Causality and Modality: The Case of "would"

Jerry R. Hobbs
Information Sciences Institute
University of Southern California
Marina del Rey, California

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 the notion of "causal complex" is introduced for a complete set of events and conditions necessary for the causal consequent to occur, and the term "cause" is used for the makeshift, nonmonotonic notion we require for everyday tasks such as planning and language understanding. Like all interesting concepts, neither of these can be defined with necessary and sufficient conditions, but they can be more or less tightly constrained by necessary conditions or sufficient conditions. The issue of how to distinguish between what is in a causal complex from what is outside it is discussed, and within a causal complex, how to distinugish the eventualities that deserve to be called "causes" from those that do not, in particular circumstances. We then examine one particular modal, the word "would", from the standpoint of its underlying causal content.

1 Causal Complexes

1.1 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. The second step is determining what we should mean by the predicate *cause*, as it appears in commonsense reasoning and lexical semantics.

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 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. The first question is dealt with in this section, and the second in Part 4. Finally the notion of *cause* is used in the analysis of the modal "would".

It should be noted at the outset that one of the aims 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. We would like to be able to say that the lack of a signature causes a contract to be invalid.

Much research in AI begins with simple intuitions about a phenomenon, but then problems are encountered, and by the time they are overcome, the formal treatment is quite complex. My aim in this work is to frontload the complexity, so that the axiomatizations that result at the end preserve the original simplicity of the intuitions.

Before beginning the formal treatment of these questions, I will describe the notational conventions used in this paper. The notation and ontology of Hobbs (1985a) 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 or state 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.

1.2 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 in this paper.) 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 what may be thought of as causal constraints and what may be thought of as noncausal, but I will not distinguish these a priori, since knowledge of the constraints is not given to us through intuition but is the hard-won result of careful investigation.

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_1. Formally,

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

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.

There is not necessarily a unique closest world w_2.

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

$$(\forall e_1, e)[\textit{change-relevant}(e_1, e) \\ \equiv (\exists w_1, w_2 \in Con(S, C))[\textit{closest-world}(w_2, w_1, e_1, C) \\ \wedge \neg[\textit{holds}(e, w_1) \equiv \textit{hold}(e, w_2)]]]$$

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 a closest world to w_1 when e_1 is turned on, then w_1 is a 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 by lifting it up again.

This axiom may be thought of as instructions for how to carry out an experiment. We want to know if a certain factor e_1 is relevant to a certain phenomenon e. We try to find situations in which e is absent and when we add the factor e_1, e is present, or in which e is present and when we add the factor e_1, e is absent.

Now we can propose the axiom

$$(\forall s, e)[\textit{causal-complex}(s, e) \supset (\forall e_1 \in s)[\textit{change-relevant}(e_1, e)]]$$

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.

1.3 Examples

Consider two simple models for this set of axioms. In the first, there is a set of light switches on a table, and a light bulb. When the right combination of

switches are toggled in the right ways, the light is on. Here the effect e is the light's being on, $\neg e$ its being off. There is an eventuality e_i for each switch s_i which is the condition of its being on; $\neg e i_i$ is the condition of its being off. A possible world is some combination of the switches being on or off. The condition e_1 of a switch s_1 being on is change-relevant to e, the light's being on, if and only if there is arrangement w_1 of switches in which switch s_1 is off, and we can turn it on and change the state e of the light. The possible world or arrangement w_2 of switches, in this case, is the arrangement in which s_1 is on and all other switches are as in w_1. The proposition $holds(e_1, w_2)$ is obviously true. The expression $\neg[holds(e, w_1) \equiv holds(e, w_2)]$ means the light changes when the switch is toggled. There are no constraints in C that mean that one switch can affect the state of another switch, so the only consequence for the switches of turning s_1 on is that s_1 is on. Thus, the only e_2 in $w_2 - w_1$ is e_1 itself, and e_1 follows trivially from $(w_1 \cap w_2) \cup C \cup \{e_1\}$, while e_1 does not follow from $(w_1 \cap w_2) \cup C$ since switches cannot influence each other. Thus the eventuality e_1 is change-relevant to e, and is therefore not ruled out as a part of some causal complex for e.

For the second example, 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—$upright'(e_i, d_i)$—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. That has not been done here.

1.4 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. Or consider a person hammering a nail. The person's arm reaching the end of its trajectory, the head of the hammer striking the head of the nail, and the nail beginning its motion into the surface are all simultaneous but 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 suggested by nonmonotonic logic. If an axiom or constraint in C allows us to draw a conclusion from e_1 to something causally prior to it, then we must be able to disable that axiom somehow, or we could reason backwards about causally prior conditions. One way to do this is to consider the axiom to be defeasible. In nonmonotonic logic this is done by including $\neg ab_i$ predications in the antecedent of a rule: $P \wedge \neg ab_i \supset Q$. Equivalently, in "Interpretation as Abduction" (Hobbs et al., 1993), most axioms are assumed to be of the form $P \wedge etc_i \supset Q$. Thus, we can change $\neg e_1$ to e_1 and prevent back-propagation of its effects by changing C into C' the following way: Suppose the constraints are expressed in disjunctive normal form. Then for every constraint that contains a negative occurrence of e_1 (or $holds(\neg e_1, w)$) and a positive occurrence of a causally prior eventuality, disjoin an abnormality predication to it—$ab_i(\ldots)$. For example, we change the rule

$$\text{bang} \supset \text{fire (i.e., } \neg\text{bang} \vee \text{fire)}$$

into

$$\text{bang} \wedge \neg ab_i(\ldots) \supset \text{fire (i.e., } \neg\text{bang} \vee ab_i \vee \text{fire)}$$

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.

Thus, from a set of constraints C and an eventuality e_1, we can define a new set of constraints in which all the constraints that would allow us to

draw conclusions about causally prior eventualities are made defeasible. It is assumed that the constraints are expressed in disjunctive normal form, and $disjunct\text{-}in(e, a)$ means that e is a top level disjunct in the expression a.

$$
\begin{aligned}
&Defeas(C, e_1) = \\
&\quad \{a \vee ab_a \mid a \in C \wedge disjunct\text{-}in(\neg e_1, a) \\
&\qquad \wedge (\exists e_0)[disjunct\text{-}in(e_0, a) \wedge causally\text{-}prior(e_0, e_1)]\} \\
&\quad \cup \{a \mid a \in C \wedge \neg[disjunct\text{-}in(\neg e_1, a) \\
&\qquad \wedge (\exists e_0)[disjunct\text{-}in(e_0, a) \wedge causally\text{-}prior(e_0, e_1)]]\}
\end{aligned}
$$

By allowing possible worlds consistent with this modified set of constraints, we can eliminate inferences, for example, from the bang to the firing, and consequently from the bang to the death.

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

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

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

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

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.

1.5 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.

1.6 Causality and Implication

There is a problem that Pearl does not address at all and that Ortiz addresses but bypasses by stipulation, that I also do not have a solution to: What principled ways are there to distinguish between causal connections and mere implicational connections? Clyde's being an elephant implies that Clyde is a mammal, but does not cause it. A stapler's being on a piece of paper causes the paper not to blow away, but it only implies that the paper is under the stapler. John's flipping a switch causes a light to go on, but John's flipping a switch only implies John *turned* the light on; they are two different descriptions of the same event.

Nevertheless, the two notions are closely related, as evinced by the fact that "because" is used to convey either of them. My view is that implication is a kind of "washed-out" causality. It is causality applied to the informational domain. Another take on the relation is that it is a variety of metonymy. If P implies Q, then for someone to think P causes them to think Q.

2 Counterfactuals

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

> If John had read the driver's manual, he would have passed the exam.

It is distinguished in English by the use of the otherwise rare subjunctive mode in the antecedent and the modal "would" in the consequent.

> If John were a millionaire, he would retire.

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 reading the driver's manual would cause him to pass the exam, and John's being a millionaire would cause 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 the definition of *closest-world* attempts to capture. 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.

My 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", causative verbs, the subjunctive mode, and, in the latter part of this paper, modal auxilliaries such as "would", within a framework of interpreting discourse by abduction.

The issue is what counts as evidence. My 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.

However useful a tool for discerning causality the counterfactual is, it does not help us with the problem of distinguishing between causality and mere implication. It is a true, implicational fact that chimpanzees are not monkeys. The following counterfactual sentence is perfectly fine, and it is based on this implicational relation:

> If Bonzo were a monkey, he wouldn't be a chimpanzee.

A classic conundrum in philosophical treatments of causality, especially those based on counterfactuals, is the problem of pre-emption. Suppose Adam and Ben both want to murder Chuck. Before Chuck walks off into the desert, Adam secretly drills a tiny hole (H) in Chuck's canteen so it will be empty (E) by the time he needs a drink and he will die (D) of thirst. Independently Ben poisons the water in the canteen (P) so when Chuck takes a drink he will die (D). Chuck walks off into the desert and dies of thirst. Did the hole H cause the death D? This is a problem for counterfactual approaches to causality because if Adam hadn't drilled the hole, Chuck still would have died, of poisoning by Ben. The hole pre-empted the poisoning.

In the framework of causal complexes, this situation presents no particular problems. There is a causal complex including H in which E is the effect. There is a causal complex including H and E in which D is the effect. There is a causal complex including P and $\neg E$ in which D is the effect. Since H occurred, the conditions for P's causal complex did not obtain, and that causal complex was thus not what resulted in D. Adam murdered Chuck. Ben only attempted to murder Chuck. (The movie "Gosford Park" is based on exactly this premise.)

3 Probability and Causality

When you flip a coin, you say that there is a 50% probability that it will come up heads. But you say this because you are ignorant of all the conditions that cause the outcome to be what it is, such as the distribution of mass inside the coin, the air currents, the force with which the coin is flipped, the distance it falls, and so on. If we knew all of these, and knew all the

relevant physical laws, we could predict the result of flipping the coin with certainty. The reason there is a 50% probability of the coin coming up heads is that there is a 50% probability of those hidden conditions being such that a heads will result.

Now consider a causal complex s for an effect e, and consider a subset s_1 of s. For simplicity, suppose s is the only causal complex that will cause e. We can talk of the probability of s_1 causing e. It is simply the probability that all the conditions in $s - s_1$ will be true.

Thus, the notion of a set s_1 of events causing an effect e with some probability can be reduced to the joint probability of the eventualities in the set $s - s_1$ occurring, within the development of causality already presented.

None of this should be taken to deny the utility of probabilistic approaches to causality. On the contrary, such approaches provide a means of reasoning about causality at a granularity intermediate between the full rigor of causal complexes and the defeasible reasoning in terms of the predicate *cause*.

4 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 how a car works, but they know to turn the key in the ignition to start it. 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 will 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.

An extreme example of these criteria is when we say that a certain virus causes influenza. Perhaps no more than one out of a million viruses actually cause damage. That is, the other conditions that make up the causal complex resulting in influenza, such as the failure of the lymphocytes to destroy the virus, are highly improbable. Yet that one virus's invasion of the cell is the highly consequential and potentially manipulable triggering causal element of the causal complex that results in influenza.

The predicate *cause* thus implies but is stronger than *causally-involved*:

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

Moreover, if we have a predicate *presumable*, meaning that its argument is presumably true, or presumably really exists, then we can state the following:

$$(1)\quad (\forall e_1, e)[\mathit{cause}(e_1, e) \supset (\exists s)[\mathit{causal\text{-}complex}(s, e) \wedge e_1 \in s \wedge (\forall e_2 \in s - \{e_1\})[\mathit{presumable}(e_2)]]]$$

If e_1 "causes" e, then e_1 is in a causal complex for e, the rest of which is presumably true.

Recall that conjunctions of eventualities are eventualities too, so that this use of *cause* covers the case of multiple causes as well. For example, to start a bonfire, we first pour starter fluid on the wood and then strike a match. The cause is the conjunction of these two actions and their temporal order.

The above axiom is not quite right. Consider again the example of the causal complex for the vase shattering that consists of two eventualities, letting go of the vase and the vase falling. Neither the letting go nor the falling is presumable. But we would want to call the letting go a cause, all by itself. Once it happens, the shattering will happen. Thus, we would like to eliminate as causes those eventualities that will occur anyway when all the causes and presumable eventualities occur. We will say those eventualities are not themselves causes, but are "triggered" by the causes within the causal complex, where we define *trigger* as follows:

$$(\forall e_1, s, e_2)[\mathit{trigger}(e_1, s, e_2) \equiv (\exists s_1)[s_1 \subset s \wedge e_2 \notin s_1 \wedge \mathit{causal\text{-}complex}(s_1, e_2) \wedge (\forall e_0 \in s_1)[\mathit{presumable}(e_0) \vee e_0 = e_1 \vee \mathit{trigger}(e_1, s, e_0)]]]$$

That is, eventuality e_1 in causal complex s triggers eventuality e_2 if and only if there is a proper subset s_1 of s not including e_2, s_1 is a causal complex for e_2, and every eventuality in s_1 is either presumable, is e_1 itself, or is triggered by e_1. This definition is recursive rather than circular since s_1 is smaller than s.

Then axiom (1) becomes

$$(\forall e_1, e)[\mathit{cause}(e_1, e) \supset (\exists s)[\mathit{causal\text{-}complex}(s, e) \wedge e_1 \in s \wedge (\forall e_2 \in s - \{e_1\})[\mathit{presumable}(e_2) \vee \mathit{trigger}(e_1, s, e_2)]]]$$

If e_1 is a cause of e, then e_1 is in a causal complex whose other members are either presumable or triggered by e_1.

I won't explicate the predicate *presumable* except to say that if an eventuality is presumable, its negation is not.

$$(\forall e)[\mathit{presumable}(e) \supset \neg \mathit{presumable}(\neg e)]$$

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

$$(\forall\, s)[\ldots \text{some long characterization of } s \ldots \\ \supset\ (\exists\, e)[q'(e, \ldots) \wedge \mathit{causal\text{-}complex}(s, e)]]$$

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 our axioms will be as follows:

$$(2) \quad (\forall\, e_1, x)[p'(e_1, x) \supset (\exists\, e)[q'(e, x) \wedge \mathit{cause}(e_1, e)]]$$

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

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

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.

Some philosophers have argued for the existence of "singular causation", that is, a specific instance of one event causing another without it being in any way an instance of a general causal principle. In Shoham's formulation (1990), there can be no causation without the presence of a general rule, since he lacks an explicit predicate for *cause*. This approach admits singular causation, although my feeling is that it does not occur; to recognize causality is to recognize a causal regularity.

Shoham 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 restate axiom schema (2) as follows:

$$(2') \quad (\forall\, e_1, x)[p'(e_1, x) \wedge \neg \mathit{ab}_i(e_1, x) \\ \supset\ (\exists\, e)[q'(e, x) \wedge \mathit{cause}(e_1, e)]]$$

where

$$(\forall e_1, x)[ab_i(e_1, x) \equiv$$
$$(\forall e, s)[\textit{causal-complex}(s, e) \wedge e_1 \in s \wedge q'(e, x)$$
$$\wedge \textit{presumable}(s - \{e_1\}) \supset (\exists e_2)[e_2 \in s - \{e_1\} \wedge \neg \textit{Rexists}(e_2)]]]$$

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 $(2')$ exactly matches the nonmonotonic causal rules that Shoham uses.

Mackie (1993) proposes his "at least INUS" definition for causality. In the INUS condition, C is a cause for E just in case there are an X and a Y such that

neither C nor X entails E
$C \wedge X$ does entail E
E does not entail Y
E does entail $C \vee Y$

That is, C is an Insufficient but Necessary condition of an Unnecessary but Sufficient condition for E. In his "at least INUS condition", X and/or Y can be empty. We thus have $(C \wedge X) \vee Y \supset E$. In the present development, C corresponds to the cause, X to the remainder of the causal complex, and Y to the disjunction of the other causal complexes for E.

Mackie also discusses the "causal field", those things in the causal complex that are assumed to be true, and thus do not count as a cause. For us, this corresponds to the presumable portion of the causal complex. Shoham similarly distinguishes between the foreground and the background in what I would call the causal complex.

Suppes (1970) proposes a probabilistic account of causation. Briefly, C is a "prima facie" cause of E if $P(C) > 0$ and $P(E \mid C) > P(E)$. This is consistent with the present account. The probability of E is the probability of one of its entire causal complexes holding. C is part of one of the causal complexes. If C has been isolated as a cause, then its occurrence is not entirely predicted by the rest of the causal complex. If C holds, then that increases the probability that the entire causal complex holds.

Suppes further goes on to eliminate "spurious" causes, that is, those causes that are in fact themselves caused by deeper causes. The two relevant conditions for B to be a spurious cause of E are

$$P(E \mid B, C) = P(E \mid C)$$
$$P(E \mid B, C) \geq P(E \mid B)$$

This eliminates causal chains, just as I do *within causal complexes* by ruling out triggered events as causes.

However, 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. But I would not want to eliminate the possibility of calling the falling a cause of the shattering, just because something caused *it*. In my framework this is handled 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.

I have said that most causal knowledge used in planning, explanation, prediction, and interpreting causal statements is expressed in terms of the predicate *cause*. By contrast, the causal knowledge used in diagnosis would more likely be knowledge about causal complexes, since we do diagnosis when the normal or usual or presumable operation of things breaks down. Although we don't know everything that is in a causal complex, we do know specific things that are, and this type of knowledge is expressed in axioms of the following form:

$$(\forall e, x)[q'(e, x)$$
$$\supset (\exists s, e_1)[\mathit{causal\text{-}complex}(s, e) \wedge p'(e_1, x) \wedge e_1 \in s]]$$

That is, when a q-type event occurs, there is a p-type event in its causal complex.

5 General Properties of Causality

Domain knowledge about what kinds of eventualities cause what other kinds of eventualities is encoded in axioms of form (2). 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:

$$(\forall e_2)[\mathit{Rexists}(e_2) \wedge \mathit{eventuality}(e_2)$$

$$\supset (\exists e_1)[Rexists(e_1) \wedge cause(e_1, e_2)]]$$

It is not uncontroversial that we would want this axiom. Certainly, very often we have no idea of what the cause of something is. For most of human history, people had no idea what caused the wind, although they may have had theories about it. There is in commonsense reasoning, one can argue, the scientifically erroneous notion of an "agent", an entity capable of initiating causal chains. Either agents could appear as the first argument of the predicate *cause*, or some primitive action on the part of agents, such as $will(a)$, would initiate the causal chain and these actions would be exempt from the axiom.

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. I will mention two.

The first 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*:

$$(\forall s, e)[causal\text{-}complex(s, e) \wedge Rexists(s) \supset Rexists(e)]$$

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*,

(3) $(\forall e_1, e)[cause(e_1, e) \wedge Rexists(e_1) \supset Rexists(e)]$

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

Axiom (3) can be used with axiom (2) to show that specific causes occurring will cause their effects to occur.

Another general necessary condition for causality is its relationship to time. Effects can't happen before causes:

$$(\forall e_1, e)[cause(e_1, e) \supset \neg before(e, e_1)]$$

Similarly,

$$(\forall e_1, e)[cause(e_1, e) \supset \neg causally\text{-}prior(e, e_1)]$$

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

$$(\forall e_1, e_2, e_3)[cause(e_1, e_2) \wedge cause(e_2, e_3) \supset cause(e_1, e_3)]$$

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.

$$\begin{aligned}(4)\quad &(\forall\, s_1, s_2, e_2, e_3)[\mathit{causal\text{-}complex}(s_1, e_2) \wedge e_2 \in s_2 \\ &\quad \wedge \mathit{causal\text{-}complex}(s_2, e_3) \wedge \mathit{consistent}(s_1 \cup s_2) \\ &\quad \supset \mathit{causal\text{-}complex}(s_1 \cup s_2, e_3)]\end{aligned}$$

Since

$$\begin{aligned}&(\forall\, e_1, e_2)[\mathit{cause}(e_1, e_2) \\ &\quad \supset (\exists\, s_1)[\mathit{causal\text{-}complex}(s_1, e_2) \wedge e_1 \in s_1]]\end{aligned}$$

we can define the function

$$\mathit{ccf}(e_1, e_2) = s_1$$

That is, $\mathit{ccf}(e_1, e_2)$ is a causal complex by virtue of which e_1 causes e_2.

Now suppose $\mathit{cause}(e_1, e_2)$ and $\mathit{cause}(e_2, e_3)$ are true. We can conclude that $\mathit{causal\text{-}complex}(s_1, e_2)$, $e_2 \in s_2$, and $\mathit{causal\text{-}complex}(s_2, e_3)$ are all true for some s_1 and s_2. If $s_1 \cup s_2$ is consistent, then by (4) we can conclude

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

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 $\mathit{cause}(e_1, e_3)$ holds. We have established the rule

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

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

$$(\forall e_1, e_2, e_3)[cause(e_1, e_2) \wedge cause(e_2, e_3) \wedge \neg ab_1(e_1, e_2, e_3) \supset cause(e_1, e_3)]$$

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; Ortiz, 1999b)

and

> John's leaving caused Sue to cry.
> Sue's crying caused her mother to be upset.
> * John's leaving caused Sue's mother to be upset.
> (Moens and Steedman, 1988; Ortiz, 1999b)

Neither of these examples is very compelling. Certainly the starred sentences are not about direct causes, but they *are* 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. And we can imagine Sue's mother complaining about the wide repercussions of John's actions—"Look what he did to me!" 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. Thus, the focus on events could be seen to result from this requirement. With the proper notion of "possible", we could state the following axioms:

$$(\forall e_1, e_2)[cause(e_1, e_2) \\ \supset possible(e_1) \wedge possible(\neg e_1) \wedge possible(e_2) \wedge possible(\neg e_2)]$$

6 Modality: The Case of "would"

6.1 An Example

The chief objection to basing a treatment of modality on causality is that causality is such a quagmire of difficulties in philosophy. These difficulties have not been solved here, but it has been possible to work around them to create a coherent and usable theory of causality. There are three reasons for this.

1. I have not attempted to *define* causality, or even causal complexes; rather I have written axioms capturing their principal characteristics, thereby constraining what a causal complex can be, without giving necessary and sufficient conditions.

2. The predicate *cause* has been defined in a way that makes its principal properties defeasible or nonmonotonic; yet there is a precise picture of how it relates to the notion of "causal complex" that underlies it.

3. I am assuming an "Interpretation as Abduction" framework, which uses a knowledge base of such defeasible axioms to arrive at interpretations of discourse. Essentially, one seeks the best proof of the explicit content of the text, where "best" is related in part to the reliability of the defeasible axioms used in the proof. Proofs are also better that make use of redundant information conveyed in different parts of the text; this encourages the linking of that information, as we will see in the example in this section and in a myriad of examples in the next section.

I will focus on the modal "would", since of all the modals, that conveys a causal relation in its purest form. Let us consider the example

I don't own a TV set. I would watch it all the time.

For the sake of exposition, let us simplify this to the slightly less idiomatic

(5) I don't own a TV. I would watch it.

The modal "would", like all modals, is with respect to a set of constraints c. We can build this into the predicate *would* by giving it a second argument—$would(e_4, c)$. In "I would watch it", my watching it is e_4 and c is the set of constraints that would result in my watching it. We can then reify the "would" situation and write $would'(e_3, e_4, c)$. This says that e_3 is the "would-ness" of situation e_4; we can think of e_3 as the hypothecality of e_4 with respect to constraints c. Then the causal content of "would" can be described by the axiom

$$cause'(e_3, c, e_4) \supset would'(e_3, e_4, c)$$

That is, if e_3 is the causal relation between some causing situation c and another, caused situation e_4, then e_3 is the "would-ness" property of e_4 with respect to the constraint c.

To interpret text (5) we need to assume our knowledge base has two specific rules involving causality.

$$cause'(e_3, e_2, e_4) \wedge bad\text{-}for(e_4, i) \wedge p'(e_2, i) \supset cause(e_3, e_1) \wedge not'(e_1, e_2)$$

$$own'(e_2, i, t) \supset cause'(e_3, e_2, e_4) \wedge use'(e_4, i, t)$$

The first is an axiom schema that says that if a situation causes you to do an action that is bad for you, that causes you not to bring about that situation. Don't do the cause if you don't want the effect. This will be instantiated below with *watch* instantiating the prediate variable p. The second rule says owning something causes you to use it.

Two more specific axioms are required. The first says that watching TV is bad for you.

$$watch'(e_4, i, t) \wedge tv(t) \supset bad\text{-}for(e_4, i)$$

The next axiom says that to use a TV set is to watch it.

$$tv(t) \wedge use'(e_4, i, t) \supset watch'(e_4, i, t)$$

One final axiom also involves causality. It says that one kind of information the adjacency of two sentences in discourse can convey is a causal relation between the eventualities described. The second sentence functions as an explanation of the first; the sentences are related by the "coherence relation" of Explanation.

$$cause(e_3, e_1) \supset CoherenceRel(e_1, e_3)$$

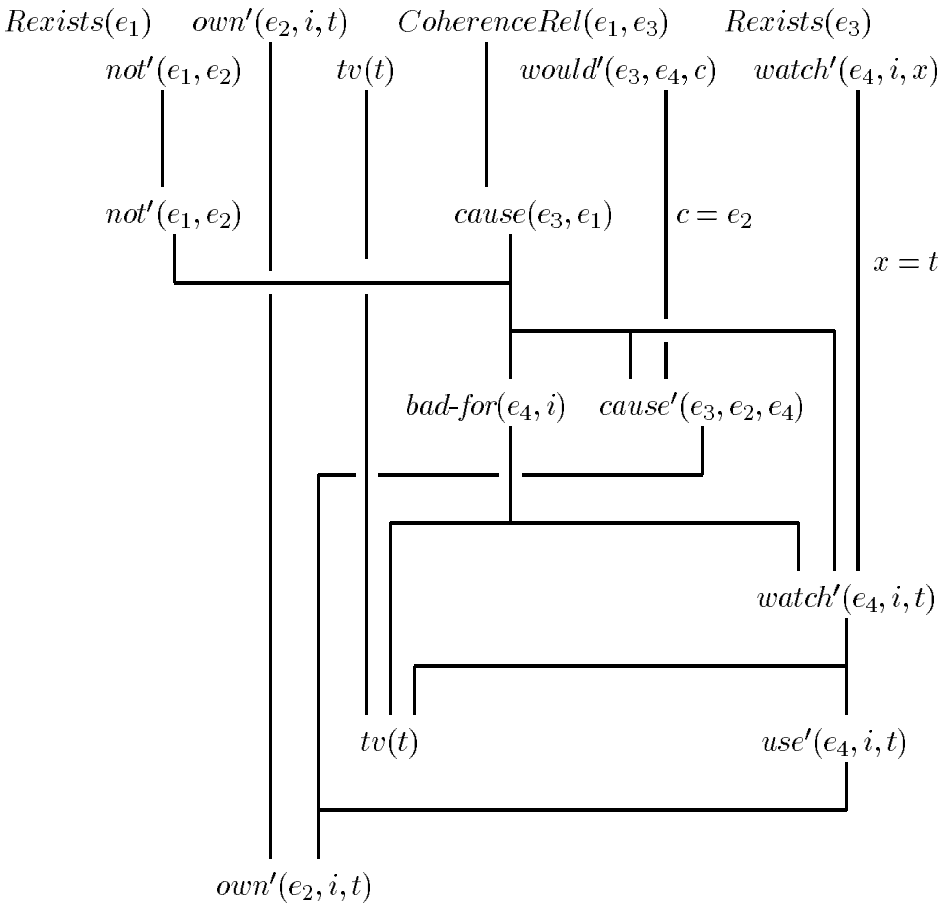

Figure 1: Interpretation of "I don't own a TV. I would watch it."

In the "Interpretation as Abduction" framework, one interprets a text by finding the best abductive proof of the logical form of the text. A proof is abductive if it allows assumptions. A proof is better insofar as it is shorter, makes fewer assumptions, uses more reliable axioms, and exploits

redundancy in the text. The logical form of the text is the conjunction of the logical forms of the sentences, conjoined with a *CoherenceRel* predication representing the information conveyed by their adjacency. The logical form of text (5) is

$$\begin{aligned} &Rexists(e_1) \wedge not'(e_1, e_2) \wedge own'(e_2, i, t) \wedge tv(t) \\ &\quad \wedge CoherenceRel(e_1, e_3) \\ &\quad \wedge Rexists(e_3) \wedge would'(e_3, e_4, c) \wedge watch'(e_4, i, x) \end{aligned}$$

That is, there exists in the real world the negation (e_1) of my owning (e_3) a TV set t, and that is related to the "would-ness" e_3 of my watching (e_4) something referred to as "it" (x), a very tortured paraphrase of (5). At this point, the constraints c have not yet been identified with e_2, the owning of a TV set, and "it" (x) has not yet been identified with the TV set.

The proof of this logical form given the above axioms is illustrated in Figure 1. We have to assume a hypothetical owning, the nonexistence of that owning, and the existence of the "would" property. Everything else can be proved from these assumptions. In the course of the proof, in order to get the best proof, the constraint c is identified with the hypothetical owning, and "it" is identified with hypothetical TV set.

For the purposes of this paper, what is most interesting about this example is the important role played by *cause* in the interpretation, and in particular, the role the causal content of "would" plays in recognizing the coherence of the text, i.e., the relation between the two sentences, and in determining the identity of the constraints associated with "would".

This example is also interesting because it has been posed as a challenge by Frank and Kamp (1997) for Discourse Representation Theory (DRT) approaches to coreference via accessibility conditions. Here the hypothetical "it" of the second sentence happily resolves to the nonexistent TV set embedded in negation in the first sentence. In the "Interpretation as Abduction" approach, the supposed accessibility conditions on pronouns are really ways of computing whether there are contradictory statements about the existence of entities in the real world, for purposes of ruling out certain coreference relations.. Here the resolution is possible because we have identified the nonexistent owning as the hypothetical cause implicit in the modal "would", and there is no contradiction.

6.2 Uses of "would"

The Data In the previous section we saw in one made-up example how a causal relation is implicit in a fragment of discourse containing "would". But if there really is such a close connection between "would" and causality, we should expect to see an explicit cause frequently when an effect is the complement of "would", because of the high degree of redundancy in discourse.

To determine if this is indeed the case, I examined 131 examples of the use of "would" in texts from seven different genres: a novel, Carson McCullers' *The Ballad of the Sad Cafe* (26 examples); business articles from the San Jose *Mercury-News* (20); articles on AIDS from *Science* magazine (9); Shakespeare's sonnets (23); transcripts of decision-making meetings in which three people are trying to make up a schedule for a visitor from a funding agency (30); and the lyrics of country and western songs (23). The aim was to get a broad coverage of types of discourse while limiting the examples to a manageable number.

Of the 131 examples, eleven, all from Shakespeare, involved the old sense of "would" meaning "want to", as in

> Tired with all these, from these would I be gone.

These were excluded from further examination.

Causes in Clauses In each of the remaining 120 examples, the effect is the eventuality described by the verb phrase in which "would" is an auxilliary. For each of these, I determined whether the cause was explicit in the surrounding discourse, and what syntactic or discourse structure conveyed this causal relation.

It turned out that in 66 of the 120 examples, or more than half, the cause was in a syntactically related part of the same clause. In 27 of these, the cause was the subject of the verb phrase. The subject described a hypothetical entity or situation whose existence would result in the effect.

(6) ... and to have sat around the property outside would have made a sorry celebration.

The causal relation is between sitting outside and the poor quality of the celebration. In the sentence

> ... the standard would allow intelligence agencies to spy on private companies and individuals.

the adoption, or existence, of the cryptography standard being discussed is the cause, the possibility of spying the effect.

It is interesting to note that in eight of the 27 examples the main verb of the verb phrase has causal content, as do "made" and "allow" in these two sentences.

A common causal pattern exhibited in these sentences is the causing of a functional property by a structural property. For example, in

(7) ... to provide generic products, particularly in software, wouldn't serve the clientele.

The cause is providing generic products, a structural description of how the company conducts its business. The effect is not serving the clientele, a higher-level functional property describing explicitly how the company relates to the outside world.

A very common type of structure-function causal relation is where the functional description is evaluative. The cause is a structural explanation and the effect is the functional property of being good or bad for some purposes. Examples (2) and (4) have this flavor. A purer example is

> It would obviously be nice to have data from more patients.

Having data from more patients is the structural cause whose effect is the niceness of the experimental study, that is, its validity or persuasiveness.

We will return to this point when considering the "would like" idiom below.

In the remaining 39 examples having the cause expressed in a syntactically related manner, the cause appears in an adjunct. In 20 of these cases, it appears in a subordinate clause with the subordinate conjunction "because", "every time", "if", "once", "were", or "when". All of these words carry at least partial causal content in that a causal relation implies them.

In

> If the rumors were right [that I still love you]
> Would you be here tonight?

the cause is my love for you and the effect is our being together, a common defeasible causal law in relationships.

The sentence

> Were he reading this article, he would have finished long ago.

exhibits a kind of causality commonly expressed in discourse – the activity is the cause of the end of the activity.

In 12 of the 39 examples, the cause is expressed in a prepositional phrase, with one of the prepositions "at", "by", "for", "from", "in" or "with", again words with possible partial causal content.

In the sentence

> The observed pattern was, however, what would be expected from sampling theory.

the effect, the expectation, is caused by one's knowledge of sampling theory. In the sentence

> At the mere mention of the words, her face would darken with shame.

the mentioning of the words is the cause of her face darkening.

In five of the 39 examples, the cause is in an infinitival adjunct, generally involving a kind of Volitional Reversal: if A causes B, then wanting B can cause one to want A. In the sentence

> And Miss Amelia would not leave him by himself to suffer with this fright.

his being along causes him to suffer a fright, so wanting him not to suffer a fright causes her to want him not to be alone.

In one example the cause is in a gerund.

> ... Miss Amelia, being rich, would not go out of her way to murder a vagabond for a few trifles of junk.

The cause is being rich and the effect is not robbing small amounts.

The final example of the 39 is the following sentence:

> ... the London routes sale is another major stumbling block before a Kerkorian-union combination would assume control of TWA.

The non-sale of the London routes is the cause of the combination not assuming control. Here the cause is in the main clause and the effect in the subordinate clause.

Causes in Discourse 66 of the 120 examples had the causes in a position syntactically related to the effect. In eleven more of the 120 examples of "would", the cause is in a clause or sentence adjacent to the effect. In five of these eleven cases, the cause is described as a possibility in an immediately preceding clause or sentence.

> Amino acid substitutions in B cell (4, 24) or cytotoxic T cell (25) epitopes can abrogate immune recognition. These escape variants would have a survival advantage that may facilitate their transmission.

The first sentence describes a possible ("can") situation in which variants of a virus escape recognition by the immune system. The second describes the consequences of this.

In two examples, the cause is described as a desire in an immediately previous clause or sentence.

> I wanna make you mine forever.
> There's nothin' on this earth I would not do.

This is another case of volitional reverse. Doing good things for a woman causes her to commit herself to the man, so the belief goes. Thus, wanting the woman to commit herself causes the man to want to do good things.

The other four examples of the eleven involve a contrastive relation between clauses or sentences. The sentence

> There were those who would have courted her, but Miss Amelia cared nothing for the love of men ...

exhibits a Violated Expectation pattern (Hobbs, 1985b). The second clause contains the cause for the expected effect in the first clause *not* to occur. In the sentence

> We have seen several indications that our recent price reductions have resulted in sales we would otherwise not have made ...

the cause is the price reductions, and the violated expected effect is not making sales.

Cognitive, Communicative, and Habitual Contexts Twenty examples of "would" in the data were in the complement of a cognitive or communicative verb—"think", "dream", "realize", "see", "presume", "wonder", "say", or "bet". For example, in the sentence

I've seen you smile more than I thought you would.

the eventuality described by the "would" verb phrase is "you smile [a certain amount]". It exists in the world of the speaker's cognition. Its existence there is an effect of some aspect of the speaker's cognitive state. In this particular case, it's a judgment about the causal connection between the degree to which a woman knows him and the degree to which she will smile at him.

In this case, the cause in general is some aspect of the cognizer's cognitive state, and the specific cause is not necessarily mentioned in the surrounding text. In five of the 20 examples, however, the cause *is* explicitly in the text. In the sentence

He said Kerkorian probably would discuss the offer with Icahn this weekend.

the effect, discussing the offer, is inside the world of the saying. But discussing the offer is a consequence of the offer itself. This is causal knowledge we have about social interaction. Thus, the cause is explict in the phrase "the offer".

Nine examples in the data involve the habitual use of "would".

Whatever happened to old-fashioned love,
The kind that would see you through?

In seven of the nine cases the time period during which the event was habitual is explicit, as it is here in the phrase "old-fashioned". In the other two it is implicit. In all these cases, we can say that the cause of the effect is some aspect of the time period in which it holds. Thus, in this example, in the old-fashioned era, people had strong values and stuck to their commitments, so their love would see you through hard times.

Hedges There are two types of expressions in the data where one could argue that the meaning of "would" has been washed out, and the modal is used merely as a hedge, generally to mitigate a request or an assertion of opinion.

The first type involves the use of "would" with one of the expressions "like", "rather", "be inclined", or "not mind". There are eight instances of this in the data. An example is

(8) I have five different subtasks I'd like my people to describe.

While these expressions are probably almost always interpreted as synonymous with "want" or "gimme", the mitigation is accomplished by the fact that "would" indicates that the thing wanted is a consequence of some stated or unstated cause.

The verb "like" is evaluative; "I like chocolate" is a positive evaluation of chocolate. We saw above that evaluative effects following from structural causes is a common sort of causal relation underlying uses of "would". We can see this case as another example. "This is good" and "I like this" are nearly synonymous. We can paraphrase (8) as something like "If my people describe my five different subtasks, then I will like that situation." In this sense, the complement of "like" provides the cause for the effect, the liking.

The second type involves the use of "would" with verbs of cognition, like "think", "guess", and "seem". This occurs six time in the data. Two examples are

(9) I would have thought you'd want to show him the stuff he likes best first, get him in a good mood.

and

(10) The rearrangement of the H3 loop would seem to be primarily a result of accommodating Tyr^{P105} of the peptide.

This also generally involves some mitigation. Utterance (9) is said in disagreement with the previous speaker, and the "would" mitigates the disagreement by making it seem dependent on causal factors beyond the control of the speaker. Indeed, three of the six examples involve disagreement. Sentence (10) is mitigated because this is scientific discourse, and, obvious as the conclusion might be from the spatial structural constraints on the entities, there is no specific experimental verification.

It is interesting to note that in four of the six examples, the complement of the cognitive verb also has "would". This is true in example (9). This means that the complement itself is an implicit causal statement, and the relevant cause there can be embedded in the thinker's thinking and function as a cause of the effect described in the complement. In example (9), the basic causal rule is that showing somebody something they like causes them to be in a good mood. By volitional reversal, since the participants in the conversation want the visitor to be in a good mood, they want to show him something he likes. We can then embed that in the speaker's cognition. His thinking that they want the visitor to be in a good mood causes him to think that they want to show him something he likes. Thus, the underlying cause of this cognitive state is explicit in the text.

Discussion Discourse is typically highly redundant. It therefore should not be surprising that the modal "would" conveys a causal relation that is often explicit in the surrounding text as well. In the 120 examples examined, this was true in 101 of them. For the listener or reader comprehending the discourse, making the connection between this cause and the complement of "would" is part of the job of comprehension. For this reason, we can say that understanding the underlying causal nature of the modal "would" is the beginning of understanding how it functions in discourse.

We saw above how the implicit causal nature of "would" enables its use as a mitigator. It conveys the sense that there is a cause beyond one's control that makes this true.

The word "would" is frequently used in describing unreal situations. This is by no means always the case in the data examined. A great many of the situations are unreal, but a great many are real, and a large number lie in between, where we don't know whether the situation is true or not. An example where the situation is clearly real is from the business news:

> He said he had no idea why the investigator would ask for the arbitration documents.

The speaker could just as correctly have said "why the investigator *asked* for the arbitration documents." The word "would" calls the reason for asking into question and thus shares information with the word "why". The reason "would" is used for describing hypothetical situations is precisely because it points to the existence of a hypothesis, i.e., the cause.

7 Summary and Future Directions

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. It also puts us in a good position to study modals such as "would" in terms of their underlying causal content.

Other modals can similarly be studied from the perspective of causality. For example, possibility is possibility with respect to a set of constraints. For something to be possible is for that set of constraints not to cause it not to occur. It would be interesting to determine how often in typical discourse the set of constraints is explicitly referred to nearby.

Much of the knowledge we use in producing and interpreting discourse is causal knowledge. It is an important enterprise to discover and characterize this knowledge, where the characterization should include common abstract patterns of causality. We have seen several such patterns in this study. Structure causing function is one of them. Volitional reverse and embedding in cognition are others. Another that figured prominently in the data was the causality resulting from tight systems of constraints. In addition, there were a number of specific domains where the analysis turned up a rich collection of causal relations. Several of the genres examined depended on complex causal relations among human relationships, emotions, and action, for example. Because of the underlying causal nature of modals, the examination of uses of modals in discourse provides an excellent window onto these causal systems.

This paper has been an attempt to lay out the basics of a coherent theory of causality and to begin an investigation of modality on that foundation. Rather than seeking to avoid all causal talk as many studies of modality do, a more promising approach is to embrace causality and use it as the basis of one's inquiries.

Acknowledgements

The author has profited from discussions with Lauren Aaronson, Cynthia Hagstron, Pat Hayes, David Israel, Srini Narayanan, and Charlie Ortiz about this work. 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] Frank, Anette, and Hans Kamp, 1997. "On Context Dependence in Modal Constructions", *Proceedings*, SALT 7, Stanford University, March 1997.

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

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

[4] Hobbs, Jerry R., 1985b. "On the Coherence and Structure of Discourse", Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

[5] Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. "Interpretation as Abduction", *Artificial Intelligence*, Vol. 63, Nos. 1-2, pp. 69-142.

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

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

[8] Moens, Marc, and Mark Steedman, 1988. "Temporal Ontology and Temporal Reference", *Computational Linguistics*, Vol. 14, No. 2, pp. 15-28.

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

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

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

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

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

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

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

[16] Suppes, Patrick, 1970. *A Probabilistic Theory of Causation*, North Holland Press.