Cognitive Science 33 (2009) 21–50

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2008.01002.x

A Causal Model Theory of the Meaning of *Cause*, *Enable*, and *Prevent*

Steven Sloman,[a] Aron K. Barbey,[b] Jared M. Hotaling[c]

[a]*Brown University*
[b]*Cognitive Neuroscience Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health*
[c]*Indiana University*

Received 31 July 2007; received in revised form 15 January 2008; accepted 25 January 2008

Abstract

The verbs *cause*, *enable*, and *prevent* express beliefs about the way the world works. We offer a theory of their meaning in terms of the structure of those beliefs expressed using qualitative properties of causal models, a graphical framework for representing causal structure. We propose that these verbs refer to a causal model relevant to a discourse and that ''A causes B'' expresses the belief that the causal model includes a link from A to B. ''A enables/allows B'' entails that the model includes a link from A to B, that A represents a category of events necessary for B, and that an alternative cause of B exists. ''A prevents B'' entails that the model includes a link from A to B and that A reduces the likelihood of B. This theory is able to account for the results of four experiments as well as a variety of existing data on human reasoning.

Keywords: Casual reasoning; Bayesian networks; Structural equations; Semantics of cause

The word *cause* has various senses both as a noun and as a verb (Wolff & Song, 2003). The focus of this paper is on how people reason about sentences that use *cause* as a verb. We propose a representation of the meaning of ''A causes B'' in terms of what such a sentence claims about the mechanism that relates A to B. A mechanism is an asymmetric relation that supports intervention (Pearl, 2000; Sloman, 2005; Woodward, 2003). If a mechanism exists from A to B (and not vice versa), then a sufficiently strong intervention on A under the right conditions would change the value of B but an intervention on B would not change the value of A. We contrast the representation of *cause* with that of *enable/allow* and *prevent*. The representations are not definitional in that mechanisms are themselves defined in terms of cause, so representing cause in terms of mechanism does not eliminate

Correspondence should be sent to Steven Sloman, Cognitive & Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912. E-mail: steven_sloman@brown.edu

all questions about the meaning of the words. Nevertheless, it does offer a handy way of distinguishing various causal verbs, theories of what people are claiming when they use causal verbs, and what conclusions they draw when they make causal statements. Our intent is not to define causation but to offer a theory of the everyday use of causal verbs in terms of how people represent structural relations among mechanisms.

We assume that a domain of discourse involves a more or less complicated system of objects, categories, and states of affairs (for simplicity, we refer to them as ''events''). Frequently, some events are causally related to other events. Our proposal is merely that to assert ''A causes B'' is to claim that A is a causal antecedent of B in that a mechanism—or set of mechanisms—exists from A to B. To assert ''A enables B'' is to claim that: (i) A is a causal antecedent of B; (ii) Some other event is an accessory conjunctive causal antecedent of B; and (iii) A is representative of a category that is necessary for B. To assert ''A prevents B'' is to claim A is an inhibitory causal antecedent of B. *Prevent* sometimes but not always entails an accessory variable. We express this theory more rigorously below using the causal Bayes nets framework.

One source of confusion in understanding the meaning of cause and associated words is that they are sometimes used in a specific and sometimes in a generic sense. One might ask, ''Was it the person, the gun, or the bullet that caused the death?'' If understood specifically, one makes a causal attribution by choosing one of the options and makes a case (e.g., the bullet and not the gun caused the death because the bullet was most proximal, cf. Hart & Honore, 1985). But one can also understand the question generically and answer, ''They were all causes'' because they were all linked in some way to the death. In this paper, we focus on the generic meaning of causal verbs.

One might believe the person, the gun, and the bullet were causes either because they were all part of a mechanism that produced the death or because they all satisfy the counterfactual ''if the cause had not been present, the death would not have occurred.'' A variety of theories exist to explain how people make causal attributions (Dowe, 2000; Halpern & Pearl, 2001; Hilton, 1990; Lewis, 1973, 1986, 2000; Mandel, 2003; Salmon, 1984; Walsh & Sloman, unpublished data). We have no intention of deciding among those theories here. Our aim is more modest: to offer a framework for distinguishing causal verbs and for reasoning about causal relations.

1. Mental model theory

Our work was motivated by an alternative theory of this domain proposed by Goldvarg and Johnson-Laird (2001). Their theory claims that causal verbs relating A and B license possibilities about the co-occurrence of the values of A and B (see Table 1). Goldvarg and Johnson-Laird denote events using capitals (e.g., A), their presence in lowercase (a) and their absence in lowercase with a tilde (∼a). They claim, for instance, that A causes B means that one of three possibilities holds: both a and b occurred, a did not but b did, or neither did. Notice that these are precisely the possibilities associated with the material conditional (see Kuhnmünch & Beller, 2005). Enable and prevent sentences differ in terms of the possibilities they allow.

Table 1
Explicit models for three causal verbs according to mental model theory: alternative possibilities (token causation) and types of possibilities (type causation)

A causes B:	
a	b
∼a	b
∼a	∼b
A enables B:	
a	b
a	∼b
∼a	∼b
A prevents B:	
a	∼b
∼a	b
∼a	∼b

The models depicted in Table 1 are the explicit models of the weak forms of the verbs. Goldvarg and Johnson-Laird (2001) also adopt the ''principle of truth'' (e.g., Johnson-Laird & Byrne, 2002), that people sometimes only use one model for reasoning, namely the first of each set. Goldvarg and Johnson-Laird posit that certain relations can also be strong such that a (for *cause*) or ∼a (for *prevents*) is necessary and sufficient for b. They also posit a temporal constraint: ''Given two states of affairs A and B, if A has a causal influence on B then B does not precede A in time.''

Goldvarg and Johnson-Laird (2001) use the same tools to model token causal statements (e.g., ''Jane causes Joe's anguish'') and type statements (e.g., ''Women like Jane cause anguish in men like Joe''). The other theories discussed in this paper follow suit.

A substantial part of the mental model theory consists of a relatively complicated set of rules for combining models in the face of multiple causal statements. For instance, if told ''A causes B'' and ''B causes C,'' each statement elicits a set of possibilities that must be rectified in order to have a single coherent set to draw conclusions from. Broadly speaking, consistent models are combined while inconsistent models and redundancies are removed. Goldvarg and Johnson-Laird (2001) state the principles that they use to combine models and point out that they have been embodied in a computer program.

2. Force dynamics theory

Force dynamics theory proposes that mental representations of causal relations reflect one of the properties of causes in the physical world. Specifically, this framework represents causal relationships in terms of configurations of forces (Barbey & Wolff, 2007; Talmy, 1988; Wolff, 2007).

Force dynamic representations reflect the interaction of two main entities: an affector and a patient (the entity acted upon by the affector). In Wolff's (2007) formulation, these entities are analyzed in terms of three dimensions: (i) the tendency of the patient for the endstate; (ii) the presence or absence of concordance between the affector and the patient; and (iii) progress toward the endstate (i.e., whether or not the endstate occurs). Table 2 summarizes how these dimensions differentiate the concepts *cause*, *allow*, and *prevent*. According to this framework, the sentence ''The explosion caused the bridge to collapse,'' for example, represents a state of affairs in which the patient (the bridge) did not have a tendency to collapse, the affector (the explosion) acted against the patient, and the result (the collapse of the bridge) occurred.

The proposed force dynamic dimensions are formally represented in the language of vectors. As Fig. 1 illustrates, the patient, B, has a tendency for the endstate, E, when the vector associated with the patient points in the same direction as the vector that specifies the endstate. Thus, the patient vector points in the same direction as the endstate vector for *allow* and *prevent*, but not in the case of *cause*. Concordance occurs when the vectors associated with the patient and affector point in the same direction. As illustrated in Fig. 1, the patient and affector are concordant for *allow*, but not in the cases of *cause* and *prevent*. Finally, the result is expected to occur when the resultant vector points in the same direction as the endstate vector, a property represented by *cause* and *allow*, but not *prevent*.

Force dynamics theory has been extended to inferences drawn from multiple causal relations (Barbey & Wolff, 2007). In the context of transitive inference, this is accomplished by representing the configuration of forces that underlie A's relationship to B, and B's relationship to C, and then linking these premises to draw a transitive inference about A's relation to C. As Fig. 2 illustrates, the *transitive dynamics model* proposes that the premises are connected by using the resultant vector in the first premise (BA) as the affector vector in the second (B_{BA}). The resultant vector points in the same direction as the affector in the second premise unless the B terms in the two premises conflict (e.g., if one is negated).

A conclusion is drawn in this framework by forming a new configuration of forces based on the two premises. Specifically, the affector in the conclusion is the affector from the first premise; the endstate vector in the conclusion is the endstate vector from the last premise; and the patient in the conclusion is the resultant of the patient vectors in the premises. The resulting configuration of vectors can then be interpreted according to the semantics for individual causal relations (see Table 2).

Table 2
Force dynamic representations of several causal concepts

	Patient Tendency for the Endstate	Affector-patient Concordance	Endstate Approached
Cause	No	No	Yes
Allow	Yes	Yes	Yes
Prevent	Yes	No	No

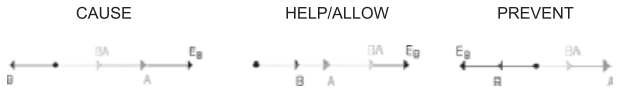

Fig. 1. Configurations of force associated with *cause*, *allow*, and *prevent*. A, the affector force; B, the patient force; BA, the resultant of A and B; E, endstate.

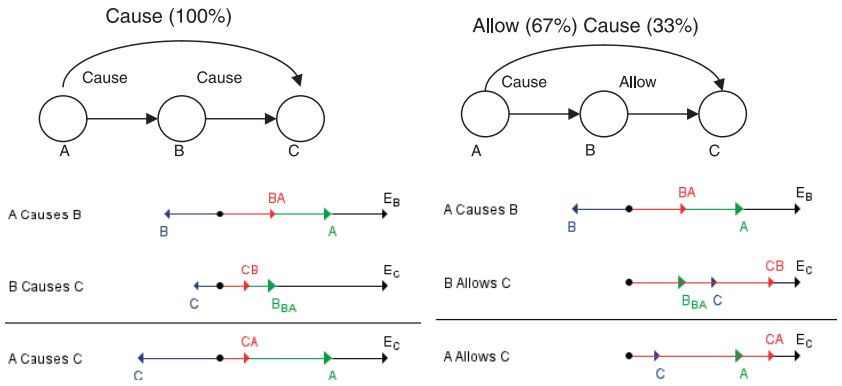

Fig. 2. Transitive arguments and configurations of force.

3. Causal model theory

The causal model theory of the meaning of *cause*, *enable*, and *prevent* makes use of the graphical formalism of causal Bayes nets (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; for a nontechnical introduction, see Sloman, 2005). The framework offers a way to represent and make inferences about causal systems using nodes and links in the form of acyclic graphs. The critical idea for our theory is the semantics of a link. A link between X and Y represents a causal mechanism that has X as one of its inputs and Y as the output. Its semantics is defined in terms of intervention (Woodward, 2003). A causal path involving one or more links exists between X and Y if intervening on X could change the value of Y (and not the converse). So the semantics of causal Bayes nets are well defined (if controversial, see Cartwright, 2002).

Our application of the theory makes little use of the technical apparatus of the Bayes nets framework. Our background assumption concerns the nature of common ground in discourse involving causal notions. We assume that discourse takes place in the context of a more-or-less shared set of assumptions about the causal relations that hold in the domain being discussed. For instance, a conversation about travel normally assumes implicitly that vehicles enable movement, that movement causes energy depletion, that money enables comfort, etc. In general, two events, A and B, could have any number of assumed causes and effects (Fig. 3a).

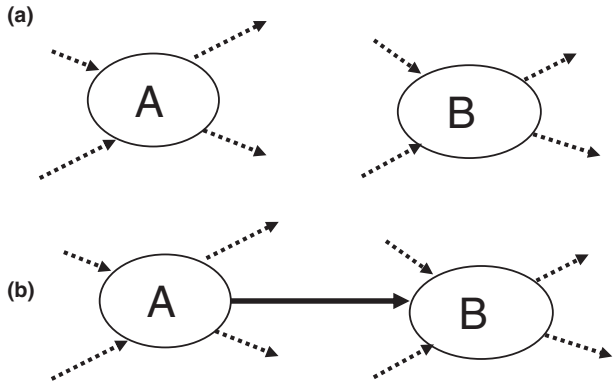

Fig. 3. (a) Discourse participants assume that events are involved in some number of causal relations. (b) ''A causes B'' introduces or emphasizes a link from A to B.

We further assume that the utterance ''A causes B'' asserts a link from A to B. It either refers to a pre-existing link or adds one that is not already there (Fig. 3b). What this means in the causal Bayes net framework is that the utterance asserts or brings into focus a mechanism that has A as an input and B as output. The mechanism may or may not have other inputs; the utterance itself does not specify. However, surrounding utterances or other aspects of context might specify. For instance, if discourse participants are looking at a machine, it might be obvious that it has other inputs. A conventional assumption in the Bayes net community corresponds to Goldvarg and Johnson-Laird's (2001) temporal constraint: The chronology of events respects their causal structure. Effects do not occur before causes.

A link in a causal Bayes net has the potential to support an intervention. Changing the value of A can change the value of B, but the converse does not hold. The intervention is only potential because physical limitations can prohibit an actual intervention. The gravitational pull of the sun is causally linked to Jupiter's orbit even though people have no way of intervening on the sun's gravitational field and would hopefully refrain from doing so even if they could. So the notion of intervention is best understood counterfactually: A person's mental representation includes a causal link from A to B if and only if that person believes that in a possible world in which A was intervened on, B would change. This assumes that the intervention on A would be drastic enough to change B and that any other variables that would disable the effect of A on B are absent.

A system of causal links in a Bayes net can be represented as a set of structural equations by applying the rule that effects are a joint function of all their causes (and assuming a single, exogenous error term). For instance, the causal model expresses the following structural relation:

$$D := f(A,B,C,\epsilon) \tag{1}$$

where ε represents uncertainty due to other variables not represented in the model. The possibility of uncertainty allows the relation between A, B, C, and D to be probabilistic. In

other words, the causal model is equivalent to a probabilistic functional relation from A, B, and C to D. The probability of each value of D is specified for any combination of values of A, B, and C.

We do not assume that f represents a linear relation. We use := rather than = to indicate that a structural equation is not merely a mathematical function relating variables A through D. In particular, one is not free to rearrange terms, causes must be isolated on one side of the equality and effects grouped on the other. In the example, the fact that D is isolated influences what operations can be performed. An intervention on D is represented by removing Equation (1) from the system of structural equations. An intervention on A, B, or C would not remove Equation (1), it would instead fix the value of the variable intervened on. An intervention only removes an equation that has the intervened-on variable as its isolated effect. More relevant to the model we present here, not all algebraic operations are legal. In particular, effects cannot be substituted for other effects; variables on the left cannot be substituted for other left-hand variables. As it happens, people show a preference for representing equations in terms of the structural equation that corresponds to the causal structure underlying the equation as opposed to its algebraic equivalents (Mochon & Sloman, 2004).

In this paper, we follow Goldvarg and Johnson-Laird (2001) in exploring only binary events that do or do not occur. We will also represent statements deterministically (ignoring ε in our formulations) for the sake of simplicity although our framework could be applied to utterances intended or understood probabilistically. Given these assumptions, the structural equation representation of ''A causes B'' is merely:

$$B := A. \tag{2}$$

In other words, if all a reasoner knows is that A causes B, then the reasoner will give B whatever value (occurs or does not occur) that A is given.

''A enables B'' makes a different assertion. It asserts that A (or an event related to A) is necessary to allow some other accessory event to affect B. It asserts: (i) A is linked to B; (ii) some accessory variable X is also linked to B (see Fig. 4); and (iii) A is necessary for B. Assertion (iii) is actually too strong. One might say ''Access to a coffee maker enables productivity.'' But clearly access to a coffee maker is not necessary for productivity. More likely, access to caffeine is what is meant. A coffee maker serves as an instance of the type of event that is necessary; it is not itself necessary. In the rest of this paper, this will be our intended meaning of ''A is necessary for B'': A is an instance of a relevant event type.

In terms of structural equations, we represent ''A enables B'' as:

$$B := A \text{ and } X. \tag{3}$$

where X is an accessory variable that is either unknown (e.g., ''Setting the DVD type to its playing speed enables the recording function to operate.'') or given by prior knowledge (e.g., ''The availability of cash enables one to really enjoy Las Vegas.'') or given by the environment. *Enables* has both a ''to help'' sense and a ''to allow'' sense. Equation (3) is intended to express both senses. In what follows, we use Equation (3) to represent not only ''A enables B'' but also ''A allows B'' and ''A will allow B.'' Our assumption is that

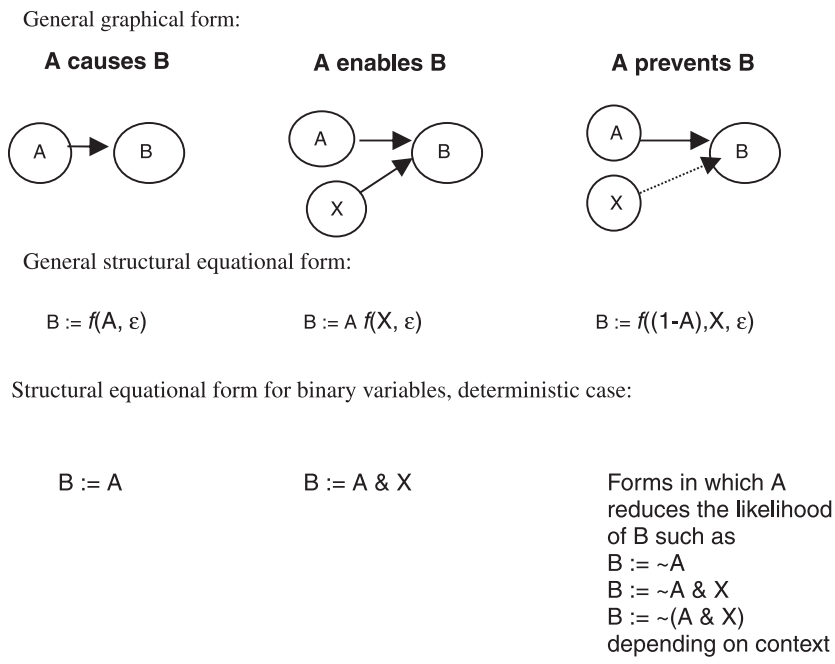

Fig. 4. A causal Bayes nets theory of the meaning of *cause*, *enable*, and *prevent*.

whatever difference holds among these phrases is irrelevant to the reasoning data that we review. Because we are considering binary events, (3) states both that A and X are necessary for B. A more general formulation would be required for nonbinary events.

Prevents differs from both *cause* and *enable* in that it operates in the domain of nonevents, the negative domain of events that do not occur. ''A prevents B'' suggests in general that A makes B less likely. We speculate that language does not articulate negative domains as well as positive ones because people have an easier time reasoning about events that do happen than those that do not. For this reason, ''A prevents B'' is vaguer than corresponding *cause* or *enable* statements (cf. Walsh & Sloman, unpublished data) in that sometimes it does suggest an accessory variable causally linked to B. For example, ''Severe punishment prevents crime'' only makes sense in the context of some set of conditions promulgating crime. However, *prevent* does not always implicate an accessory variable (''Anxiety prevents clear thinking''). Therefore, we allow some freedom in the interpretation of ''A prevents B.'' In the deterministic, binary case with no accessory variable implicated, we simply have:

$$\mathrm{B} := \sim\mathrm{A}. \tag{4a}$$

However, when an accessory variable is implicated we have several forms, including:

$$\mathrm{B} := \sim\mathrm{A} \text{ and } \mathrm{X}. \tag{4b}$$

$$B := \sim(A \text{ and } X). \quad (4c)$$

The critical property is that A must reduce the likelihood of B when X is present or unknown.

For the specific arguments that we examine here, the causal model theory reduces to Boolean algebra. Causal model theory remains relevant. First, it provides the motivation for and source of Equations (2–4) and in fact causal graphs can be used directly to derive predictions. Second, causal model theory can be applied to a range of arguments, including arguments involving uncertainty, different structural relations, intervention, and arguments with arbitrarily valued variables. Any generalization of our theory to such arguments could apply the same causal model framework but not Boolean algebra. Moreover, structural equations are constrained by their semantics by virtue of being asymmetric: variables on the right of the '':='' sign are causes and those on the left are effects. As a result, not just any substitution is legal. For instance, a common cause model would be represented as

$$A := C$$

$$B := C$$

but this does not mean that A is equivalent to B. In general, only causes can be substituted for, not effects.

4. Testing the theories

4.1. Two-premise arguments

The theories can be fruitfully compared and contrasted by focusing on Experiment 4 of Goldvarg and Johnson-Laird (2001). Participants were given two premises that related three variables via a pair of causal verbs, for example:

A causes B
B allows C

Participants were asked ''what, if anything, follows?'' All pairs of arguments of the form A relates to B and B relates to C drawn from the four relations *causes*, *allows*, *prevents*, and *not causes* (''not A causes B'') were used. That makes 16 argument forms. Each statement related familiar psychological terms in an unfamiliar way (e.g., ''Obedience allows motivation to increase.''). The modal responses from Goldvarg and Johnson-Laird for each argument form are shown in Table 3. The response predicted by mental model theory was the majority response in 15 of the 16 cases.

To apply our model to the task, we make two processing assumptions, first, that premises are combined via substitution. For instance, following Equation (2), we represent premises:

A causes B
B causes C

Table 3
Data from Goldvarg and Johnson-Laird (2001) Experiment 4. Percentage of modal conclusion drawn by participants (n = 20) for 16 two-premise arguments

	First Premise			
Second Premise	A Causes B	A Allows B	A Prevents B	Not A Causes B
B causes C	A causes C 100	A allows C 90	A prevents C 95	Not A causes C 100
B allows C	A allows C 95	A allows C 95	A prevents C 100	Not A allows C 100
B prevents C	A prevents C 100	A allows not C 70	A prevents C 75	Not A prevents C 100
Not B causes C	A prevents C 45	A allows not C 60	A causes C 85	Not A prevents C 75

as:

$$B := A. \tag{5}$$

$$C := B. \tag{6}$$

We can substitute (5) into (6) to obtain

$$C := A$$

which, according to causal model theory, is the representation of ''A causes C.'' So the theory predicts that people will respond ''A causes C'' to this argument as 20⁄20 participants did. As discussed above, the only permissible substitutions are causes for their effects. Simple substitution works here because we are only considering simple binary functional relations from causes to effects. In the more general case described in Equation (1), the functions in the two premises would have to be composed.

The second processing assumption is implicit in all causal reasoning theories. It is that for an argument of the two-premise form:

P relation Q
Q relation R,

people generate a conclusion that relates P and R with their presented valence, either positive (occurs) or negative (does not occur).

We derive predictions for the remaining 15 tested arguments using causal model theory in Appendix A (appendices are available online at: http://www.cogsci.rpi.edu/CSJarchive/Supplemental/index.html). On 13 out of 16 problems, causal model theory agrees with mental model theory's predictions and with the data. The transitive dynamics model also agrees with

mental model theory on 13 (see Barbey & Wolff, 2007). We focus on the three problems for which the predictions of the causal and mental model theory diverge.

First, causal model theory predicts the conclusion ''A prevents C'' in response to the premises:

A allows B
B prevents C

but 14/20 Princeton students said ''A allows not C'' as predicted by mental model theory and the transitive dynamics model. However, in Barbey and Wolff's (2007) replication, 15/19 Emory students said, ''A prevents C.'' *Allows* has a deontic sense involving permission. The two groups may have differed in their likelihoods of this deontic reading. Alternatively, the difference may be due to the greater range of statements in Barbey and Wolff.

The second diverging problem has a very similar pattern:

A allows B
Not B causes C.

Causal model theory predicts ''A prevents C'' but 12/20 of Goldvarg and Johnson-Laird's students said, ''A allows not C'' as predicted by mental model theory. The transitive dynamics model predicts the conclusion ''Not A causes C.'' But in Barbey and Wolff's (2007) replication, 14/19 students said, ''A prevents C.''

The final problem is the following:

A prevents B
B prevents C

In both the original study and the replication, most participants concluded ''A prevents C'' as predicted by mental model theory. Causal model theory predicts the conclusion ''A causes C.'' The transitive dynamics model predicts either this conclusion or ''A allows C.'' We are surprised by the data as it is easy to generate examples with a compelling ''A causes C'' conclusion. For example,

Distractions prevent concentration
Concentration prevents accidents.

Clearly, the conclusion that distractions cause (or allow) accidents is more plausible than the conclusion that distractions prevent accidents. In general, everyday examples that have *cause* or *allow* rather than *prevent* conclusions are much easier to generate. In fact, Barbey and Wolff (2007) found that sentences with *allows* were the modal conclusion. This leads us to conclude that the Goldvarg and Johnson-Laird (2001) and Barbey and Wolff's (2007) data were produced by an atmosphere effect induced by the experimental materials or procedure: Participants are carrying over the verb in both premises to the conclusion.

4.2. More two- and three-premise arguments from Barbey and Wolff

Barbey and Wolff (2007) further evaluated the transitive dynamics and mental model theories in the context of two- and three-premise causal arguments (for a summary of their predictions, see Tables 5 and 6 of Barbey & Wolff, 2007). We review the modal responses from this study, although all three theories motivate multiple conclusions for some problems (see Barbey & Wolff, 2007).

For the two-premise arguments, the transitive dynamics model predicts the modal response for 27 out of 32 two-premise arguments (significantly greater than chance by a binomial test, $p < .001$) and mental model theory predicts the modal response for 25 of them ($p = .002$). We derive the predictions of the causal model theory for these arguments in Appendix B. The theory predicts the same number as the transitive dynamics theory, though not significantly more than mental model theory ($Z < 1$).

For the three-premise arguments, Barbey and Wolff (2007) show that the mental model and transitive dynamics theories correctly predict the modal response on 9 and 10 of the 15 problems, respectively (not significantly greater than chance). To derive predictions for the causal model theory we had to drop the second processing assumption because it applies only to two-premise arguments. We instead made the processing assumption that participants generated conclusions that related the first (A) term to the last (D) term, giving each term a positive valence. The idea is that reasoners simplify the representation of complex three-premise arguments by limiting themselves to positive values of the key terms that they are relating (cf., Goldvarg & Johnson-Laird, 2001).

To illustrate a derivation for a three-premise argument, consider the premises:

A causes B
B causes Not C
C causes Not D

We translate these into structural equations just as before:

$$B := A$$

$$\sim C := B$$

$$\sim D := C$$

Substituting the first equation into the second, we get $\sim C := A$. We will assume that causes in 3-premise arguments are taken to be necessary so that substituting this into the third, we get $\sim D := \sim A$. The assumption that people relate positive values along with the assumption of necessity takes us from $\sim D := \sim A$ to $D := A$. This is the representation of ''A causes D,'' which is what the theory predicts people will say.

The causal model theory successfully predicts 13 out of 15 of the three-premise arguments ($p = .007$) as shown in Appendix C. This is marginally better than the other theories (using one-tailed tests, $Z = 1.65$, $p < .05$ in comparison to mental model theory and $Z = 1.29$, $p < .1$ for transitive dynamics).

In sum, causal model theory provides at least as good an account of the reviewed findings from Goldvarg and Johnson-Laird (2001) and Barbey and Wolff (2007) as the mental models and transitive dynamics theories. The theories can be further evaluated on the basis of other data and on their relative conceptual virtues like parsimony and generalizability to other scenarios and paradigms. In what follows, we report four experiments testing implications of causal model theory that distinguish it from other theories and then address the remaining data reported by Goldvarg and Johnson-Laird. In the General Discussion, we discuss the relative merits of the three theories and how they may relate.

5. Experiment 1: Does enable imply an accessory variable?

A critical claim of the causal model theory is that *enable* implies an accessory variable, whereas *cause* does not. We tested this claim by giving people statements involving *cause* and others involving *enable* both with and without an accessory variable and asking them to draw a conclusion. Here's an example of the cause condition:

> Imagine you enter a classroom and the instructor is saying about a group of people:
> A. When stress occurs, stress causes fixation.
> You are told that stress is present. Based on sentence A., would you conclude that fixation will occur?

The enable conditions were identical except that, in the accessory-absent case, Statement A read, ''A. When stress occurs, stress enables fixation.'' And in the enable-accessory-present condition, Statement A read, ''A. When attention occurs, attention causes fixation, but only when stress enables it.''

According to causal model theory, judgments will not depend on the presence of an accessory variable in enable conditions because ''enable'' implies that the accessory variable already exists; it should not matter whether the accessory variable is explicitly mentioned in the experimental materials. Therefore, participants should be uncertain whether the effect occurred or not with enable relations because they will wonder whether the accessory variable was present. Mentioning the accessory variable should not reduce uncertainty much because we do not state whether it occurred. Hence, we predict: (i) that participants will be more likely to say ''yes'' for *cause* than *enable* (either with or without an accessory variable) and (ii) the presence of an accessory variable will not matter for *enable*.

In contrast, the transitive dynamics theory does not predict any systematic difference among the three conditions. The presence of an accessory cause does not in and of itself change any of the model's parameters (see Table 2) and the conditions do

not differ in ways that are relevant for the model. Mental model theory turns out to make the same predictions as causal model theory. Mental model theory does require that the accessory variable be represented explicitly and this turns out to make a difference to its predictions. The accounts of these theories are spelled out in greater detail in the discussion of Experiment 1.

5.1. Method

5.1.1. Materials

To keep the level of abstraction constant across the two domains, type (as opposed to token) casual relations were used. To clarify that the relations concerned types, we attributed the sentences expressing the relations to a classroom instructor on the assumption that participants would expect an instructor to focus on relations among broad categories. To illustrate, in one problem participants were asked to ''imagine you enter a classroom and the instructor is saying about a collection of chemicals.'' Then they were given one of the following three sentences:

Cause: A. When magnetism occurs, magnetism causes ionization.
Enable accessory absent: A. When magnetism occurs, magnetism enables ionization.
Enable accessory present: A. When conductivity occurs, conductivity causes ionization, but only when magnetism enables it.

Then participants were asked, ''You are told that magnetism is present. Based on sentence A., would you conclude that ionization will occur?''

Terms for all problems were chosen from the psychological and physical domains. Arguments were constructed from the following list of word pairs counterbalanced across the three conditions: vibration/radiation, motion/flexibility, magnetism/ionization, resistance/combustion, heating/pressure, solubility/erosion, motivation/anxiety, depression/memory, stress/fixation, relaxation/compliance, social reinforcement/paranoia, and addiction/insomnia.

5.1.2. Participants

Fifty-one participants were recruited using an advertisement in an online newspaper for Brown University students. Participants were entered into a $40 lottery. One participant chosen at random won the entire sum. The data from one person who did not answer all questions were eliminated because they were incomplete.

5.1.3. Design and procedure

Participants were tested via an online survey. Half the participants were presented with six psychological arguments followed by six physical arguments, while the other half were given the reverse order. Within each domain, three arguments came from our experimental conditions (i.e., cause, enable accessory absent, enable accessory present). The remaining three arguments were fillers containing the relations *prevents accessory absent*, *prevents*

accessory present, and *causes absence of*. Responses were given on a 1–7 scale from ''definitely not'' to ''definitely yes.''

5.2. Results

Means and standard errors for all three conditions are shown in Fig. 5. As predicted, mean confidence ratings for *cause* were higher than the average of the *enable accessory absent* conditions (5.57 vs. 4.65). A 3 (relation: cause, enable accessory absent, enable accessory present) × 2 (domain: psychological, physical) × 2 (order: psychological first, physical first) mixed model analysis of variance (ANOVA) was conducted. A marginally significant main effect of *relation* was found, $F(2,96) = 2.78$, $p = .067$. There was no significant effect of domain, $F(1,48) = 1.71$, n.s. No reliable effect of presentation order was observed, $F(1,48) = 1.14$, n.s., but there was a significant interaction between relation and order ($F(2,96) = 3.25$, $p < .05$)) due to participants who received psychological arguments first, giving lower *cause* ratings than did *physical-first* participants (4.84 vs. 6.30). A significant three way interaction from was also found, $F(2,96) = 5.18$, $p < .05$. No other interactions were reliable.

Planned comparisons confirmed that *cause* ratings were reliably higher than *enable accessory absent* and *enable accessory present*, $t(49) = 4.33$, $p < .001$. Against expectations, the presence of the accessory variable made a small difference for *enable*, $t(49) = 2.07$, $p < .05$.

Item analyses were also conducted. A 3 (relation: cause, enable accessory absent, enable accessory present) × 2 (domain: psychological, physical) × 2 (order: psychological first, physical first) ANOVA replicated the main effect of *relation*, as well as the interaction between *relation* and *order*. No other significant effects were found.

Planned comparisons by item confirmed *cause* ratings to be higher than *enable accessory absent* and *enable accessory present*, $t(11) = 5.97$, $p < .001$. Consistent with the predictions of causal models theory, the presence of an accessory variable did not have a significant effect on *enable* ratings, $t(11) = 1.2$, n.s.

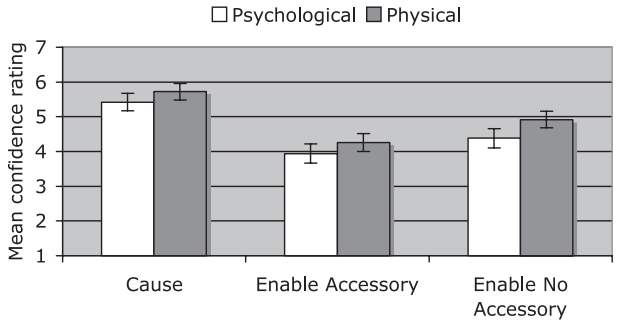

Fig. 5. Mean confidence that effect will occur in Experiment 1 for three conditions and two domains. 1 means no and 7 means yes. Standard error bars are shown.

5.3. Discussion

The predictions of the causal model theory were largely confirmed. First, certainty in the effect was greater for cause than for enable sentences. Second, the presence of an accessory variable had only a small effect for enable sentences, a difference that was not significant across items and that was smaller than the difference between enable and cause judgments.

Mental model theory is also consistent with the results. Knowing that the cause has occurred, the only viable model of ''A causes B'' is

A B

However, there two viable models of ''A enables B'':

A B
A ~ B

Therefore, the conclusion that B occurred is only licensed in the first case. If we add a statement of a third accessory variable (C causes B) in the enables case, there remain models where B occurs and others where B does not occur. This is consistent with the second result that the presence of an accessory variable made little difference for enable sentences.

The transitive dynamics theory cannot predict these results. That theory predicts certainty in the effect when the presence of the affector results in the endstate being approached. In the example, the affector is magnetism and the endstate is ionization in all three conditions. A transitive dynamics theorist could either argue that the endstate is approached in all three conditions because the statements in the problems imply that the criteria of application of the verbs (as shown in Table 2) are met. In that case, the theory predicts that participants should have been certain of the effect in all conditions, a prediction not borne out by the data. Or the theorist could argue that not enough information has been given to decide if the criteria have been met. In that case, the theory makes no prediction.

Barbey and Wolff (2007) and Wolff (2007) distinguish *cause* from *enable* by arguing that *enable* represents a chain of prevent relations (A prevents B and B prevents C implies that A enables C). In this sense, enable can be interpreted as incorporating an accessory variable. However, the current experiment did not use such a structure. Furthermore, A prevents B and B prevents C can also imply that A causes C according to the transitive dynamics model depending on the relative sizes of the vectors B and C. So the reduction of the relations to chains of other relations will not succeed in explaining the results of our experiment as we do not specify the relative sizes of B and C.

6. Experiment 2: Does cause imply an accessory variable?

Experiment 2 tests causal model theory's prediction that *cause* does not entail an accessory variable. We test this claim by giving people statements involving *cause* both with and without an accessory variable and asking them to draw a conclusion. Otherwise the

materials were like those of Experiment 1. Here is an illustration of a case without an accessory variable:

> Imagine you enter a classroom and the instructor is saying about a group of people:
> *Cause accessory absent*: A. When relaxation occurs, relaxation causes obedience.
> Question:
> You are told that relaxation is present. Based on sentence A., would you conclude that obedience will occur?

The corresponding case with an accessory variable has a different sentence A:

> *Cause accessory present*: A. Relaxation causes obedience when attention enables it.

Because *cause* does not assume an accessory variable, participants should not think about it unless it is explicitly mentioned. If it is not mentioned, the theory predicts that participants will respond with certainty that the effect will occur. However, mentioning an accessory variable (*attention* in the example) should make participants aware of a source of uncertainty that they did not previously consider and therefore they should give judgments closer to the midpoint of the response scale. In sum, we predict that participants will be more likely to say ''yes'' for cause without an accessory variable than with one.

This experiment also serves to distinguish the causal model and mental model theories, depending on the interpretation given to the materials. Knowing that the cause has occurred, the only viable mental model of ''A causes B'' is

A B

This is the only relevant model in the condition with no accessory variable and it suggests that B occurred. In the experiment, the accessory variable was introduced as an enabler of the effect. Call the enabler C. Then the only model consistent with ''A causes B,'' ''C enables B,'' and the presence of A is

A C B

This again implies that B occurred. So if the sentence ''A causes B when C enables it'' is interpreted to mean ''A causes B, C enables B,'' then mental model theory predicts that people should expect B in the presence and in the absence of an accessory variable and predicts no difference between the conditions.

However, a version of mental model theory is consistent with the results on a different interpretation of the experimental materials.[1] If the sentence ''A causes B when C enables it'' is interpreted to mean that the causal relation between A and B is effective only when C is present, then this would permit the possibility of B not occurring despite the presence of A (as long as C is absent). This makes use of a different sense of *enables* than the theory of Goldvarg and Johnson-Laird (2001) reviewed above.

The transitive dynamics theory again predicts no differences between the conditions. It does not distinguish inference about an effect of a cause in the presence versus the absence of an accessory variable.

6.1. Method

6.1.1. Materials

Arguments were constructed from the following list of physical and psychological word pairs: vibration/radiation, resistance/combustion, pressure/heating, solubility/erosion, stress/fixation, relaxation/obedience, paranoia/social reinforcement, addiction/insomnia. Here's an example of a physical argument:

> Imagine you enter a classroom and the instructor is saying about a collection of chemicals:
>
> A. When solubility occurs, solubility causes erosion.
>
> Question:
>
> You are told that solubility is present. Based on sentence A., would you conclude that erosion will occur?

6.1.2. Participants

Eighty-seven participants were recruited and paid for their participation as in Experiment 1. The data from two people were eliminated because they were incomplete.

6.1.3. Design and procedure

Participants again completed a Web-based survey. Each participant was presented with four arguments from each experimental condition. Half the participants were presented with all four psychological arguments followed by all four physical arguments, while the other half were given the reverse order. Otherwise the method was identical to Experiment 1.

6.2. Results and discussion

Means and standard errors for both conditions are shown in Fig. 6. As predicted, mean confidence ratings for *cause accessory absent* were higher than *cause accessory present* (5.94 vs. 3.51). A 2 (relation: accessory absent, accessory present) × 2 (domain: psychological, physical) × 2 (order: psychological first, physical first) mixed model ANOVA showed a significant effect of *relation*, ($F(1,83) = 193.29$, $p < .001$). Presentation order also affected judgments ($F(1,83) = 5.75$, $p < .05$), with *psychological-first* ratings being lower than *physical-first* ratings (4.49 vs. 4.94). There was no effect of domain, $F(1,332) = p$ .51, n.s., but there was a significant interaction between relation and, domain, $F(1,83) = 3.27$, $p < .05$. Planned comparisons confirmed that *cause accessory absent* ratings were higher than *cause accessory present*, $t(1,84) = 13.91$, $p < .001$. Item analyses showed an identical pattern.

The prediction of causal model theory was confirmed. Certainty in the effect was greater for cause sentences when no accessory variable was mentioned than when one was. For reasons given above, the transitive dynamics model is not consistent with

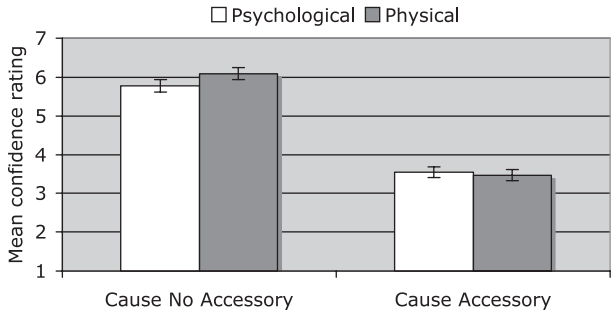

Fig. 6. Mean confidence that effect will occur in Experiment 2 for two conditions and two domains. 1 means no and 7 means yes. Standard error bars are shown.

this result. Mental model theory requires a special interpretation of *enables* to explain it.

7. Experiment 3: Replication and extension

The results of Experiments 1 and 2 cannot be directly compared because the materials used different syntactic structures. Experiment 3 is therefore an attempt to replicate the main findings of Experiments 1 and 2 using more comparable materials. We included both cause and enable statements with and without an accessory cause.

We used only two vignettes: one in the psychological and one in the physical domain. The four conditions for the physical vignette differed only in a single target sentence as follows:

Cause accessory absent condition: Magnetism causes ionization.

Cause accessory present condition: Magnetism causes ionization when conductivity enables it.

Enable accessory absent condition: Magnetism enables ionization.

Cause accessory present condition: Magnetism enables ionization when conductivity causes it.

In each case, participants were then told that magnetism is present and asked, based on the target sentence, whether he or she would conclude that ionization will occur.

Predictions parallel those of Experiments 1 and 2: According to causal model theory, participants should believe that the cause (magnetism) is sufficient for the effect (ionization) only in the cause-accessory-absent condition because in all other cases another variable should be assumed necessary and the value of that variable is unknown. So they should infer ionization with confidence only in the cause-accessory-absent condition. The same derivations apply to mental model theory as those given earlier. If we interpret the theory to represent the cause-accessory-present case as ''A causes B, C enables B,'' then the theory predicts that people should infer ionization in both cause conditions but neither enable

condition. The theory can explain it if we admit other interpretations. Again the transitive dynamics model makes no clear prediction.

7.1. Method

7.1.1. Materials

Two vignettes were used. The physical one was just described. In the psychological one, the target cause/enabler was ''relaxation,'' the effect was ''obedience,'' and the accessory variable was ''attention.''

7.1.2. Participants

One hundred twenty volunteers were recruited from cognitive science classes at Brown University. Thirty were assigned to each condition.

7.1.3. Design and procedure

Both independent variables, type of relation (cause versus enable) and accessory variable (present versus absent) were manipulated between-participants. Each participant filled out a one-page questionnaire asking them two questions in the same condition, a physical and a psychological one. The physical one was always asked first. In answer to the question, ''Would you conclude that ionization will occur/they will be obedient?'' participants circled a number between 1 (NO) through 4 (CAN'T TELL) to 7 (YES).

7.2. Results and discussion

The means for Experiment 3 are shown in Fig. 7. The causal model theory predictions were once again confirmed. People were more willing to infer the effect in the cause-accessory-absent condition than in other conditions. In all three remaining conditions, they expressed great uncertainty.

These conclusions are supported by two analyses of variance. For the physical problem, there was an overall effect of type of relation, $F(1,116) = 12.34$, $MSe = 2.43$,

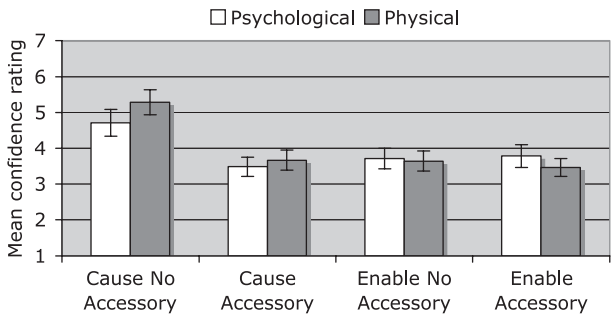

Fig. 7. Mean confidence that effect will occur in Experiment 3 for four conditions and two domains. 1 means no and 7 means yes. Standard error bars are shown.

$p = .001$, of accessory variable, $F(1,116) = 14.04$, $p < .001$, and, most importantly, a significant interaction, $F(1,116) = 7.26$, $p < .01$. For the psychological problem, there was not an overall effect of type of relation, $F(1,116) = 1.46$, MSe = 2.77, n.s. There was an effect of accessory variable, $F(1,116) = 6.36$, $p < .05$. The interaction between the two was very close to the standard significance level, $F(1,116) = 3.90$, $p = .051$. Most crucially, planned comparisons comparing cause-accessory-absent to the mean of the other three conditions were highly significant both for the physical problem, $F(1,116) = 33.03$, $p < .001$, and for the psychological problem, $F(1,116) = 10.84$, $p = .001$.

These results confirm the conclusions we drew for Experiments 1 and 2 and extend them to slightly different sentence forms. They again support causal model theory over the original version of mental model theory and the transitive dynamics theory.

8. Experiment 4: Labeling

All of the data that we have discussed so far concern how people reason about relations already labelled *cause*, *enable*, or *prevent*. But all three theories at hand (mental model, transitive dynamics, and causal model) offer predictions about production as well, how people will label relations given different conceptual structures. Causal model theory's predictions are a function of causal structure: If A alone is the source of B, then people should describe the relation as ''A causes B.'' But if some other variable X is necessary for A to have an effect on B, then people should assert ''A enables B.'' Experiment 4 asks people to label a relation between two variables after reading a vignette describing a structural model that includes the variables. Mental model theory is insensitive to these structural relations and thus does not predict an effect of the presence of an accessory variable on how people label. For reasons we spell out in the Discussion, the transitive dynamics model can be interpreted to make the same predictions as causal model theory in this experiment.

8.1. Method

8.1.1. Materials

Vignettes describing causal relationships were chosen from both the psychological and physical domains. These could have a single cause (1 link) or an additional accessory variable (2 links). Here's an example of a physical vignette with the 2-link version in italics:

> John discovered that if he pressed button A then light B would come on. *He also found that this would happen only if he turned switch C to the on position.*
>
> Question:
>
> Which of the following five statements best describes the relation between A and B?
>
> A causes B
> A enables B

A helps B
A allows B
A prevents B

8.1.2. Participants

Seventy participants were recruited using the same method as in Experiments 1 and 2 except that participants were entered into a $50 lottery.

8.1.3. Design and procedure

Web-based surveys were again used. Surveys consisted of four five-vignette blocks. One Physical block was followed by two Psychological blocks, followed by the remaining Physical block. Half of the participants were presented with blocks of three 2-link vignettes followed by two 1-link vignettes. The other half had blocks of three 1-link vignettes followed by two 2-link vignettes. The order of causes and enablers were counterbalanced within the 2-link condition. For example, the *accessory first* complement to the above vignette read:

> John discovered that if switch A was in the on position, when he pressed button C then light B would come on.

Half of the participants in all conditions first received two *cause-first* blocks followed by two *accessory-first* blocks. The other half were presented with the opposite order.

8.2. Results

The means for Experiment 4 are shown in Figs. 8 and 9. Since *enables*, *helps,* and *allows* are all ways of expressing enabling relations, all three responses were collapsed for analysis. As predicted, the proportion of cause responses was greater for 1-link than 2-link models, and, conversely, the proportion of enable responses was greater for 2-link than 1-link models. *Prevent* responses were rare, accounting for 4.95% of 1-link responses and 2.96% of 2-link responses, and will thus be omitted from further analysis.

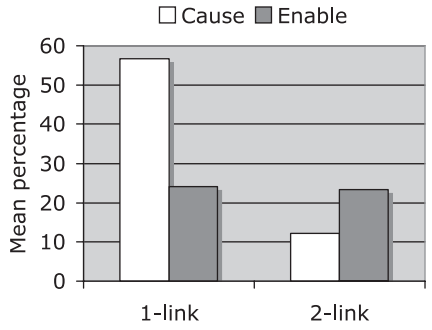

Fig. 8. Percent of participants choosing *cause* and *enable* in Experiment 4 for one versus two links in the psychological domain.

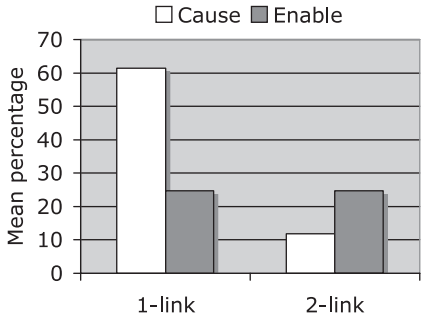

Fig. 9. Percent of participants choosing *cause* and *enable* in Experiment 4 for one versus two links in the mechanical domain.

Two chi-squared tests of independence confirmed the reliability of the cause-enable differences in both the psychological and physical domains. For psychological events, *cause* responses were more frequent for 1-link models and *enable* responses were more frequent for 2-link models, $\chi^2(1,\ n = 464) = 31.98,\ p < .001$. Correspondingly, in the physical domain, *cause* responses were more frequent with no accessory variable and *enable* responses were more frequent in the presence of an accessory variable, $\chi^2(1,\ n = 509) = 52.52, p < .001$.

8.3. Discussion

The predictions of causal model theory were again confirmed. When presented with 1-link models of causal relationships people tended to say, ''A causes B''. However, with the introduction of an accessory variable people responded that ''A enables B.''

It might seem that our theory makes an obviously incorrect prediction, that whenever people know about an accessory cause, they will consider a target variable to be an enabler and not a cause. How for instance can causal model theory explain why people say the match caused the fire when they know oxygen must have been present? The answer is that the relevant causal model for determining which causal verb to choose is not the causal model associated with the entirety of a person's knowledge but with the discourse model depicted in Fig. 3. Even though people may know that oxygen must be present for a fire, it is background knowledge that is generally not explicit in a discourse. In contrast, when oxygen's causal role is attended to, the presence of a spark must be specified explicitly to understand why fire occurred. Hence, oxygen needs an accessory cause to understand its causal role and that is why it is understood as enabling fire rather than causing it. Of course, in an environment normally without oxygen, in which its presence must be specified explicitly to understand the occurrence of fire, causal model theory predicts that the spark would be talked about as enabling fire rather than causing it. Cheng (1997) offers a different view of causing and enabling in terms of which contrast set to the actual event is most natural. The two views may be compatible.

We see no way for mental models theory to explain the results of Experiment 4. The task for the theory is to explain how sentences like those in our experiment describing causal structure without causal terms (i.e., cause, allow, prevent) elicit labels of *cause* versus *enables*. To do so, the theory requires either a theory of natural language understanding or a theory mapping causal structure into causal language. Goldvarg and Johnson-Laird (2001) do not offer either.

The transitive dynamics model can explain the results of Experiment 4. According to this framework, 1-link models represent the configuration of forces underlying a causal event (see Fig. 1). Causal statements expressed by 2-link models could represent A's relation to B while incorporating X. This is accomplished by representing the affector, A, and the patient formed from the resultant of X and B, in relation to the endstate, B. The fact that A and X are jointly necessary for B implies that affector A and the patient formed from the resultant of X and B point in the direction of endstate B. As a result, the transitive dynamics model can predict that this configuration of forces represents ''A enables/allows B'' (Fig. 1).

9. Other data

Goldvarg and Johnson-Laird (2001) offer four other experiments as support for their theory. In this section, we suggest how those data can also be explained by causal model theory.

In their Experiment 1, they asked participants to list the possibilities associated with three causal terms (*will cause*, *will prevent*, *will allow*, and *will allow not*). Participants were told to list what is possible and impossible in terms of the four combinations of cause and effect. They found that people listed possibilities generally consistent with mental model theory (a conclusion not accepted by Kuhnmünch & Beller, 2005). In each case, the modal response was the list of possibilities associated with the ''strong'' interpretation of the causal relations, according to which the cause is necessary and sufficient for the effect. On this interpretation, *cause* and *allow* converge to the same set of possibilities, as do *prevent* and *allow not*.

To derive a prediction from causal model theory, we assume that people will list the possibilities consistent with the structural equation model for each verb. For instance, *cause*, represented by Equation (2), states that B := A. So if A = a, then B = b and if A = ∼a, then B = ∼b. This is consistent with the possibilities

$$\begin{array}{cc} a & b \\ \sim a & \sim b \end{array}$$

and this was precisely the modal response for the *will cause* sentences. On the assumption that *will allow* has the same representation as *enable*, the causal model theory predicts the modal response for *will prevent*, *will allow*, and *will allow not* using the corresponding structural equations.

In their Experiment 2, Goldvarg and Johnson-Laird (2001) described the effect of pairs of variables on an outcome and asked students to identify which variable was the cause and which the enabler. In one case, students were told

> Given that there is good sunlight, if a certain new fertilizer is used on poor flowers, then they grow remarkably well. However, if there is not good sunlight, poor flowers do not grow well even if the fertilizer is used on them.

We extract the following intended meaning from this description: Flower growth (G) is promoted by a certain new fertilizer (F). Therefore,

$$G := F. \tag{7}$$

However it turns out that sunlight (S) is necessary for the fertilizer to have its effect. In causal model terms this implies:

$$G := S \ \& \ F. \tag{8}$$

Equation (8) suggests that both sunlight and fertilizer are enablers. However, participants were forced to choose one enabler and one cause. As (7) provides reason to identify fertilizer as a cause, the prediction is that fertilizer is the cause and sunlight the enabler. This is what most people said. Kuhnmünch and Beller (2005) offer a fuller analysis of the Goldvarg and Johnson-Laird paragraphs arguing that the assignment of variables as causes or enablers can be fully attributed to linguistic markers without the need to assume any form of conceptual representation.

In a second illustration, Goldvarg and Johnson-Laird (2001) gave the following description:

> Given the use of a certain new fertilizer on poor flowers, if there is good sunlight then the flowers grow remarkably well. However, if the new fertilizer is not used on poor flowers, they do not grow well even if there is good sunlight.

The intended meaning of this description seems to be that flower growth is promoted by sunlight. However, it turns out that fertilizer is necessary for the sunlight to have its effect. Representations of these two statements are

$$G := S. \tag{9}$$

$$G := F \ \& \ S. \tag{10}$$

Equation (10) suggests that both sunlight and fertilizer are enablers, but (9) provides reason to identify sunlight as a cause. Being forced to choose, participants should make sunlight the cause and fertilizer the enabler. This was the modal response.

In their Experiment 3, Goldvarg and Johnson-Laird (2001) asked participants to make inferences with various argument forms. One form was the following:

> A will cause B

A
Does B follow (yes or no)?

To represent the first premise, causal model theory asserts via Equation (2) that B := A. The second premise asserts that A occurred. Hence, B occurred and people should answer ''yes.'' Ninety-three percent did. We make the same assumptions as earlier to represent the remaining arguments: Equation (3) for *will allow* and Equation (4a) for *prevent*. Note that Equation (3) does not give a value for the effect if the accessory variable is not specified. As it was not specified by Goldvarg and Johnson-Laird, this would predict uncertainty. Uncertainty should lead to variable responding when people are forced to provide a yes or no response as they were in this experiment. On these assumptions, causal model theory makes the same predictions as mental model theory and is consistent with the data.

Experiment 5 gave participants problems like the following:

One of these assertions is true and one of them is false:
Marrying Evelyn will cause Vivien to relax.
Not marrying Evelyn will cause Vivien to relax.
The following assertion is definitely true:
Vivien will marry Evelyn.
Will Vivien relax?

Sixty-eight percent of Princeton undergraduates said yes, and the remaining ones said that it is impossible to know.

Causal model theory's operating principle is that people do not reason using propositional logic and hence it has no good way of representing ''One of these assertions is true and one of them is false.'' We see this as a virtue as people seem to have trouble with it too. To model this problem, we would set up two causal models, one stating that marrying Evelyn causes Vivien to relax and the other that not marrying Evelyn causes Vivien to relax. On being told that Vivien will marry Evelyn, the first mechanisms will deliver the proposed effect that Vivien will relax. This explains why 68% of students came to that conclusion. Another interpretation of the two initial premises is that marrying Evelyn is independent of Vivien's state of relaxation. If this is how the problem is represented, then knowing they got married reveals nothing and an appropriate response would be ''impossible to know.'' This could explain the remaining 32% of responses.

Goldvarg and Johnson-Laird (2001) present an analogous problem involving *prevent* rather than *cause*. It is amenable to the same kind of analysis. They also present two control problems whose data fall out directly from Equations (2) and (4a).

10. General discussion

In sum, we have presented a novel theory of the meaning of *cause*, *enable*, and *prevent* framed in the language of causal Bayes nets. The theory posits that all three relations

introduce or refer to links that support intervention in a background causal model serving as common ground in a discourse. The theory can be distinguished from other theories of the meaning of these terms, mental model theory (Goldvarg & Johnson-Laird, 2001) and transitive dynamics theory (Barbey & Wolff, 2007), in positing that *enable* but not *cause* assumes an accessory variable. Both *cause* and *enable* increase the probability of effects, whereas *prevent* decreases the probability.

The causal model theory fares well in empirical tests. It fits the data from both two- and three-premise conclusion production tasks reported by Goldvarg and Johnson-Laird (2001) and Barbey and Wolff (2007) at least as well as the other theories. Moreover, we report four experiments that support our proposals about *cause* and *enable* by showing that an inference about an effect in a *cause* sentence is sensitive to the presence of an accessory variable but an inference in an *enable* sentence is not. Mental model theory cannot explain the results of Experiment 4 and may or may not be consistent with Experiments 2 and 3. The transitive dynamics model cannot explain the results of Experiments 1, 2, and 3. We also reviewed four additional experiments from Goldvarg and Johnson-Laird that had people reason about causal terms. With appropriate assumptions, the causal model theory offers accounts of those data.

Besides offering an account of the experiments reported here, causal model theory has an important conceptual advantage. It is the simplest of the theories for deriving predictions. The inferential steps required to draw conclusions are trivial for the arguments discussed here and even with more complicated arguments. They consist of the most basic kind of composition, mostly just substituting one variable for another in a two-variable equality. In contrast, the derivation procedures in mental model theory consist of a list of nonobvious principles that are used to implement a computer program. These authors regularly have to appeal to the computer program to derive predictions. For us, the program is essentially a black box. Force dynamics theory requires a set of vectors additions. The additions themselves are simple, but knowing which vectors to add is not.

Causal model theory has the additional advantage that it extends directly to situations involving multiple causes, enablers, or preventions. The theory already assumes a background of other causes and effects. Any variables added to a discourse either introduce a new structural equation if they are effects, or introduce a variable to the right-hand side of an existing equation if they are causes, or both. In contrast, the number of mental models required to represent new variables increases exponentially. The transitive dynamics model represents situations involving multiple causal relations. Each new causal relation, however, introduces a configuration of forces that need to be represented and integrated with existing configurations.

Moreover, causal model theory can be naturally extended to probabilistic causal relations (like ''smoking causes lung cancer''). Indeed, causal Bayes nets are probability models. Nothing additional is needed except to add an error term to the structural equations [already included in Equation (1)]. A mental model theory of probabilistic reasoning has been proposed (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). However, that model allows the possibility of any combination of events. The probability of a proposition corresponds to the proportion of models in which it occurs. Such a representation cannot be

integrated with Goldvarg and Johnson-Laird's (2001) representation of causal verbs which depends on the absence of specific types of model. Like mental model theory, the transitive dynamics model's representation of *cause*, *allow*, and *prevent* is deterministic. However, uncertainty can arise concerning the absolute magnitudes of the force vectors when multiple causal relations are combined. For example, multiple conclusions follow from the argument ''A *causes* B and B *allows* C'' depending on the relative magnitude of the patient vectors. The resulting probabilistic representations are thus limited to a particular class of arguments (see Barbey & Wolff, 2007).

The advantages of causal model over mental model theory result from causal model theory's representation of causal structure in contrast to mental model theory's representation of extensional structure. Causal model theory represents functional relations, whereas mental model theory represents sets of possibilities. Functional relations compose with each other and with additional variables economically and transparently. Possibilities do not. Of course, these are not mutually exclusive forms of representation. People may well be able to represent both. The advantage of representing possibilities is direct knowledge of the state of the world and of possible worlds; the advantage of functional representation is direct knowledge of the governing mechanisms. If people do represent both, the representations are potentially isomorphic and in that sense they could provide identical information. But as indicated by the distinct predictions reviewed above, causal model and mental model theory are not isomorphic. Nevertheless, the data that we have reviewed do not rule out the possibility that each explains a distinct form of reasoning.

Causal model theory has the additional advantage that it provides a language and inferential machinery for distinguishing observation and intervention (Pearl, 2000; Spirtes et al., 1993). When one observes an event, that event is diagnostic of its causes. The event is not diagnostic of its causes if it is produced by an agent's intervention. People are highly sensitive to this logic (Sloman & Lagnado, 2005; Sloman, 2005; Waldmann & Hagmayer, 2005). The other theories reviewed here have no natural way of making this distinction.

The aspect of mental model theory that we are most uncomfortable with is the near identity it claims between *cause* and material implication as indicated by the identical set of possibilities that each is associated with. They differ only with respect to the temporal constraint that effects do not precede causes according to Goldvarg and Johnson-Laird (2001). The temporal constraint makes sense for *cause* because it relates events; it does not make sense for material implication as it relates atemporal propositions. Identifying the meaning of conditional statements with the material conditional is already tendentious (Bennett, 2003; Evans & Over, 2004). Extending the material conditional to the meaning of cause is extreme. The material conditional $p \subset q$ is equivalent to the proposition not-p OR q. To say ''A causes B'' seems at minimum to refer to a relation between A and B, not to an event that occurs whenever A does not or B does.

Sloman (2005) makes the argument that causal models serve as the fundamental representational form for human thought. People understand the world by understanding its mechanisms and how they relate to one another. People reason by figuring out which events are likely to come about by virtue of those mechanisms. The current

theory shows that this perspective can accommodate what people mean by causal verbs and how we use them to reason.

Note

1. We thank Phil Johnson-Laird for this point (personal communication, Nov. 20, 2007).

Acknowledgments

We thank Phil Fernbach Tom Griffiths, Phil Johnson-Laird, Julie Sedivy, and Phil Wolff for discussions on this topic. Jean-François Bonnefon pointed out a critical problem with the theory that has hopefully been corrected. This research was supported by NSF Award 0518147 to Steven Sloman. Experiment 1 was presented at the 2006 Psychonomic Society Annual Meeting.

References

Barbey, A. K., & Wolff, P. (2007). Learning causal structure from reasoning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 713–718). Mahwah, NJ: Erlbaum.

Bennett, J. (2003). *A philosophical guide to conditionals*. New York: Oxford University Press.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, *53*(3), 411–453.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Dowe, P. (2000). *Physical causation*. New York: Cambridge University Press.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, England: Oxford University Press.

Goldvarg, Y., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.

Halpern, J. Y., & Pearl, J. (2001). Causes and explanations: A structural-model approach—Part I: causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 194–202). San Francisco: Morgan Kaufmann.

Hart, H. L., & Honore, A. M. (1985). *Causation in the law*, 2nd ed. Oxford, England: Oxford University Press. (Original work published 1959).

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*, 65–81.

Johnson-Laird, P. N., & Byrne, R. (2002). Conditionals: A theory of meaning, pragmatics and inference. *Psychological Review*, *109*, 646–678.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

Kuhnmünch, G., & Beller, S. (2005). Distinguishing between causes and enabling conditions—through mental models or linguistic cues? *Cognitive Science*, *29*, 1077–1090.

Lewis, D. (1973). *Counterfactuals*. Oxford, England: Blackwell.

Lewis, D. (1986). *Philosophical papers: volume II*. Oxford, England: Oxford University Press.

Lewis, D. (2000). Causation as influence. In J. Collins, N. Hall & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 182–187). Cambridge, MA: MIT Press.

Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual, and covariational reasoning. *Journal of Experimental Psychology: General*, *132*, 419–434.

Mochon, D., & Sloman, S. A. (2004). Causal models frame interpretation of mathematical equations. *Psychonomic Bulletin & Review*, *11*, 1099–1104.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge, England: Cambridge University Press.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Sloman, S. A., & Lagnado, D. (2005). Do we ''do''? *Cognitive Science*, *29*, 5–39.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*, 49–100.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Motivation, and Cognition*, *31*, 216–227.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *1*, 82–111.

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, *47*, 276–332.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.