_{X \leftarrow x, \vec{Y} \leftarrow \vec{y}}, \vec{u})$, where $(M, \vec{u}, \vec{Y} \leftarrow \vec{y}) \in \mathcal{K}$. Our actual definition of blame is just this definition, except that, when computing the expectation, we do not count situations in AC where ψ was already true or X was already x. To understand why, suppose that we are trying to compute the degree of blame of Suzy's throwing the rock for the bottle shattering. Assume that we are interested in a period in a bottle in the period between time 0 and time 1, and the bottle was actually shattered at time 1. We certainly don't want to say that Suzy's throw was to blame if Suzy didn't throw between time 0 and time 1 or if the bottle was already shattered at time 0. So suppose that Suzy does in fact throw between time 0 and time 1, and at time 0, she considers the following four situations to be equally likely:

- (M_1, u_1), where the bottle was already shattered before Suzy's throw;
- (M_2, u_2), where the bottle was whole before Suzy's throw, and Suzy and Billy both hit the bottle simultaneously (as described in the model in Figure 1);
- (M_3, u_3), where the bottle was whole before Suzy's throw, and Suzy's throw hit before Billy's throw (as described in the model in Figure 2); and
- (M_4, u_4), where the bottle was whole before Suzy's throw, and Billy did not throw.

To compute the degree of blame assigned to Suzy's throwing the rock for the bottle shattering, we ignore $(A/, u_1)$, because the bottle is already shattered in (M, u_1) before Suzy's action. The degree of responsibility of Suzy's throw for the bottle shattering is 1/2 in (M_2, u_2) and (M_3, u_3), and is 1 in $(M_4, 7/4)$. It is easy to see that the degree of blame is $\frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot 1 = \frac{1}{2}$.

Definition 3.5 *The* degree of blame of setting X to x for φ relative to epistemic state $(\mathcal{K}, \Pr)$, *denoted* $\mathrm{db}(\mathcal{K}, \Pr, X \leftarrow x, \varphi)$, *is* $\sum_{\{(M,\vec{u}) \in \mathcal{K}:\, (M,\vec{u}) \models X^0 \neq x \wedge \neg \varphi^0\}} \mathrm{dr}((M, \vec{u}), X^1 = x, \varphi^1) \Pr((M, \vec{u}))$.

Example 3.6 Consider again the example of the firing squad with ten excellent marksmen. Suppose that marksman 1 knows that exactly one marksman has a live bullet in his rifle. Thus, he considers 10 situations possible, depending on who has the bullet. Let p_L be his prior probability that marksman i has the live bullet. Then the degree of blame of his shot for the death is p_i. The degree of responsibility is cither 1 or 0, depending on whether or not he actually had the live bullet. Thus, it is possible for the degree of responsibility to be 1 and the degree of blame to be 0 (if he ascribes probability 0 to his having the live bullet, when in fact he does), and it is possible for the degree of responsibility to be 0 and the degree of blame to be 1 (if he mistakenly ascribes probability 1 to his having the bullet when he in fact does not).

Example 3.7 The previous example suggests that both degree of blame and degree of responsibility may be relevant in a legal setting. Another issue that is relevant in legal settings is whether to consider actual epistemic state or to consider what the epistemic state should have been. The former is relevant when considering intent. To see the relevance of the latter, consider a patient who dies as a result of being treated by a doctor with a particular drug. Assume that the patient died due to the drug's adverse side effects on people with high blood pressure and, for simplicity, that this was the only cause of death. Suppose that the doctor was not aware of the drug's adverse side effects. (Formally, this means that he docs not consider possible a situation with a causal model where taking the drug causes death.) Then, relative to the doctor's actual epistemic state, the doctor's degree of blame will be 0. However, a lawyer might argue in court that the doctor should have known that treatment had adverse side effects for patients with high blood pressure (because this is well documented in the literature) and thus should have checked the patient's blood pressure. If the doctor had performed this test, he would of course have known that the patient had high blood pressure. With respect to the resulting epistemic state, the doctor's degree of blame for the death is quite high. Of course, the lawyer's job is to convince the court that the latter epistemic state is the appropriate one to consider when assigning degree of blame.

4 The Complexity of Computing Responsibility and Blame

In this section we present complexity results for computing the degree of responsibility and blame for general recursive models.

4.1 Complexity of responsibility

Complexity results for computing causality were presented by Eiter and Lukasiewicz 2002a; 2002b. They showed that the problem of detecting whether $X = x$ is an actual cause of ψ is Σ_2^P-complete for general recursive models and NP-complete for binary models [Liter and Lukasiewicz, 2002b]. (Recall that Σ_2^P is the second level of the polynomial hierarchy and that binary models are ones where all random variables can take on exactly two values.) There is a similar gap between the complexity of computing the degree of responsibility and blame in general models and in binary models.

For a complexity class *A*, $\mathbf{FP}^{\mathbf{A}[\log n]}$ consists of all functions that can be computed by a polynomial-time Turing machine with an oracle for a problem in *A*, which on input *x* asks a total of $O(\log |x|)$ queries (cf. [Papadimitriou, 1984]). We show that computing the degree of responsibility of $X = x$ for ψ in arbitrary models is $\mathbf{FP}^{\Sigma_2^P[\log n]}$-complete. In [Chockler *et al*., 2003], we show that computing the degree of responsibility in binary models is $\mathbf{FP}^{\mathbf{NP}[\log n]}$-complete.

Since there are no known natural $\mathbf{FP}^{\Sigma_2^P[\log n]}$-complete problems, the first step in showing that computing the degree of responsibility is $\mathbf{FP}^{\Sigma_2^P[\log n]}$-complete is to define an $\mathbf{FP}^{\Sigma_2^P[\log n]}$-complete problem. We start by defining one that we call MAXQSAT_2.

Recall that a *quantified Boolean formula* [Stock meyer, 1977] (QBF) has the form $\forall X_1 \exists X_2 \ldots \psi \ldots$, where X_1, X_2,... are propositional variables and ψ is a propositional formula. A QBF is *closed* if it has no free propositional variables. TQBF consists of the closed QBF formulas that are true. For example, $\forall X \exists Y (X \Rightarrow Y) \in \text{TQBF}$. As shown by Stockmeyer 1977, the following problem QSAT_2 is Σ_2^P-complete:

$$\text{QSAT}_2 = \{\exists X \forall Y \psi(X, Y) \in TQBF : \psi \in 3\text{-CNF}\}.$$

That is, QSAT_2 is the language of all true QBFs of the form $\exists \vec{X} \forall \vec{Y} \psi$, where ψ is a Boolean formula in 3-CNF.

A *witness f* for a true closed QBF $\exists \vec{X} \forall \vec{Y} \psi$ is an assignment / to X under which $\forall Y \psi$ is true. We define MAXQSAT2 as the problem of computing the maximal number of variables in X that can be assigned 1 in a witness for $\exists \vec{X} \forall \vec{Y} \psi$. Formally, given QBF $\Phi = \exists \vec{X} \forall \vec{Y} \psi$, n e $\mathbf{MAXQSAT}_2(\Phi)$ to be *k* if there exists a witness for that assigns exactly *k* of the variables in X the value 1, and every other witness for Φ assigns at most $k' < k$ variables in A' the value 1. If $\Phi \notin \text{QSAT}_2$, then $\mathbf{MAXQSAT}_2(\Phi) = -1$

Theorem 4.1 *MAXQSAT_2 is $FP^{\Sigma_2^P[\log n]}$-complete.*

The proof of Theorem 4.1 shows explicitly how to reduce the computation of any $\mathbf{FP}^{\Sigma_2^P[\log n]}$ problem to MAXQSAT2. Given a polynomial-time Turing machine *M* with oracle in Σ_2^P that on an input of size *n* makes $O(\log n)$ oracles queries, we construct a formula $\Phi = \exists \vec{X} \forall \vec{Y} \varphi$, where ψ is a Boolean formula in 3-CNF such that given $\mathbf{MAXQSAT}_2(\Phi)$ we can compute the output of *M* in polynomial time. Essentially, the formula ψ describes the output of *M* for all possible sequences of answers for oracle queries and $\mathbf{MAXQSAT}_2(\Phi)$ gives the correct sequence of answers. Since the total number of oracle queries is $O(\log n)$, the number of all possible sequences of answers is polynomial in the size of the input, and thus the size of Φ is also polynomial in the size of the input. The construction is somewhat complicated; see the full paper for details.

Similarly to MAXQSAT2, we define $\mathbf{MINQSAT}_2(\exists \vec{X} \forall \vec{Y} \psi)$ to be the minimum number of variables in X that can be assigned 1 in a witness for $\exists \vec{X} \forall \vec{Y} \psi$ if there is such a witness, and $|\vec{X}| + 1$ otherwise. It is easy to see that MINQSAT2 has the same complexity as MAXQSAT_2, since

$$\mathbf{MAXQSAT}_2(\exists \vec{X} \forall \vec{Y} \psi) = |\vec{X}| - \mathbf{MINQSAT}_2(\exists \vec{X} \forall \vec{Y} \overline{\psi}),$$

where (0) is obtained from ψ by replacing each propositional variable $X \in \vec{X}$ with its negation.

Using a reduction from MINQSAT_2, we can prove the desired complexity result.

Theorem 4.2 *The degree of responsibility is $FP^{\Sigma_2^P[\log n]}$-complete for general recursive causal models.*

Proof: Due to lack of space, we present the proof here in a somewhat abridged form. Membership in $\mathbf{FP}^{\Sigma_2^P[\log n]}$ can be proved using an argument similar to the one used in [Chockler *et al,* 2003] to prove membership of the degree of responsibility for binary causal models in $\mathbf{FP}^{\mathbf{NP}[\log n]}$

The proof that computing the degree of responsibility is $\mathbf{FP}^{\Sigma_2^P[\log n]}$-hard essentially follows from an argument in [Eiter and Lukasiewicz, 2002a] showing that QSAT_2 can be reduced to the problem of detecting causality. In fact, their argument actually provides a reduction from MINQSAT_2 to the degree of responsibility. Given a QBF of the form $\exists \vec{X} \ldots$ ψ Eiter and Lukasiewicz construct a causal model *M* whose endogenous variables include X', Y, and a fresh variable X^*. They consider a context *u* in which all variables in X' get the value 0. They show that $\forall \vec{X} \exists \vec{Y} \psi$ is true iff X'* = 0 is a cause of ψ in (M, *u).* If X* = 0 is indeed a cause, then the set *W* in AC2 must be a subset of X. That is, if there is a partition *(W,Z)* and a setting *(x',w')* satisfying AC2 showing that $X^* = 0$ is a cause of ψ, then *W* must be a subset of X. Moreover, the variables in *W* that change value from 0 to 1 in *w'* are precisely those such that an assignment f that assigns 1 to just those variables makes $\forall \vec{Z} \psi$ true. It then easily follows that $\mathbf{MINQSAT}_2(\exists \vec{X} \forall \vec{Y} \psi) = i < |\vec{X}| + 1$ iff $dr((M, u), X^* = 0, \psi) = 1/(i + 1)$, and $\mathbf{MINQSAT}_2(\exists \vec{X} \forall \vec{Y} \psi) = |\vec{X}| + 1$ iff $dr((M, u), X^* = 0, \psi) = 0$. □

4.2 Complexity of blame

Given an epistemic state (K, Pr), where K consists of N possible situations, each with at most *n* random variables, the straightforward way to compute $db(\mathcal{K}, \Pr, X \leftarrow x, \varphi)$ is by computing $dr((M, \vec{u}), X^1 = x, \varphi^1)$ for each $(M, \vec{u}) \in \mathcal{K}$ such that $(M, \vec{u}) \models X^0 \neq x \wedge \neg\varphi^0$, and then computing the expected degree of responsibility with respect to these situations, as in Definition 3.5. Recall that the degree of responsibility in each model is determined by using a binary search thus uses $O(\log n)$ queries in each model in K. Since the number of models is *N,* we get a polynomial time algorithm with *N* log n oracle queries. The type of oracle depends on whether the models are binary or general. For binary models it is enough to have an NP oracle, whereas for general models we need a Σ_2^P**-oracle.** We do not have a matching lower bound but, as we show in the full paper, any binary-search style algorithm for computing the degree of blame requires $\Omega(N \log n)$ oracle queries, for *N — o(n).*

Acknowledgment We thank Michael Ben-Or and Orna Kupferman for helpful discussions.

References

[Chockler et al., 2003] H. Chockler, J. Y. Halpern, and O. Kupferman. Causality and responsibility in temporal logic model checking. Unpublished manuscript. Available at http://www.cs.cornell.edu/home/halpcni/papers/resp.ps, 2003.

[Collins *et ai,* 2003] J. Collins, N. Hall, and L. A. Paul, editors. *Causation and Counter/actuals.* MIT Press, Cambridge, Mass., 2003.

[Eiter and Lukasiewicz, 2002a] T. Eiter and T. Lukasiewicz. Causes and explanations in the structural-model approach: tractable cases. In *Proc. Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI2002),* pages 146—153,2002.

[Eiter and Lukasiewicz, 2002b] T. Eiter and T. Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence,* 142(1):53-89, 2002.

[Hall, 2003] N. Hall. Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and Counter/actuals.* MIT Press, Cambridge, Mass., 2003.

[Halpern and Pearl, 2001a] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part 1: Causes. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001),* pages 194-202, 2001. The full version of the paper is available at http://www.cs.cornell.edu/home/halpern.

[Halpern and Pearl, 2001b] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. In *Proc. Seventeenth International Joint Conference on Artificial Intelligence (IJCAI '01),* pages 27-34, 2001. The full version of the paper is available at http://www.cs.cornell.edu/home/halpern.

[Hume, 1739] D. Hume. *A Treatise of Human Nature.* John Noon, London, 1739.

[Papadimitriou, 1984] C.H. Papadimitriou. The complexity of unique solutions. *Journal of ACM,* 31:492-500, 1984.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, 2000.

[Stockmeyer, 1977] L. J. Stockmeyer. The polynomial-time hierarchy. *Theoretical Computer Science,* 3:1-22, 1977.

[Zimmerman, 1988] M. Zimmerman. *An Essay on Moral Responsibility.* Rowman and Littlefield, Totowa, N.J., 1988.