Causality

Jerry R. Hobbs
Artificial Intelligence Center
SRI International

Abstract

We do things in the world by exploiting our knowledge of what causes what. But in trying to reason formally about causality, there is a difficulty: to reason with certainty we need complete knowledge of all the relevant events and circumstances, whereas everyday reasoning tasks we need a more serviceable but looser notion that does not make such demands on our knowledge. In this work I introduce the notion of "causal complex" for a complete set of events and conditions necessary for the causal consequent to occur, and use the term "cause" for the makeshift, nonmonotonic notion we require for everyday tasks such as planning and language understanding. I discuss the issue of how to distinguish between what is in a causal complex from what is outside it, and within a causal complex, how to distinugish the eventualities that deserve to be called "causes" from those that do not, in particular circumstances. In addition, I discuss some general properties of causality, such as transitivity, and how related notions such as "prevent", "enable", and "maintain", can be defined in terms of "cause".

1 Causal Complexes

Introduction: It is natural to say that when you flip a light switch, you cause the light to go on. But it would not happen if a whole large system of other conditions were not in place. The wiring has to connect the switch to the socket, and be intact. The light bulb has to be in good working order. The switch has to be connected to a system for supplying electricity. The power plant in that system has to be operational. And so on. Flipping the light switch is only the last small move in a large-scale system of actions and conditions required for the light to go on.

I will take as my starting point that people are able to recognize that a particular effect is caused by some "causal complex". By "causal complex" I mean some collection of eventualities (events or states) whose holding or happening entails that the effect will happen. People may not know a priori what events or eventualities are in the causal complex or what constraints or laws the world is operating under. But they are able to reason to some extent about what may or may not be part of that causal complex. The first step in coming to a clear account of causality is deciding how to talk about such causal complexes and what criteria there are for deciding what eventualities are in or out of the causal complex for a particular effect.

But "causal complex" is not the most serviceable of notions, since it requires complete knowledge of what of all the relevant events and circumstances. for everyday reasoning tasks, such as planning and language understanding, we need a looser notion that does not make such demands on our knowledge. That is the notion of "cause". So the second step in this investigation is determining what we should mean by the predicate *cause*, grounded in the more rigorous theory of causal complexes.

Splitting the inquiry like this into an investigation of causal complexes and an investigation of the predicate *cause* leads us to see two principal questions about causality that have been addressed in the literature:

1. How do we distinguish between what eventualities are in a causal complex from those that are outside it.

2. Within a causal complex, how do we distinguish the eventualities that deserve to be called "causes" from those that do not.

Lewis (1973), Ortiz (1999b), Simon (1952, 1991), and Pearl (2000) are primarily concerned with the first question. Mackie (1993) and Shoham (1990) are primarily concerned with the second. The first question leads one to examine counterfactuals. The second leads one to introduce nonmonoticity. It is because Simon deals with the first and Shoham with the second that in their exchange (Shoham 1990, 1991; Simon, 1991) they largely talk past one another. I deal with the first question in this section, and the second in Part 3.

It should be noted at the outset that the aim of this paper is the development of a theory of causality that will work equally well for phyiscal causality and other types of causality, such as social, political, and economic causality, and the causality of folk psychology. It should work in any domain where we attribute the occurrence of events to underlying causal principles. Moreover, possible causes should be permitted to be not just actions, but also agentless events and states, such as tornadoes, the slipperiness of the floor, and a signature's not being present on a document.

Before beginning the formal treatment of these questions, I will describe the notational conventions used in this paper. The notation and ontology of Hobbs (1985) will be employed. Briefly, corresponding to every predication $p(x)$ there is an eventuality e which is the eventuality of p being true of x. The expression $p'(e, x)$ says that e is the eventuality of p being true of x. Thus, $tall'(e, John)$ says that e is the eventuality of John's being tall. An eventuality may or may not be exist in a particular possible world. The predication $holds(e, w)$ says that eventuality e exists in world w. The predication $Rexists(e)$ says that eventuality e exists in the real world.

A possible world can be thought of as consisting of a set of eventualities that does not contain both an eventuality and its negation. A possible world may or may not be restricted to a particular moment in time.

Since eventualities correspond to predications, it makes sense to talk about the conjunctions and negations of eventualities, and they are eventualities too. Thus, $and'(e, e_1, e_2)$ says that e is an eventuality that exists when the eventualities e_1 and e_2 exist, and $not'(e, e_1)$ says that e is an eventuality that exists when the eventuality e_1 does not exist. (In the following development, when $not'(e, e_1)$ holds, e will normally be abbreviated to $\neg e_1$.) For a set s of eventualities to hold in a world w, the conjunction of the eventualities in s must hold in w, and thus each of the eventualities in s must hold in w.

The use of this notation allows us to work entirely in first-order logic. When one's primary focus is a particular phenomenon, like causality, a special-purpose logic that highlights the special features of the phenomenon may be justified, and indeed most formal explorations of causality have taken place in such special-purpose logics. But when, as in this research, the effort is part of the larger enterprise of developing an account of natural language understanding and/or reasoning in everyday life, it is better to have a simple and uniform logic for all phenomena, and that is what the introduction of eventualities gives us.

Change Relevance: With this background, let us now perform a thought experiment. Think of the world at any given instant as made up of a very very large set of eventualities which obtain at that instant. (In Pearl (2000) random variables taking on specific values correspond to the eventualities.) Suppose the eventuality e is one of the eventualities that obtain, and we wish to understand the causal complex of which e is an effect.

Now suppose we can reach into this world and cause an eventuality e_1 to become true. That is, we toggle $\neg e_1$ into e_1. This change will propagate along what after the fact we think of as "causal chains", changing other eventualities into their negations. But not everything will change. The effects of any single change tend to be quite local, in some sense of "local". Suppose e is changed in this process. Then we know that making e_1 true is relevant to e. In attempting to identify a causal complex that causes e, we have learned that there is one that has e_1 in it.

Let S be the set of possible worlds. Let $Con(S, C)$ be the subset of S containing the worlds that respect some set C of constraints. The constraints may be thought of as background knowledge, or if the thought experiments are real, then simply the laws that would operate to bring about the consequences of a change. I will be as silent as possible about the structure of C. In particular, C may contain both causal and noncausal constraints.

We first need to define a predicate *closest-world*. We want to say that a world w_2 is a closest world to world w_1 with respect to an eventuality e_1 and a set of constraints C if everything in w_2 different from w_1 is a consequence of adding e. Formally,

$$
\begin{aligned}
&(\forall\, w_2, w_1, e_1, C)\\
&\quad [\mathit{closest\text{-}world}(w_2, w_1, e_1, C)\\
&\quad\equiv [\neg \mathit{holds}(e_1, w_1) \wedge \mathit{holds}(e_1, w_2)\\
&\qquad \wedge (\forall\, e_2 \in w_2 - w_1)[[(w_1 \cap w_2)\\
&\qquad\quad \cup C \cup \{e_1\} \supset e_2]\\
&\qquad\quad \wedge \neg[(w_1 \cap w_2) \cup C \supset e_2]]]]
\end{aligned}
$$

That is, eventuality e_1 doesn't hold in w_1 but it does hold in w_2, and for every eventuality e_2 in the difference between w_2 and w_1, e_2 follows from the common core of w_1 and w_2 and the constraints C together with the eventuality e_1, but does not follow from the common core of w_1 and w_2 and the constraints C alone.

Then we can define *change-relevant* as follows, where the set of possible worlds S and the set of constraints C are fixed:

$$
\begin{aligned}
&(\forall\, e_1, e)[\mathit{change\text{-}relevant}(e_1, e)\\
&\quad \equiv (\exists\, w_1, w_2 \in \mathit{Con}(S, C))\\
&\qquad [\mathit{closest\text{-}world}(w_2, w_1, e_1, C)\\
&\qquad \wedge \neg[\mathit{holds}(e, w_1) \equiv \mathit{hold}(e, w_2)]]]
\end{aligned}
$$

That is, to show that an eventuality e_1 is change-relevant to an effect e, find two possible worlds w_1 and w_2 such that e_1 doesn't hold in w_1 and does hold in w_2 but where the two worlds are otherwise as close as possible, given the constraints C, and where the effect e holds in one world and not in the other. In other words, there is some situation in which turning e_1 on will toggle e.

It is not necessarily the case that if w_2 is the closest world to w_1 when e_1 is turned on, then w_1 is the closest world to w_2 when e_1 is turned off. The constraints in C may cause the changes to propagate more in one direction than in the other. As a result, it does not follow from the definitions that if e_1 is change-relevant to e then so is $\neg e_1$. Events have consequences, and sometimes it is not possible to fix things by merely undoing what triggered the damage. If I drop a vase and it shatters, I can't fix it just be lifting it up again.

Now we can propose the axiom

$$
\begin{aligned}
&(\forall\, s, e)[\mathit{causal\text{-}complex}(s, e)\\
&\quad \supset (\forall\, e_1 \in s)[\mathit{change\text{-}relevant}(e_1, e)]]
\end{aligned}
$$

That is, if a set s of eventualities is a causal complex for an effect e, then all of the eventualities in s are change-relevant to e. Toggling them can change e under the right circumstances. This axiom does not define the notion of a causal complex, but it does constrain it. Change relevance is only a necessary condition for an element of a causal complex; it is not sufficient.

Examples: Consider a line of dominos where they are all close enough to their neighbors to knock them down. The constraints C enforce this. The two possible states for each domino d_i will be being upright, e_i—$\mathit{upright}'(e_i, d_i)$—and being knocked over to the right, $\neg e_i$. Let the effect e be the eventuality of the domino d at the right end being upright.

First let us ask if the eventuality $\neg e_1$ of the leftmost domino d_1 being down is change-relevant to e. To show that it is, we need to find closest worlds w_1 and w_2 such that e holds in one and $\neg e$ in the other. Let w_1 be the world in which all the dominos are upright. In particular, d and d_1 are upright, so e and e_1 hold. Now let us introduce $\neg e_1$ into w_1; that is, we knock d_1 down to the right. Because of the constraints, all the other dominos will go down. Thus the closest world to w_1 after introducing $\neg e_1$ under constraints C is the w_2 in which all the dominos are down. In particular, d is down, so $\neg e$ holds in w_2, and the definition of *change-relevant* is satisfied.

Now let us ask if the eventuality e_1 is change-relevant to e. Let w_1 be any world in which d_1 is down to the right; that is, $\neg e_1$ holds. Because of the constraints, all the other dominos will be down, and in particular, $\neg e$ will hold. Now introduce e_1; that is, set d_1 upright. The constraints as stated do not entail that any other domino will thereby become upright, and thus the closest world to w_1 is the one in which only d_1 is upright. The eventuality $\neg e$ still holds, and thus e_1 is change-irrelevant to e.

If however we augment the constraints C with "frame axioms" that say that the only way a domino can be down is if its neighbor knocked it down, then e_1 *is* change-relevant to e. Ortiz (1999a) builds a solution to the frame problem into his treatment of causality, so that frame axioms do not have to be explicitly stated. I have not done that here.

Temporal Order and Causal Priority: We can say that if an eventuality is a member of a causal complex for an effect, then it is "causally involved" in the effect.

$$
(\forall\, e, e_1)[\mathit{causally\text{-}involved}(e_1, e)
$$

$$\equiv (\exists\, s)[\mathit{causal\text{-}complex}(s, e) \wedge e_1 \in s]]$$

A further constraint on causes and effects is that the cause cannot happen after the effect.

$$(\forall\, e_1, e)[\mathit{causally\text{-}involved}(e_1, e) \supset \neg \mathit{before}(e, e_1)]$$

However, the facts about causal flow are not entirely determined by knowing the times of events. There are cases where events occur simultaneously, but we have clear intuitions about what caused what. For example, flipping the switch and the light coming on are perceptibly simultaneous (except on airplanes). Yet clearly it is flipping the switch that causes the light to go on. When a person hammering a nail, the person's arm reaches the end of its trajectory, the head of the hammer strikes the head of the nail, and the nail begins its motion into the surface simultaneously, but the events have a clear causal order.

It may be that the criteria for causal flow for many such cases can be spelled out in various domain theories with varying specificity. At a general level, we can sometimes make judgments of causal flow between eventualities within a larger causal complex. We certainly want it to be the case that the eventualities in a causal complex for an event are causally prior to the effect, unless there is a feedback loop. Thus,

$$(\forall\, e_1, e_2)[\mathit{causally\text{-}involved}(e_1, e_2) \wedge \neg \mathit{causally\text{-}involved}(e_2, e_1) \supset \mathit{causally\text{-}prior}(e_1, e_2)]$$

Given a causal complex s for an effect e, consider two eventualities e_1 and e_2 in s. There may be cases where we know that e_1 is itself in a causal complex s_1 for e_2 and not vice versa. In this case, we would know that e_1 is causally prior to e_2, independent of information about time.

Causal priority is related to temporal order in that if e_1 is causally involved in e and occurs before e, then it is causally prior to e:

$$(\forall\, e_1, e)[\mathit{causally\text{-}involved}(e_1, e) \wedge \mathit{before}(e_1, e) \supset \mathit{causally\text{-}prior}(e_1, e)]$$

Pearl (2000) faces the problem of causal flow in his treatment of counterfactuals and causality. He models his causal complexes as Bayesian networks. The links in these networks have an intrinsic directionality. There is nothing in the definition of Bayesian networks that requires this directionality to respect the direction of causal flow, but in the examples one sees, they generally do. Because of this, when he looks for the closest world, he can simply excise the links into the node whose value he wants to change, and the backwards propagation of the change is prevented.

Ortiz (1999b) has also dealt with the problem of causal flow by stipulation. He divides his constraints into two sets, those used for prediction, L_P, and those used for explanation, L_E. In the former, inference follows the direction of causal flow; in the latter, it goes against causal flow. When constructing the nearest counterfactual world to the real world, he favors suspending the laws in L_E over those in L_P, thereby preventing, or at least discouraging, propagation of changes against causal flow. (In my view, explanation is not a process of deduction but of abduction. Thus, the same forward rules would be used for both prediction and explanation, but in explanation they would be used abductively by back-chaining over them.)

None of the development here precludes circular causation, or feedback loops. Suppose two books are leaning against each other. The first book's leaning against the second is in the causal complex causing the second to be upright, and vice versa.

Now we can see that the definitions of *closest-world* and *change-relevant* are not as tight as our intuitions will allow. Suppose e_1 causes e and is the only possible cause of e. That is, in the constraints C there is a constraint that whenever one occurs, the other does too, but we nevertheless know that e_1 is causally prior to e. Then e_1 and e occur in all the same worlds, and each is change-relevant to the other, by our current definitions, and there is no distinguishing which is in the causal complex for the other.

To deal with this problem, we first need a way of loosening the constraints on possible worlds. Then we need to stipulate that when we toggle an eventuality, the changes cannot propagate against causal flow.

Consider for example the situation in which someone fires a gun, a loud bang happens, and someone dies. We want to know if the loud bang is causally implicated in the death. The constraints that the set of possible worlds must respect are that the firing occurs if and only if the bang occurs, and that the firing occurs if and only if the death occurs. If the bang occurs, then by the first constraint, so does the firing, and thus by the second constraint so does the death. Therefore, the closest world to the world in which nothing happens is the world in which everything happens, and the bang is change-relevant to the death. Thus it is not ruled out as part of the causal complex that causes the death. Yet we know the firing is causally prior to the bang, and that when we infer the firing from the bang, we are propagating changes against causal flow.

We would like to stipulate in the definition of *closest-world* that in selecting w_2 we only change eventualities that are not causally prior to e_1. But this will normally involve breaking some of the constraints (e.g., bang implies firing), and thus there is no such possible world in the set of possible worlds that respect the constraints C. We need to alter the set of constraints to a weaker set C'.

One way to modify C into C' is, in the style of nonmonotonic logic, to introduce $\neg ab_i$ predications in the antecedents of rules that permit inference from eventualities e_1 to causally prior eventualities. For example, we change the rule

$$\text{bang} \supset \text{fire}$$

into

$$\text{bang} \wedge \neg ab_i(\ldots) \supset \text{fire}$$

Then when we change $\neg e_1$ to e_1, it will be possible to maintain consistency between the altered constraints C' and the condition e_1 by assuming the abnormality predications. We can call this modified set of constraints $Defeas(C, e_1)$.

We can now modify the definition for *closest-world* is as follows:

$$\begin{aligned}
&(\forall w_2, w_1, e_1, C)[closestworld(w_2, w_1, e_1, C)\\
&\quad \equiv w_1 \in Con(S, C)\\
&\qquad \wedge w_2 \in Con(S, Defeas(C, e_1))\\
&\qquad \wedge \neg holds(e_1, w_1) \wedge holds(e_1, w_2)\\
&\qquad \wedge (\forall e_0 \in w_2)\\
&\qquad\quad [causally\text{-}prior(e_0, e_1) \supset e_0 \in w_1]\\
&\qquad \wedge (\forall e_2 \in w_2 - w_1)\\
&\qquad\quad [[(w_1 \cap w_2) \cup Defeas(C, e_1)\\
&\qquad\qquad \cup \{e_1\} \supset e_2]\\
&\qquad\quad \wedge \neg[(w_1 \cap w_2) \cup Defeas(C, e_1)\\
&\qquad\qquad \supset e_2]]]
\end{aligned}$$

The definition of *change-relevant* has to be modified as well, since now world w_2 need only respect the weakened constraints:

$$\begin{aligned}
&(\forall e_1, e)[change\text{-}relevant(e_1, e)\\
&\quad \equiv (\exists w_1 \in Con(S, C),\\
&\qquad\quad w_2 \in Con(S, Defeas(C, e_1)))\\
&\qquad [closest\text{-}world(w_2, w_1, e_1, C)\\
&\qquad \wedge \neg[holds(e, w_1) \equiv hold(e, w_2)]]]
\end{aligned}$$

Now the closest world to the world in which nothing happens is the world in which only the bang happens, and the bang is not change-relevant to the death.

Structure in Causal Complexes: The two key features of causal complexes are that if all the eventualities in the causal complex obtain, then the effect will occur, and there is nothing in the causal complex that is irrelevant to the effect. This characterization allows for internal causal structure in causal complexes.

For example, if someone lets go of a vase, the vase will fall, and when it hits the floor it will shatter. Suppose only the letting go and the falling are relevant eventualities to the effect of shattering. The falling alone constitutes a causal complex for the shattering. It's relevant to the shattering, and if it happens, the shattering happens. Similarly, the letting go alone constitutes a causal complex for the shattering. It's relevant to the shattering, and if it happens, the shattering happens. Finally, the letting go and the falling together constitute a causal complex for the shattering. They are both relevant to the shattering, and if they both happen, the shattering happens.

In a causal complex s for e, it may be the case that there is a subset s_1 of s and an eventuality e_1 in s and not in s_1 such that s_1 is a causal complex of e_1, in which case $s - \{e_1\}$ is also a causal complex for e. For example, since the letting go causes the falling, the letting go is a sufficient causal complex for the shattering. The letting go, in a sense, is more "ultimate" than the falling.

Causal complexes can sometimes be composed. If s_1 is a causal complex for e and contains the eventuality e_1 and s_2 is a causal complex for e_1 and s_1 and s_2 are consistent, then $s_1 \cup s_2$ is a causal complex for e.

2 Counterfactuals

A counterfactual statement is a conditional whose antecedent is counter to the truth, as in

> If John weren't a millionaire, he wouldn't have retired.

It is distinguished in English by the use of the otherwise rare subjunctive mode in the antecedent and the modal "would" in the consequent. In the philosophical and more recently in the AI literature (e.g., Lewis, 1973; Ortiz, 1999b; Pearl, 2000), counterfactuals are taken to give us reliable insights into the facts about causality. Thus, John's being a millionaire caused him to retire.

Counterfactuals are certainly related to causality. In the definitions of *closest-world* and *change-relevant*, suppose e holds in w_2. w_2 results from the occurrence of e_1 in w_1. If this had not occurred, then we would still be in w_1, in which e does not hold. That is, if e_1 had not occurred, then e would not have occurred—the counterfactual. Pearl observes that a counterfactual is a natural language way of saying that one eventuality should be changed and everything else should remain the same, insofar as possible. This is what we try to capture in the definition of *closest-world*. Ortiz develops a rich notion of counterfactual reasoning, including an impressive account of how to minimize the changes triggered by the counterfactual, and defines causality in terms of that.

The position in this paper however is that, aside from the hints it gives us for formalizing causality, counterfactuals in English are just a particular kind of English expression, with no special or priveleged status. I have not adopted the position that we have clear intuitions about the use of counterfactuals that give us special access to facts about causality. Rather I am seeking to develop a clear and coherent theory of causality, where one of the ultimate aims of the theory is to provide the predicates and axioms required for characterizing and relating lexical items with a causal flavor, such as the conjunctions "if", "because", and "so", modal auxilliaries such as "would", and causative verbs, as well as the subjunctive mode, within a framework of interpreting discourse by abduction.

The issue is what counts as evidence. Here the view is that a challenging counterfactual is not direct evidence against a particular theory of causality, but rather a challenge for characterizing "if", "would", and the subjunctive mode in terms of that theory.

3 What the Predicate *cause* Means

We could use predications of the form *causal-complex*(s, e) for encoding our causal knowledge, where we then spell out the nature and interaction of the elements of s. The problem with this is that we rarely know or need to know the entire causal complex for many of the effects in our lives. Very few people really know all that goes into making an electric light work, although they know that flipping the switch will generally turn it on.

The most common situation is one in which nearly all of the causal complex is in place for the effect to occur, and we must only figure out the one or two last steps to complete it. Or in seeking to explain an effect, nearly all of the causal complex normally holds, and only one or two eventualities are in doubt, and they thus constitute the explanation for the effect. Or a causal statement is made in discourse, and to verify its plausibility we don't need to verify the truth of the entire causal complex but only that part of it that is not true normally.

A causal complex will contain a large number of eventualities that are defeasibly true, or assumably true if there is no evidence to the contrary, or normally true, or true with high probability. We may say that in all these cases the eventualities are "presumable" or "presumably true". Only the remainder of the eventualities will be exercised in most reasoning, explanation, interpretation, and planning. The predicate *cause* should be for these latter eventualities. The axioms in which the predicate *cause* occurs will be central in the use of any commonsense knowledge base, whereas appeals to axioms for the predicate *causal-complex* will be relatively rare.

Which eventualities are presumable is very much dependent upon the task that is being performed, the situation or context, and/or the knowledge base being used. Shoham (1990) points out that what we specify as causal laws are just those causal relationships that prove to be useful in everyday reasoning. In part this depends on probabilities; we can ignore as presumable those factors that hold with high probability and focus just on those factors that are in doubt—it is the latter that get expressed in terms of *cause*. In addition, the choice of causal laws depends on utility; even if a factor normally holds, if its not holding would result in catastrophic consequences, we would usually want to reason about it as well and thus would express its role in terms of *cause*. Shoham gives the example of firing a cartridge that is probably a blank. An eventuality is generally a cause if it is manipulable by agents, and thus of use in planning, although there are certainly agentless causes such as a bare wire causing a fire. If an event is the final, triggering event that completes the causal complex and precipitates the effect, it is often identified as a cause. Actions required at significantly lower frequency than other actions are often taken to be presumable; to drive a car, we have to both fill the tank and turn the key in the ignition, but generally we take the former to be presumable.

A causal complex thus consists of presumable eventualities and causes. In addition, there may be eventualities that are triggered by the causes in the presence of the presumable eventualities; we will not call them causes. In the causal complex consisting of letting go of a vase and its falling, the former is the cause and the latter is merely triggered.

Our causal knowledge could be stated in a form using the predicate *causal-complex*:

$$
\begin{aligned}
&(\forall s)[\ldots \text{long characterization of } s \ldots \\
&\quad \supset (\exists e)[q'(e,\ldots) \\
&\qquad \wedge \mathit{causal\text{-}complex}(s,e)]]
\end{aligned}
$$

If we were to do so, we could state our knowledge with certainty. The axioms would be monotonic.

However, we are more likely to learn and use facts about especially useful or manipulable elements of causal complexes, for which the predicate *cause* is appropriate, and the form of the axioms will be as follows:

$$
(1) \quad \begin{aligned}
&(\forall e_1, x)[p'(e_1,x) \supset (\exists e)[q'(e,x) \\
&\qquad \wedge \mathit{cause}(e_1,e)]]
\end{aligned}
$$

That is, p-type things normally cause q-type things. Since we have not made the entire causal complex explicit, this axiom will only be defeasible (as will be most axioms in a commonsense knowledge base).

Axioms schema (1) is the typical form for general causal knowledge. For specific instances of one eventuality causing another, the typical form is the following:

$$
\begin{aligned}
&(\exists e_1, e, x)[p'(e_1,x) \wedge q'(e,x) \\
&\quad \wedge \mathit{cause}(e_1,e) \wedge \mathit{Rexists}(e_1)]
\end{aligned}
$$

That is, e_1 is the eventuality of p's being true of x, e is the eventuality of q's being true of x, e_1 causes e, and e_1 really exists.

Shoham (1990)lists as one of his criteria for a notion of causality that it be nonmonotonic and that the cause not necessarily imply the effect. Separating out the notions of *causal-complex* and *cause* as distinct and deriving *cause* as we have from *causal-complex* makes these properties of *cause* fall out. More precisely, we can state axiom schema (1) as follows:

$$
(1') \quad \begin{aligned}
&(\forall e_1, x)[p'(e_1,x) \wedge \neg ab_i(e_1,x) \\
&\quad \supset (\exists e)[q'(e,x) \wedge \mathit{cause}(e_1,e)]]
\end{aligned}
$$

where

$$
\begin{aligned}
&(\forall e_1, x)[ab_i(e_1,x) \equiv \\
&\quad (\forall e, s)[\mathit{causal\text{-}complex}(s,e) \\
&\qquad \wedge e_1 \in s \wedge q'(e,x) \\
&\qquad \wedge \mathit{presumable}(s - \{e_1\}) \\
&\quad\ \supset (\exists e_2)[e_2 \in s - \{e_1\} \\
&\qquad \wedge \neg \mathit{Rexists}(e_2)]]]
\end{aligned}
$$

That is, causality from p to q will fail when in any causal complex s for q, some presumable eventuality e_2 in s does not obtain; otherwise, p causes q. The form of $(1')$ exactly matches the nonmonotonic causal rules that Shoham uses.

Causal chains play a very important role in commonsense reasoning. Letting go of a vase is the cause of the falling of the vase, which is the cause of its shattering. We would not want to eliminate the possibility of calling the falling a cause of the shattering, just because something caused *it*. In this framework we can handle this by relativizing the notion of triggering to causal complexes. Since the more proximate cause is itself in a causal complex that does not include the more ultimate cause, it can be a cause by virtue of that smaller causal complex. Thus, the letting go is a cause by virtue of the causal complex consisting of only the letting go, or by virtue of the causal complex consisting of the letting go and the falling, but the falling is a cause only by virtue of the causal complex consisting only of the falling.

4 General Properties of Causality

Domain knowledge about what kinds of eventualities cause what other kinds of eventualities is encoded in axioms of form (1). These are usually very specific to domains—e.g., flipping switches causes lights to go on. These are the most common sufficient conditions for causality. The predicate *cause* appears in the consequent.

A candidate for a *general* sufficient condition is the idea that every eventuality has a cause. The axiom would be stated as follows:

$$
\begin{aligned}
&(\forall e_2)[Rexists(e_2) \wedge eventuality(e_2) \\
&\quad \supset (\exists e_1)[Rexists(e_1) \\
&\qquad \wedge cause(e_1, e_2)]]
\end{aligned}
$$

It is not uncontroversial that we would want this axiom. Certainly, very often we have no idea of what the cause of something is.

There is not very much that can be concluded from mere causality, without any further details. That is, there seem to be very few axioms stating general necessary conditions for causality, in which *cause* is in the antecedent. One candidate relates causality and existence in the real world. If we were to state this in its strongest, monotonic form, we would use the predicate *causal-complex*:

$$
\begin{aligned}
&(\forall s, e)[causal\text{-}complex(s, e) \\
&\qquad \wedge Rexists(s) \\
&\quad \supset Rexists(e)]
\end{aligned}
$$

If s is a causal complex for an effect e and s really exists, then e really exists. When we state this using the predicate *cause*,

$$
\begin{aligned}
(2) \quad &(\forall e_1, e)[cause(e_1, e) \wedge Rexists(e_1) \\
&\qquad \supset Rexists(e)]
\end{aligned}
$$

the axiom is only defeasible, because it requires the rest of e's causal complex, the presumably true part, to be actually true.

Axiom (2) can be used with axiom (1) to show that specific causes occurring will cause their effects to occur.

Now we come to the question of whether *cause* should be transitive:

$$
\begin{aligned}
&(\forall e_1, e_2, e_3)[cause(e_1, e_2) \\
&\quad \wedge cause(e_2, e_3) \supset cause(e_1, e_3)]
\end{aligned}
$$

Let us analyze the question in terms of causal complexes. Suppose we know that an eventuality e_1 is a member of a set s_1, which is a causal complex for eventuality e_2, which is in a causal complex s_2 for eventuality e_3. It is possible that $s_1 \cup s_2$ is inconsistent. Shoham's example is that taking the engine out of a car (e_1) makes it lighter (e_2) and making a car lighter makes it go faster (e_3), so taking the engine out of the car makes it go faster. The problem with this example is that the union of the two causal complexes is inconsistent. A presumable eventuality in s_2 is that the car has a working engine. When it is consistent, then we can say that $s_1 \cup s_2$ is a causal complex for e_3.

$$
\begin{aligned}
&(\forall s_1, s_2, e_2, e_3)[causal\text{-}complex(s_1, e_2) \\
&\qquad \wedge e_2 \in s_2 \\
&\qquad \wedge causal\text{-}complex(s_2, e_3) \\
&\qquad \wedge consistent(s_1 \cup s_2) \\
&\quad \supset causal\text{-}complex(s_1 \cup s_2, e_3)]
\end{aligned}
$$

Let $ccf(e_1, e_2)$ be a causal complex by virtue of which e_1 causes e_2. Now suppose $cause(e_1, e_2)$ and $cause(e_2, e_3)$ are true. We can conclude that $causal\text{-}complex(s_1, e_2)$, $e_2 \in s_2$, and $causal\text{-}complex(s_2, e_3)$ are all true for some s_1 and s_2. If $s_1 \cup s_2$ is consistent, then by this axiom we can conclude

$$causal\text{-}complex(s_1 \cup s_2, e_3)$$

By the definition of *cause*, all the eventualities in $s_1 - \{e_1\}$ and $s_2 - \{e_2\}$ are presumably true, and e_2 is triggered by e_1 in the causal complex $s_1 \cup s_2$. Thus, we can identify e_1 as the cause in $s_1 \cup s_2$ for e_3. This means that $cause(e_1, e_3)$ holds. We have established the rule

$$
\begin{aligned}
&(\forall e_1, e_2, e_3)[cause(e_1, e_2) \\
&\qquad \wedge cause(e_2, e_3) \\
&\qquad \wedge consistent(ccf(e_1, e_2) \cup ccf(e_2, e_3)) \\
&\quad \supset cause(e_1, e_3)]
\end{aligned}
$$

If we take $\neg consistent(ccf(e_1, e_2) \cup ccf(e_2, e_3))$ to be the abnormality condition for the axiom, then we can state the defeasible rule

$$
\begin{aligned}
&(\forall e_1, e_2, e_3)[cause(e_1, e_2) \\
&\qquad \wedge cause(e_2, e_3) \wedge \neg ab_1(e_1, e_2, e_3) \\
&\quad \supset cause(e_1, e_3)]
\end{aligned}
$$

That is, causality is defeasibly transitive.

This rule is used heavily in commonsense reasoning for deducing causal chains between an effect and its ultimate cause.

Several writers have argued against the transitivity of causality on the basis of examples like

The cold caused the road to ice over.
The icy road caused the accident.
* The cold caused the accident.
(Hart and Honoré, 1985)

This example is not very compelling. Certainly the starred sentence is not about direct causes, but it *is* about indirect causes. Very frequently, newspapers attribute some number of deaths to a heat wave, even though the direct causes might be a variety of medical and other conditions. Direct causality is of course not transitive.

Shoham (1990) believes that *cause* should be antisymmetric and antireflexive. Two eventualities cannot cause each other, and an eventuality cannot cause itself. I am not sure of this. If two books are leaning against each other and keeping each other in an upright position, it seems quite reasonable to say that the one book's condition of leaning toward the other is causing the other's condition of leaning toward the first. If this instance of symmetry is allowed, then reflexivity follows. Each book's position is causing its own position, though not directly. It is possible to view this as a reasonable statement, in spite of its initial implausibility.

There is a strong temptation in writing about causality to confine one's self to events, that is, changes of state. This is surely not adequate, since we would like to be able to say, for example, that the slipperiness of the floor caused John to fall, and that someone spilling vegetable oil on the floor caused the floor to be slippery. A state like slipperiness can be both a cause and an effect. Nevertheless, there is something to the temptation. Whether a state is a cause or effect will not normally become an issue unless there is the possibility of a change into or out of that state. That requires that both the state and its negation be possible, and the focus on events could be seen to result from this requirement.

5 Other Causal Predicates

There are a number of words in English that involve causality, such as "prevent", "enable", "let", "help", and "maintain". Ortiz (1999a) has examined these words and defined the concepts in his framework. I will present what I believe is a simpler and more natural characterization of the concepts and relate them to his approach.

In brief, in his treatment of counterfactuals and causality, Ortiz (1999b) defines three relations among events. The first is "counterfactual implication" and accommodates the selective lifting of constraints and the frame problem. The second is the stronger relation of "strongest antecedent condition", which subsumes counterfactual implication but also takes care of the problem of pre-emption, where, for example, if A had not caused E, then B would have. The third relation is *cause*, which is strongest antecedent condition but in addition excludes some special relations such as "method-of".

"Prevent": The natural definition of "prevent" is simply "cause not":

$$(\forall e_1, e_2)[prevent(e_1, e_2) \\ \equiv cause(e_1, \neg e_2)]$$

Ortiz's definition is close to this:

$$occurs(e_1, t_1) - SAC- > \\ \neg holds(possible(occurs(e_2, t_2)), t_2)$$

There are two things to note about this expression. He uses "strongest antecedent condition" rather than causality, and what is negated is the holding of the possibility of the occurrence of e_2. It may be that the former is due to the latter. The latter is possibly due to his reluctance to have negative events such as $\neg e_2$. But certainly if e_1 causes $\neg e_2$ and e_1 occurs, then (defeasibly) it is not possible for e_2 to occur, by axiom (1). The same defeasibility accompanies *prevent* that accompanies *cause*. If one accepts Ortiz's intuition that "method-of" relations don't count as instances of causality, then he has made the definition of *prevent* too weak. There should be a "not-method-of" restriction on *prevent*, or else it would be the case, for example, that

John's sitting prevented him from standing up.

If one admits negative events in one's ontology, as I do, then there is no reason not to simply adopt the "cause not" definition of *prevent*.

"Enable": One common intuition is that e_1 enables e_2 just in the case that if e_1 didn't happen, that would cause e_2 not to happen. That is, for e_1 to enable e_2 is for $\neg e_1$ to cause $\neg e_2$.

$$(\forall e_1, e_2)[enable(e_1, e_2) \\ \equiv cause(\neg e_1, \neg e_2)]$$

Another common intuition about enablement is that the enabling conditions for some effect are the eventualities in the causal complex for the effect other than the cause. In turning on a light, the circuits being intact and the bulb not being burnt out are enabling conditions, while flipping the switch is the cause.

$$(\forall e_1, e_2)[enable(e_1, e_2) \\ \equiv (\exists s)[causal\text{-}complex(s, e_2) \\ \wedge e_1 \in s \wedge \neg cause(e_1, e_2)]]$$

If e_1 is a presumable eventuality in a causal complex s, then because it is change-relevant, toggling it to $\neg e_1$ while changing nothing else in s will toggle e_2 to $\neg e_2$. Thus, $(s - \{e_1\}) \cup \{\neg e_1\}$ contains a causal complex for e_2. Only e_1 in s is toggled, and because e_1 is presumable, $\neg e_1$ is not presumable. Thus, $\neg e_1$ constitutes a cause of $\neg e_2$. The two definitions of *enable* are compatible.

Ortiz (1999a) rewrites enablement in the following manner:

$$occurs(e_1, t_1) - SAC- > \\ holds(possible(occurs(e_2, t_2)), t_2)$$

He paraphrases this as "if $[e_1]$ had not occurred then $[e_2]$ would not have been possible", very similar to my formulation. Again, Ortiz has avoided the use of negative events by basing it on counterfactual implication in which negation is implicit.

Toward the end of his discussion of enablement, he notes that it is natural to attribute enablement only to conditions that are "part of some assumed background knowledge". Whereas this is an informal remark at the end of his development, I have made this a key feature of the development by the distinction between causal complexes and the predicate *cause*.

"Let": The simplest candidate for a meaning for "let" is "not cause not":

$$(\forall e_1, e_2)[let(e_1, e_2) \\ \equiv \neg cause(e_1, \neg e_2)]$$

Eventuality e_1 happening lets eventuality e_2 happen if and only if e_1 happening does not cause e_2 not to happen. Similarly, an agent e_1 lets eventuality e_2 happen if and only if e_1 does not cause e_2 not to happen.

Ortiz defines "let" as $\neg prevent(e_1, e_2)$ where the manner of the not preventing is refraining. The essential condition on refraining is that e_2 be possible. It is reasonable to constrain uses of "let" to cases where e_2 is possible. In all likelihood, right now, Noam Chomsky is not causing you not to read this exposition, but it seems strange to say that he is letting you read it. In most cases, it would not be possible for him to keep you from reading it. It might be argued however that the possibility of e_2 results from a pragmatic principle. We don't make assertions unless their negations are also possible. It would be true that Noam Chomsky is letting you read the exposition, but uninformative, since he is not in a position to prevent you.

"Help": We can speak of an agent a helping another agent b achieve a goal e, or we can speak of a set of events s helping another event e to come about. The former can be defined in terms of the latter.

I would argue that one set of events helps, in the latter sense, another event come about if that set is a subset of a causal complex for the other event.

$$(\forall s_1, e)[help(s_1, e) \\ \equiv (\exists s)[subset(s_1, s) \\ \wedge causal\text{-}complex(s, e)]]$$

Ortiz (1999a) defines *help* in terms of resources. An event e_2 requires certain resources, and e_1 helps e_2 if it enables e_2 to occur with a limited use of resources. Resources can be such things as time, energy, steps, money, or number of agents required. The two definitions are not so far apart if we think of using a bit of the resources as an eventuality that is part of the causal complex. Then s_1 would be acts of utilizing the resources to achieve e, and s_1's occurring would make the achievement of e require less resources, namely, only $s - s_1$ rather than s. "Hinder" can be defined as helping not.

"Maintain": The simplest possible definition of *maintain* is "cause not change":

$$(\forall e_1, e)[maintain(e_1, e) \\ \equiv cause(e_1, \neg change(e, \neg e))]$$

This relates two eventualities. We could then define the notion of an agent a maintaining an eventuality e in terms of *maintain*:

$$(\forall a, e)[maintain_1(a, e) \equiv (\exists e_1)[agent(a, e_1) \wedge maintain(e_1, e)]]$$

Ortiz analyzes maintaining by an agent a of an eventuality e as a's successfully hindering $\neg e$. That is, an agent must be expending resources, at least intermittantly, to hinder $\neg e$, and all the while e must obtain.

For both his definition and ours, it is not enough that e remains true. There must be some causal role on the part of the agent or the agent's actions. For Ortiz, this is spelled out in terms of resource consumption. My characterization is more neutral, making use only of the predicate *cause*. Of course, Ortiz's characterization would imply mine.

6 Summary

Causality cannot be defined in terms of necessary and sufficient conditions. But we can specify a number of necessary conditions and a number of sufficient conditions. What I have tried to do in this paper is explicate some of these conditions.

The key move has been to distinguish between *causal complexes*, which can be reasoned about monotonically but can rarely be completely explicated, and *causes*, which constitute the bulk of our causal knowledge but must be reasoned about defeasibly. This has led to more precise characterizations of some of the properties of causality, such as transitivity, and to definitions of other causal predicates that preserve the simplicity of our initial intuitions.

Acknowledgements

The author has profited from discussions with Lauren Aaronson, Pat Hayes, and Charlie Ortiz about this work, and from the suggestions of two anonymous referees. The research was funded in part by the Defense Advanced Research Projects Agency under Air Force Research Laboratory contract F30602-00-C-0168 and in part by the National Science Foundation under Grant Number IRI-9619126 (Multimodal Access to Spatial Data).

References

[1] Hart, H. L. A., and Tony Honoré, 1985. *Causation in the Law*, Clarendon Press.

[2] Hobbs, Jerry R. 1985. "Ontological Promiscuity." *Proceedings, 23rd Annual Meeting of the Association for Computational Linguistics*, pp. 61-69. Chicago, Illinois, July 1985.

[3] Lewis, David K., 1973. *Counterfactuals*, Harvard University Press, Cambridge, Massachusetts.

[4] Mackie, John L., 1993. "Causes and Conditions", in E. Sosa and M. Tooley, Eds., *Causation*, Oxford University Press, pp. 33-55.

[5] Ortiz, Charles L., 1999b. "Explanatory Update Theory: Applications of Counterfactual Reasoning to Causation", *Artificial Intelligence*, Vol. 108, Nos. 1-2, pp. 125-178.

[6] Ortiz, Charles L., 1999a. "A Commonsense Language for Reasoning about Causation and Rational Action", *Artificial Intelligence*, Vol. 111, No. 2, pp. 73-130.

[7] Pearl, Judea, 2000. *Causality*, Cambridge University Press.

[8] Shoham, Yoav, 1990. "Nonmonotonic reasoning and Causation", *Cognitive Science*, Vol. 14, pp. 213-252.

[9] Shoham, Yoav, 1991. "Remarks on Simon's Comments", *Cognitive Science*, Vol. 15, pp. 301-303.

[10] Simon, Herbert A., 1952. "On the Definition of the Causal Relation", *The Journal of Philosophy*, Vol. 49, pp. 517-528.

[11] Simon, Herbert A., 1991. "Nonmonotonic reasoning and Causation: Comment", *Cognitive Science*, Vol. 15, pp. 293-300..