Causality, Responsibility, and Blame: A Structural-Model Approach

Joseph Y. Halpern
Computer Science Department
Cornell University, Ithaca, NY 14853, USA
halpern@cs.cornell.edu
`www.cs.cornell.edu/home/halpern`

Abstract

This talk will provide an overview of work that I have done with Hana Chockler, Orna Kupferman, and Judea Pearl [1, 2, 10, 9] on defining notions such as causality*,* explanation*,* responsibility*, and* blame*. I first review the Halpern-Pearl definition of causality—what it means that A is a cause of B—and show how it handles well some standard problems of causality. This definition of causality (like most in the literature), views causality as an all-or-nothing concept. Either A is a cause of B or it is not. I show how it can be extended to take into account the degree of responsibility of A for B. For example, if someone wins an election 11–0, each person is less responsible for his victory than if he had won 6–5. Finally, I show how this notion of degree of responsibility can be used to provide insight into model checking notions such as* coverage*.*

This talk will provide an overview of work that I have done with Hana Chockler, Orna Kupferman, and Judea Pearl [1, 2, 10, 9] on defining notions such as *causality*, *explanation*, *responsibility*, and *blame*, showing how they can be applied in a number of contexts, of which perhaps the most relevant here is model checking. I briefly review the issues here. The interested reader is encouraged to consult the papers on which the talk is based for more intuition, details, and examples.

Causality is a topic that has long been discussed in the philosophy literature. It is a surprisingly subtle notion. There have been many attempts to define what it means for an event A to be an *actual cause* of an event B, going back to [13], and continuing to the present (see, for example, [6, 15] for some recent work). The problem of defining actual causality is important far beyond philosophy. For example, it is highly relevant in legal reasoning [11]. If someone sues an automobile manufacturer if a car flips over, the court must establish whether the car design was a cause of the car flipping over, as opposed to bad driving or slick road conditions. Perhaps less obviously, the notion of causality has also arisen in formal verification. For example, Groce et al. [7] introduce a notion of *error explanation*, which tries to make precise whether a line of code is a cause of an error.

My talk will start by a discussion of the problem of defining the notion of actual cause, and introduce a definition of actual cause due Judea Pearl and me [10]. Like many other definitions of causality going back to Hume [13], 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. Consider the following example (due to Hall [8]): 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.

The Halpern and Pearl (HP from now on) definition deals with this problem by idea is that A is a cause of B if B counterfactually depends on C *under some contingency*. For example, 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.)

I will show how the HP definition deals well with many of the standard problematic examples in the philosophy literature. The HP definition also has another significant advantage. All the standard definitions of causality (including the HP definition) treat causality as an all-or-nothing concept. While there may be more than one cause for an event B, an event A is either a cause of B or it is not. As a consequence, thinking only in terms of causality does 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.

Hana Chockler and I [1] show how the HP definition can be extended to define a notion of *responsibility* that takes this distinction into account. To understand the intuition behind our definition, note that it is clear, using the standard counterfactual definition that each voter how votes for Mr. B is a cause of him winning in the cause of the 6–5 vote. Had any of the voters for Mr. B (counterfactually) changed his or her vote, then Mr. B would not have won. Under the HP definition, each voter for Mr. B is also a cause of him winning in the case of the 11–0 vote. 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.

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 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.

Thus, the notion of responsibility gives us a more fine-grained way of thinking about causality. If A is not a cause of B, then A's degree of responsibility for B is 0; if A is a cause of B, then A's degree of responsibility for B is strictly greater than 0, and can be as high as 1. Degree of responsibility does not work like probability. In the case of the 6–5 victory, each of the the six people who voted for Mr. B is a cause of him winning, and each has degree of responsibility 1. In the case of the 11–0 victory, again, each of the eleven people who voted for Mr. B is a cause of him winning, and each has degree of responsibility 1/6.

In determining causality and 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. With both causality and responsibility, 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 Chockler and I called *blame*: the epistemic state.[1] 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.

Chockler and I 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 $1/10$.[2]

While the notions of degree of responsibility and blame are crude, they do seem to capture some of our intuitions. They can certainly be applied in legal settings such as the car flipping example. Of most interest to us here is the application of degree of responsibility in model checking. A model checker verifies the correctness of a finite-state sys-

[1] In English we use both the terms "responsibility" and "blame" for a number of related but distinct notions that we were trying to tease apart. We used the term "responsibility" for the non-epistemic version and "blame" for the epistemic version, just for definiteness. The key point is that these are distinct notions.

[2] This example is due to Tim Williamson.

tem with respect to a desired behavior by checking whether a labeled state-transition graph that models the system satisfies a specification of this behavior [5]. If a system does not satisfy a specification, model checkers typically provide a counterexample showing why. These counterexamples can be essential in detecting subtle errors in complex designs [4]. On the other hand, when a system does satisfy the specification, most model-checking tools terminate with no further information to the user.

In the last few years, however, there has been growing awareness that further analysis may be necessary in the latter case. One approach that has been used is *coverage estimation*. Roughly speaking, a component or a state is *covered* by a specification ψ if changing this component falsifies ψ (see [12, 3]). For example, if a specification requires that $AG(req \rightarrow AFgrant)$ (every request is eventually followed by a grant on every path) holds at an initial state, and there is a path in which *req* holds only in one state, followed by two states both satisfying *grant*, then neither of these two states is covered by the specification (changing the truth of *grant* in either one does not render the specification untrue). On the other hand, if there is only one state on the path in which *grant* holds, then that state is covered by the specification. The intuition is that the presence of many uncovered states suggests that either the specification the user really desires has more requirements than those explicitly written (for example, perhaps the specification should really require a correspondence between the number of requests and grants), or that the system contains redundancies, and can perhaps be simplified (for example, perhaps there should be only a single grant on the path). This approach has already proven to be effective in practice, detecting design errors that escape early verification efforts in industrial settings [12].

Coverage considers can be viewed as trying to answer the question of *what causes the system to satisfy the specification*. Indeed, the main definitions of coverage in the literature are inspired by counterfactual dependence: a state s is p-covered by the specification ψ if, had the value of the atomic proposition p been different in state s, then ψ would not have been true. Chocker, Kupferman, and I [2] show that the initial definition of coverage [12] and its generalization [3] can be understood in terms of causality, as can a number of interesting variants. Doing so gives significant insight into unresolved issues regarding the definition of coverage, and leads to potentially useful extensions of coverage. Moreover, thinking in terms of degree of responsibility allows a finer-grained analysis of coverage that gives even more insight.

Consider for example the specification *EXp* (p holds at the next state of some path starting from the initial state). There seems to be a qualitative difference between a system where the initial state has 100 successors satisfying p and one where there are only two successors satisfying p. Although, in both cases, no state is p-covered by the specification, intuitively, the states that satisfy p play a more important role in the case where there are only two of them than in the case where there are 100 of them. That is, each of the two successors is more responsible for the satisfaction of EXp than each of the 100 successors.

Note that having a low degree of responsibility is not necessarily a bad thing in the context of fault tolerance. A state with a high degree of responsibility is intuitively a critical state. To ensure that a system can cope with unexpected hardware or software faults, such as a power failure, a link failure, or a Byzantine failure [14], we may want to ensure that no state has a high degree of responsibility. To increase fault tolerance, we want states to be uncovered. On the other hand, while we may not want to have nodes with degree of responsibility 1, since that implies a single point of failure, a degree of responsibility of $1/100$ implies perhaps unnecessary redundancy.

As I hope this talk will make clear, the notions of causality, responsibility, and blame can be applied usefully in a number of settings. While we have taken preliminary steps, I believe that much more can be done.

References

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

[2] H. Chockler, J. Y. Halpern, and O. Kupferman. What causes a system to satisfy a specification? Unpublished manuscript. Available at http://www.cs.cornell.edu/home/halpern/papers/resp.ps, 2003.

[3] H. Chockler, O. Kupferman, and M. Vardi. Coverage metrics for temporal logic model checking. In *Tools and Algorithms for the Construction and Analysis of Systems*, number 2031 in Lecture Notes in Computer Science, pages 528 – 542. Springer-Verlag, 2001.

[4] E. Clarke, O. Grumberg, K. McMillan, and X. Zhao. Efficient generation of counterexamples and witnesses in symbolic model checking. In *Proc. 32nd Design Automation Conference*, pages 427–432. IEEE Computer Society, 1995.

[5] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. MIT Press, Cambridge, Mass., 1999.

[6] J. Collins, N. Hall, and L. A. Paul, editors. *Causation and Counterfactuals*. MIT Press, Cambridge, Mass., 2004.

[7] A. Groce, S. Chaki, D. Kroening, and O. Strichman. Error explanation with distance metrics. *International Journal on Software Tools for Technology Transfer (STTT)*. accepted for publication.

[8] N. Hall. Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and Counterfactuals*. MIT Press, Cambridge, Mass., 2004.

[9] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science*, pages 843–887, 2005.

[10] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *British Journal for Philosophy of Science*, 2005.

[11] H. L. A. Hart and T. Honoré. *Causation in the Law*. Oxford University Press, Oxford, U.K., second edition, 1985.

[12] Y. Hoskote, T. Kam, P.-H. Ho, and X. Zhao. Coverage estimation for symbolic model checking. In *Proc. 36th Design Automation Conference*, pages 300–305, 1999.

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

[14] N. Lynch. *Distributed Algorithms*. Morgan Kaufmann, San Francisco, 1996.

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