Cognitive Science 29 (2005) 5–39

Do We "do"?

Steven A. Sloman[a], David A. Lagnado[b]

[a]*Cognitive and Linguistic Sciences, Brown University*
[b]*Department of Psychology, University College, London*

Received 10 June 2003; received in revised form 20 May 2004; accepted 24 May 2004

Abstract

A normative framework for modeling causal and counterfactual reasoning has been proposed by Spirtes, Glymour, and Scheines (1993; cf. Pearl, 2000). The framework takes as fundamental that reasoning from observation and intervention differ. Intervention includes actual manipulation as well as counterfactual manipulation of a model via thought. To represent intervention, Pearl employed the *do* operator that simplifies the structure of a causal model by disconnecting an intervened-on variable from its normal causes. Construing the *do* operator as a psychological function affords predictions about how people reason when asked counterfactual questions about causal relations that we refer to as *undoing,* a family of effects that derive from the claim that intervened-on variables become independent of their normal causes. Six studies support the prediction for causal (A causes B) arguments but not consistently for parallel conditional (if A then B) ones. Two of the studies show that effects are treated as diagnostic when their values are observed but nondiagnostic when they are intervened on. These results cannot be explained by theories that do not distinguish interventions from other sorts of events.

Keywords: Causal models; Reasoning; Causal reasoning; Counterfactuals; Bayes nets

1. Introduction

Counterfactual reasoning can entail a type of imagined intervention. To reason causally about other possible worlds—about the way things might otherwise be or have been—requires changing a model of the actual world by making some counterfactual assumption thus creating a model of a different world. One way to do this is to treat a counterfactual assumption as if it were an actual intervention, that is, as an imagined intervention. To do so, some variable in the model of the actual world must be set to the value specified by the imaginary intervention.

This possibility has deep implications for theories of counterfactual reasoning about cause, because causal intervention has a logic all its own. What we can infer from intervened-on vari-

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

ables is different from what we can infer from variables whose states are observed. On one hand, we can infer a lot about effects under intervention. Indeed, the ability to infer the state of a manipulated variable's effects enables causal induction. We have known since Bacon (1620) that manipulation is a more effective way than observation to induce cause (see Reichenbach, 1956, for discussion) because intervention on an independent variable limits the possible reasons for a subsequent effect. On the other hand, intervention on a variable does not support inference about the variable's (normal) causes (Lewis, 1986; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). If an agent intervenes on a causal system to set a variable to some value, that value tells us nothing about the variables that normally cause it because they are not responsible on this occasion. We know nothing more about them than we did before the intervention. This induced independence of cause and effect predicts an effect we call *undoing.*

To illustrate, yellow teeth and lung disease are correlated across individuals because they are both influenced by smoking. If you intervene on the world by taking up smoking, then you should predict an increase in the probabilities that your teeth will be yellow and that you will get lung disease. Similarly, if you are astute enough not to smoke, then if you imagine taking up smoking, you should predict a corresponding increase in the probabilities of yellow teeth and lung disease in the imagined counterfactual world. A forward inference from cause (smoking) to effect is warranted. In contrast, if you whiten your teeth, or just imagine whitening your teeth, a backward inference is not warranted. Whitening your teeth will not by itself affect smoking behavior, and neither will it change the probability of lung cancer. Intervening on the effect renders it independent of its normal causes by undoing the relation between them. Notice that observation has no such effect. Observing that teeth are white does have diagnostic implications for smoking: If someone's teeth are white, he or she is less likely to be a smoker.

People are known to take advantage of the benefits of intervention when learning about causal structure (Gopnik et al., 2004; Lagnado & Sloman, 2004; Sobel, 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). In this article, we examine whether people obey this calculus of intervention when making inferences about counterfactuals in the context of causal relations. Do they undo the informational linkage of an effect from its causes? We contrast what people infer about a cause when its effect is governed by intervention as opposed to being observed. We also contrast inference in the context of causal relations to logical inference. *Intervention* is a causal concept relating to the effects of an agent's action. It is not a logical concept concerning truth preservation. Hence, the calculus of intervention should only apply in situations understood as causal. We investigate whether people are more likely to show undoing effects with unambiguously causal arguments than with logical ones whose causal status is less clear.

1.1. Representing causality: observation versus action (seeing versus doing)

Intervention has emerged as a key concept in decision making (e.g., Meek & Glymour, 1994; Nozick, 1995), the philosophy of explanation (e.g., Woodward, 2003), and graphical modeling of both probabilistic and deterministic processes (e.g., Pearl, 2000; Spirtes et al., 1993). We here provide a brief, essentially nontechnical introduction to the representation of intervention in graphical models.

Graphical representations of a causal system include a probability distribution over the variables of the system and a graph that represents the dependency structure of the distribution. Bayes nets are a particular form of graphical model that use only directed, acyclic graphs. An arc between two variables on a graph indicates marginal dependence between those variables. The absence of an arc implies independence or conditional independence. In particular, if no path exists from one variable to another, then each value of each variable is independent of all values of the other variable. One benefit of graphs is that they afford qualitative reasoning about dependency (e.g., how does the rate of unemployment depend on presidential decisions?). The benefits graphs provide to computation come from the fact that a graph implies an efficient factorization of the probability distribution.

To illustrate, say you believe that eating (T) is the cause of George's degree of fatness (G), and that it is George's mother's attitude towards food (M) that determines how much George eats. This could be represented by the following causal model:

(G1)

A path of causal influence obtains between M and G, although they are not linked directly (they are independent conditional on T). What this model implies is the following factorization: The probability distribution relating M, T, and G can be expressed in a relatively simple form requiring only one marginal probability, $\Pr\{M = m\}$, and two conditional probabilities:

$$P\{M = m, T = t, G = g\} = P\{M = m\} \bullet P\{T = t | M = m\} \bullet P\{G = g | T = t\}.$$

Pearl (1988) provides a technical introduction to Bayes nets with proof and explanation.

Spirtes et al. (1993; see also Pearl, 2000) used Bayes nets to construct a normative theoretical framework for reasoning, not just about dependency, but also about the more specific relation of cause (Glymour, 2001, and Sloman & Lagnado, 2004, provide accessible introductions). What distinguishes this framework from other graphical models is that arcs on causal graphs represent the mechanisms that determine the effects of intervention. To make transparent the distinction between observing a change in value and intervening to change the value of a variable, Pearl introduces a new operator, the *do*(•) operator. This distinction is what allows correlations to be distinguished from causal relations in the framework and permits causal reasoning about both actual and counterfactual events.

1.1.1. Seeing

Observation can be represented in the usual probabilistic way, using conditional probabilities defined over events. The probability of observing an event (say, that George is fat) under some circumstance (he eats too much) can be represented as the conditional probability that a random variable G is at some level *g* when T is observed to take some value *t*. The textbook definition of such a conditional probability is

$$P(G = g \mid T = t) = \frac{P\left(G = g \,\&\, T = t\right)}{P\left(T = t\right)}.$$

Conditional probabilities are generally invertible. As long as the probability of G is greater than 0, Bayes' rule tells us how to calculate the converse using marginal probabilities:

$$P(T = t \mid G = g) = P(G = g \mid T = t)\frac{P(T = t)}{P(G = g)} \tag{1}$$

Conditional probabilities are invertible because they reflect mere correlation. Whether understood as degrees-of-belief (Savage, 1954), willingness-to-bet (De Finetti, 1980), or relative frequencies (Von Mises, 1957), conditional probabilities are representations of the extent to which variables go together, regardless of why they go together.

1.1.2. Doing

Representing why variables are related requires a representation of how variables influence one another, their causal roles. A causal representation must tell us the effects of action. If I wiggle X, will Y change? In other words, a causal representation must be what Pearl (2000) calls an "oracle for intervention," a predictor of the effect of action. To represent action, Pearl uses the *do*(•) operator to set the value of a manipulated variable. *do*(X = x) has the effect of setting the variable X to the value x. Less obviously, the *do* operator also changes causal structure. Specifically, it *undoes* one or more causal relations: It changes the graph representing causal relations by removing any directed links from other variables to X (i.e., by cutting X off from the variables that normally cause it). The motivation is that an intervention by an external agent to set X = x renders other potential causes of x irrelevant. Most significantly, no diagnostic inference should be made about the normal causes of effect x because the agent is overriding their influence.

For example, intervening on the system represented by G1 to change George's degree of fatness (via surgery, say) would not affect how much he eats because there is no causal link from G to T. This can be corroborated by running an experiment: Take George to the operating table, change his degree of fatness and then measure how much he eats. Measurements taken in the context of action, as opposed to observation, would reflect the probability that T = t under the condition that *do*(G = *g*), an *interventional probability:*

$$P\{T = t \mid do(G = g)\}$$

obtained by, first, constructing a new causal model by removing any causal links to G:

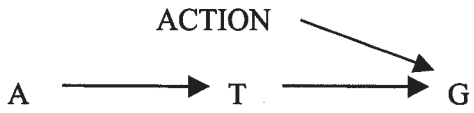
(G2)

Again, G is being caused by the agent, not by T, so the system should temporarily be represented without the causal mechanism linking T to G. Because the *do* operation removes the link between T and G in the graph, they are rendered probabilistically independent. The result is that G should be uninformative about T:

$$P\{T = t \mid do(G = g)\} = P(T = t).$$

In contrast, a causal relation could not in general be determined by just observing how much George eats and how fat he is because observation provides merely correlational information.

The effect of the *do* operator could alternatively be represented by introducing a new variable to represent an intervention that has the power to override the normal causes of the acted-on variable. Graphically, G2 could support the same inferences as the following graph:



where the value of G is a joint function of T and ACTION such that the effect of ACTION renders T irrelevant. Indeed, Spirtes et al. (1993) distinguish observation from intervention in this manner, by conditioning on distinctive "policy" variables. Such a representation has the advantage of generalizing to situations in which an action is not determinative but merely influences a variable along with its usual causes. Such cases are beyond the scope of this article, however. The critical disadvantage of such a representation is that it fails to represent the independence of the acted-on variable from its causes (the independence of T and G in this case). This independence is what simplifies both learning and inference in the context of action, and its depiction is one of the main purposes of graphical representations (Pearl, 1988). The issue here is merely one of depiction, whether the *do* operation is represented as removing a link or as adding a node along with a link. Either way, a special operation is required to distinguish the representations of action and observation. This special operation is the focus of this article.

1.2. Purpose of this article

We report six experiments intended to test whether people understand and are sensitive to the logic of intervention, the *do* operator in particular, when asked counterfactual questions about causal relations. In each case, we predict that a counterfactual event will not be treated as diagnostic of its normal causes (an *undoing* effect). The experiments examine two propositions: first, that people interpret interventions in a causal system by disconnecting the intervened-on variable from its causes. In other words, intervened-on variables are treated as independent of their causes. This implies both that intervention on an effect should not determine whether a cause could occur (tested in Experiment 1) or change what the value of the cause would be if known (Experiments 2 and 6). The article also examines some of the conditions under which people treat counterfactuals as interventions; in particular, whether a theory of counterfactual inference suggested by Pearl (2000) correctly predicts how people interpret subjunctive conditional questions about causal events. This is a focus of Experiments 3 to 5. Experiments 2 and 6 contrast intervention and observation directly.

Theories of belief updating that are not supplemented with the logic of intervention, whether probabilistic or deterministic, commonly treat effects as always diagnostic of their causes. This is illustrated by theories that employ Bayesian updating to determine the probability of a cause given its effect (e.g., the likelihood of a disease, given symptoms), from the probability of the effect given the cause, along with base rates of cause and effect (e.g., Anderson, 1990). This also characterizes theories that reduce causal reasoning to a form of conditional

reasoning (e.g., Goldvarg & Johnson-Laird, 2001). Such theories hold that argument forms such as modus tollens should hold:

If cause then effect.

No effect.

Therefore, no cause.

But as we have seen, such a conclusion is not warranted if the effect was prevented through intervention. The force of the logic of intervention is that it undermines the diagnosticity of effects whenever they come about through direct intervention on a causal system. However, the logic of intervention does not apply when reasoning is not intended to be causal. For instance, in a context of purely deductive inference (such as proving a theorem), modus tollens should always hold. In most of our experiments, we investigate how the framing of a counterfactual influences undoing effects by comparing reasoning about relations that are explicitly causal to reasoning about relations that could be interpreted as logical conditionals.

All experiments present participants with a set of premises and then ask them to judge the validity of a particular conclusion based on a counterfactual supposition. Alternative psychological theories for similar argument forms exist: mental model theory (Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Byrne, 2002), mental logic theory (e.g., Braine & O'Brien, 1991), and noninterventional Bayesian analysis. We consider each framework's fit to the data in the General discussion (Section 8). The causal modeling framework applies to both deterministic and probabilistic causal relations. Five of the experiments involve deterministic relations, and one experiment examines probabilistic arguments (Experiment 2).

2. Experiment 1

One implication of undoing is that causes are still able to occur even when effects are prevented. This experiment examines whether people are sensitive to this aspect of interventional logic by asking a counterfactual question about a simple causal scenario:

Causal. There are three billiard balls on a table that act in the following way: Ball 1's movement causes Ball 2 to move. Ball 2's movement causes Ball 3 to move.

The causal model underlying this scenario looks like this:

If an agent intervenes to prevent Ball 2 from moving, Ball 3 could no longer move because its cause is absent. As long as the action is attributed to the intervention of an agent from outside the system, the causal modeling framework states that the event should be represented as *do*(Ball 2 movement = no). As the cause of Ball 2 is no longer Ball 1 but the agent, a new causal model is relevant:

Ball 1 is rendered independent of Ball 2, and therefore Ball 1 should retain the ability that Ball 3 should lose, the ability to move. In particular, the causal modeling framework predicts, by virtue of undoing, that respondents should answer the question

(1) Imagine that Ball 2 could not move, would Ball 1 still move? Circle one of the 3 options: It could. It could not. I don't know.

by circling "It could." They should also answer

(2) Imagine that Ball 2 could not move, would Ball 3 still move? Circle one of the 3 options: It could. It could not. I don't know.

by circling "It could not."

The pair of predictions depends on the balls being related causally. They do not hold, for example, if the relation is construed as a logical conditional, such as the material or deontic conditional, although they should hold for causal conditionals that express the causal relations depicted. Specifically, we predict that they would hold for the

Causal conditional. There are three billiard balls on a table that act in the following way: If Ball 1 moves, then Ball 2 moves. If Ball 2 moves, then Ball 3 moves.

We also predict that they would not hold for the

Logical conditional. Someone is showing off her logical abilities. She is moving balls without breaking the following rules: If Ball 1 moves, then Ball 2 moves. If Ball 2 moves, then Ball 3 moves.

Predictions for the logical conditional depend on a theory of logical reasoning. We derive predictions for two such theories in the General discussion (Section 8). The causal modeling framework makes no claim that people will succeed at responding logically and merely allows that people may treat noncausal conditionals as different from causal conditionals or causal statements of any sort.

2.1. Method

The causal, causal conditional, and logical conditional scenarios presented previously were each tested on a different group of 20 participants. All 60 were students at the Université de Provence Aix-Marseilles. They were volunteers, approached on campus and presented with one of the three scenarios on a sheet of paper. The sheet also asked both of the counterfactual questions (identical in all conditions). The study was conducted in French. The French translations used can be found in the Appendix. Participants were given as much time as they desired to respond.

2.2. Results

Choice data are shown in Table 1. The results in the causal condition were as predicted. The vast majority responded that Ball 1 could move if Ball 2 could not (18 of 20 participants) and

Table 1
Percentages of participants giving each response to two questions in the three billiard ball problems of Experiment 1

	Yes	No	I don't know
Ball 1 moves			
Causal	90	5	5
Causal conditional	90	5	5
Logical conditional	45	55	0
Ball 3 moves			
Causal	5	90	5
Causal conditional	25	70	5
Logical conditional	30	70	0

Note. "Ball 1 moves" and "Ball 3 moves" refer to Questions 1 and 2, respectively.

that Ball 3 could not (also 90%). These percentages differ significantly from chance, $z = 5.96$; $p < .001$. An identical proportion responded that Ball 1 could move with the causal conditional, but only 45% did so with the logical conditional, also a highly significant difference, $z = 3.04$; $p < .01$. Despite great consistency when the relation between balls was causal, responses were highly variable when the relation was described as logical. More variability was observed in the response to the question about Ball 3 with both conditional scenarios. Most participants said that Ball 3 could not move, fewer with conditionals than in the causal condition, but this difference is not significant, 70% versus 90%; $z = 1.58$; *ns.*

2.3. Discussion

As predicted by the causal modeling framework, a strong undoing effect was observed with causal scenarios wherein the counterfactual absence of an effect was not treated as diagnostic of whether its cause could occur, whether the causal relations were described directly or using conditionals. Moreover, most participants correctly stated that an effect would not obtain without its cause. Responses under a logical frame were less consistent. Participants were split in judging whether Ball 1 could or could not move; 70% made an inference about Ball 3 consistent with causal logic (Ball 3 could not move), but 30% did not.

This experiment varied only a single scenario, and the results may, of course, depend on attributes of the specific scenarios. In particular, the ease of imagining independent mechanisms governing cause and effect may be important. Balls are easily imagined to move independently. The effect might not have occurred, or might have been considerably weaker, had the relation concerned a chain causing the wheel of a bicycle to turn.[1] In such a case, people might also assume that the wheel causes the chain to move. All we can conclude from this experiment is that most people can treat effects as nondiagnostic of their causes, not that they always do.

In our causal conditions, one might argue that our participants did not draw inferences from Ball 2's lack of movement, not because they disconnected Ball 2 from its cause, Ball 1, but because they assumed that Ball 2 required a further enabling condition beyond Ball 1, whose absence prevented Ball 2, but not Ball 1, from moving. This would be an instance of the ACTION variable discussed previously. If the sole motivation of this enabling condition were to isolate

Ball 2 from its causes, then it would be equivalent to the *do* operation, although it would be a representation that does not make the induced independence between Balls 1 and 2 explicit.

3. Experiment 2

The causal modeling framework applies to probabilistic arguments as well as deterministic ones. Indeed, the logic of the *do* operation is identical for the two types, and therefore the undoing effect should hold in both. Experiment 2 extends the effect to probabilistic arguments. In accordance with this shift from deterministic to probabilistic arguments, a probability response scale was used. Experiment 2 also attempts to generalize the effect of undoing from Experiment 1. Experiment 1 shows that intervention on an effect makes the effect nondiagnostic of whether its cause *could* happen. Experiment 2 examines if intervention renders the effect nondiagnostic of whether the cause *would* have happened given that it did.

The experiment uses the same simple chain structure as Experiment 1:

Three variables were crossed factorially, Type of counterfactual question (interventional, observational, unspecified) × Type of relation in the premises (causal versus conditional) × Scenario (abstract, rocket ship, smoking). The type of counterfactual was manipulated to assess the causal modeling framework's prediction of an undoing effect with intervention, but not observation. The unspecified case is a control condition in which the origin of the relevant variable value is not specified.

> In the abstract causal condition participants were given the following premise set:
> When A happens, it causes B most of the time.
> When B happens, it causes C most of the time.
> A happened.
> C happened.

The interventional questions asked the following with a 1 to 5 response scale:

(i) Someone intervened directly on B, preventing it from happening. What is the probability that C would have happened?
(ii) Someone intervened directly on B, preventing it from happening. What is the probability that A would have happened?

The causal modeling framework predicts an undoing effect in Question (ii). When assessing a counterfactual that supposes that an agent's action prevented B from occurring, participants should mentally sever the link from A to B and, thus, not treat B's absence as diagnostic of A. Because they were told that A had happened, this would correspond to relatively high responses, potentially near the extreme of the response scale (5) and, at minimum, greater than the midpoint (3). In contrast, responses to Question (i) should show a reduction in belief about the occurrence of C. The intact causal link from B to C, coupled with the counterfactual supposition that B does not occur, should lead to responses at or below the scale midpoint.

We contrast this to observational questions:

(i) What is the probability that C would have happened if we observed that B did not happen?
(ii) What is the probability that A would have happened if we observed that B did not happen?

Unlike the interventional case, these questions explicitly state that B's nonoccurrence was observed, implying that B was not intervened on. Therefore, B should be treated as diagnostic of A; we do not expect undoing. The judged probability of A, Question (ii), should be substantially lower than with interventional questions. B's nonoccurrence makes C less likely, so the judged probability in Question (i) should again be low.

The unspecified (control) probability questions were as follows:

(i) What is the probability that C would have happened if B had not happened?
(ii) What is the probability that A would have happened if B had not happened?

The answer to Question (i) should again be low for the same reasons as the other cases. The answer to Question (ii) will reveal people's propensity to treat probability questions with modal verbs such as "hadn't" as interventional versus observational. We expect responses to be in between those of the interventional and observational conditions.

We again included corresponding conditional versions to compare causal to logical reasoning frames. Abstract conditional premises were as follows:

If A is true, then B is likely to be true.
If B is true, then C is likely to be true.
A is true.
C is true.

Corresponding interventional questions were

(i) Someone intervened and made B false. What is the probability that C would be true?
(ii) Someone intervened and made B false. What is the probability that A would be true?

If people systematically infer the contrapositive with conditionals then their responses to Question (ii) should be low, consistent with propositional logic. But if participants use the fact of intervention as a cue to interpret the conditional statements causally, we should instead see the undoing effect; responses should prove compatible with causal logic. The correct response to Question (i) is ambiguous. The second premise has no implications when B is false, so people might infer that C remains true, or else they might be confused and just choose to express uncertainty. To fully compare causal and logical frames, we also included observational conditionals and unspecified conditionals, created using the previously mentioned conditional premises, but asking probability questions that parallel the questions asked with causal premises for corresponding conditions.

We tested each of these six conditions using three different scenarios, the previously mentioned abstract scenario and a scenario concerning physical causality:

When there is gas in the rocket ship's fuel tank, it causes the engine to fire most of the time.
When the engine fires, most of the time it causes the rocket ship to take off.

The rocket ship's fuel tank has gas in it.
The rocket ship takes off.

Using a medical scenario:

Smoking causes cancer most of the time.
Cancer causes hospitalization most of the time.
Joe smokes.
Joe is hospitalized.

In sum, Experiment 2 contrasts causal to conditional premises, examines three varieties of observation or intervention, and uses three different scenarios, each of a different type, all in the context of probabilistic premises. It also uses a probability response scale, allowing examination of the undoing effect, even when people have the option of expressing complete uncertainty by using the midpoint of the scale.

3.1. Method

3.1.1. Design

All variables were combined factorially: Causal versus conditional premises × Type of intervention (intervention, observation, unspecified) × Scenario (abstract, rocket ship, smoking). All variables were manipulated between-participants except scenario. For half the scenarios, the question about the first variable (A in the abstract scenario) came before the other question; for the other half, question order was reversed. The order of scenarios was roughly counterbalanced across participants.

3.1.2. Participants

We tested 151 Brown University undergraduates using the same questionnaire format as previous studies. We also tested 125 volunteer participants on the Internet, using an identical questionnaire. They were obtained by advertising on various Web sites related to psychological science. We obtained no identifying information about these participants. An approximately equal number of Web and non-Web participants were tested in each condition.

3.1.3. Procedure

The instructions urged participants to assume that the relations presented were the only ones relevant by stating at the outset of each problem, "Please treat the following as facts. Assume that there are no factors involved outside of those described below." The instructions for the response scale read, "Please respond to the following questions, using an integer scale from 1 to 5 where: 1 = *very low,* 2 = *low,* 3 = *medium,* 4 = *high,* 5 = *very high.*" Participants worked at their own pace and were given as much time as they desired to answer the questions.

3.2. Results

Brown University students and Web participants gave the same pattern of responses, and therefore we collapsed their data. Mean probability judgments are shown in Table 2 averaged

Table 2
Mean probability judgments on a 1–5 scale for two questions in Experiment 2

	Causal Version		Conditional Version	
Type of Counterfactual	*P(C|~B)*	*P(A|~B)*	*P(C|~B)*	*P(A|~B)*
Interventional	2.3	3.9	3.0	4.1
Observational	2.3	2.7	2.8	3.3
Unspecified	2.4	3.2	2.6	3.0

Note. *P(C|~B)* = Question i; *P(A|~B)* = Question ii.

across the three scenarios. The overall patterns were similar across scenarios except that judgments in the rocket ship scenario tended to be lower than for the other scenarios, especially for the question about Variable C (concerning whether the rocket ship would take off if the engine fired).

When questions were interventional, undoing was observed in both the causal and conditional versions. The mean judged P(A|~B) were appreciably higher than the scale midpoint, 3.9 and 4.1, respectively, $t(48) = 7.75$; standard error [*SE*] = .12 and $t(47) = 8.32$; *SE* = .13; both *p*s < .0001. Intervening explicitly to prevent B did not cause participants to eliminate belief in the occurrence or truth of A. More important, the judged probability of A was higher in the interventional than in the observational conditions, $t(96) = 6.86$; *SE* = .17 and $t(93) = 3.64$; *SE* = .22 in the causal and conditional cases, respectively, both *p*s < .001.

In the causal case, the nonoccurrence of B suggested to participants that its effect did not occur either: mean P(C|~B) of 2.3, significantly lower than 3, $t(48) = 4.36$; *SE* = .15; $p = .0001$. In the conditional case, the probability of C, given that its antecedent B was made false, was judged completely unknown (the scale midpoint), even though participants had been told that C was true. The difference between causal and conditional responses to the question about C may result from a few logically sophisticated participants who realized that B's falsity has no bearing on the truth of C with the conditional version, even though B's nonoccurrence did suggest the nonoccurrence of C with the causal version.

With observational questions, the pattern in the causal and conditional versions was similar. P(A|~B) judgments (2.7 and 3.3 in the causal and conditional conditions, respectively) were not statistically distinguishable from the midpoint of the scale, $t(48) = 2.23$; *SE* = .13 and $t(46) = 1.58$; *SE* = .18, both *ns.* Moreover, these were both higher than corresponding P(C|~B) judgments, $t(48) = 3.19$; *SE* = .12 and $t(46) = 3.28$; *SE* = .13, both *p*s < .01. In other words, in the observational condition, the negation of B was treated as removing any evidence in favor of A, but as providing evidence against C. Consistent with the causal modeling framework, participants treated observations as correlational evidence and did not exhibit an undoing effect. Instead, they treated observations that B did not occur as diagnostic that A did not occur and predictive that C would not occur.

When the nature of the intervention was unspecified, again little difference was observed between the causal and conditional conditions. The undoing effect was again not significant in either condition in the sense that the mean P(A|~B) judgments (3.2 and 3.0, respectively) did not differ from the midpoint of the response scale (3), $t(41) = 1.5$; *SE* = .16; *ns,* and $t < 1$, re-

spectively. Participants were not sure about Event A when told B had not happened or that B was false. However, both judgments were higher than corresponding P(C|~B) judgments, $t(41) = 5.09$; $SE = .17$; $p < .0001$ and $t(40) = 3.40$; $SE = .13$; $p < .01$, respectively, suggesting that the negation of B did reduce belief in the occurrence or truth of C to some extent, consistent with a causal reading of the B–C relation. Overall, the data suggest that participants treated the simple negation of B as observational, not interventional.

The parallel tendencies among the probability judgments in the causal and conditional conditions and their consistency with the causal modeling framework suggest that, in this experiment, the conditional relations tended to be interpreted as causal. This is a natural interpretation, particularly for the medical and rocket ship scenarios.

3.3. Discussion

In Experiment 2, the undoing effect manifested as a higher probability judgment that the cause occurred when the effect was prevented than when the effect was observed to be absent. This is weaker than the effect predicted by the causal modeling framework, that a prevented effect should be independent of its cause. On this view, the probability of A should have been the same as it was before supposing that B did not occur and that A was known to have occurred. So the judged probability of A should have been very high. Should participants have reported the maximal value of 5? Perhaps, but such a prescription assumes absolute confidence in understanding the causal model and the described events, and in the interpretation of the questions. It also assumes that participants are willing to use the endpoints of the response scale. Our finding is that on average participants were only willing to choose the second highest response scale value (4). As such, this is evidence only for a weak form of the undoing effect, that intervention renders the effect less diagnostic of its causes than observation does.

The fact that negating B caused people to reduce the judged probability of C suggests that people were interpreting these scenarios causally. Unlike Experiment 1, this happened even when the statements were conditional. Obviously, conditional statements can indicate causal relations. Our post hoc guess is that our rocket ship and medical scenarios elicited such an interpretation. The probabilistic context may itself have contributed to the causal reading because a probabilistic relation is not easily understood as logical.

4. Experiment 3

We now examine a question beyond whether people obey the logic of intervention and ask whether the *do* calculus is consistent with the way people interpret counterfactual questions when intervention is not explicit. Pearl (2000) presented a theory of inference from counterfactual assumptions that states that counterfactual probabilities are derived using a three-step procedure of abduction, action, and prediction: (a) abduction (updating): beliefs about the world are updated by taking into account all evidence given in the context; (b) action (the *do* operation): variables representing the counterfactually manipulated state are changed to reflect the counterfactual assumption, and a new causal model (of the counterfactual world) is created by modifying a model of the actual world through undoing: The counterfactually changed vari-

ables are disconnected from their normal causes; (c) prediction (inference): counterfactual reasoning occurs over the new model represented by the surgically altered network using updated knowledge. In other words, Pearl spells out how to reason about counterfactuals, imagined intervention on a causal model, in analogy to reasoning about actual intervention. His theory turns out to be a more precise, testable version of Lewis's (1986) theory of counterfactuals.

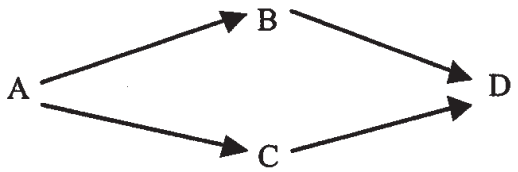
Consider the following set of causal premises in which A, B, C, and D are the only relevant events:

A causes B.
A causes C.
B causes D.
C causes D.
D definitely occurred.

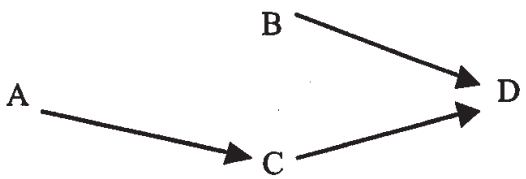
On the basis of these facts, answer the following two questions:

(i) If B had not occurred, would D still have occurred?___ (yes or no)
(ii) If B had not occurred, would A have occurred?___ (yes or no)

Answers to these questions can be derived from Pearl's (2000) analysis of counterfactuals. The causal relations among the variables can be depicted as follows:



In Step 1, abduction, we update our beliefs from the givens. We know that D has occurred. This implies that B or C or both occurred, which in turn implies that A must have occurred. A is the only available explanation for D. Because A occurred, B and C both must have occurred. Therefore, all four events have occurred. The rules of ordinary logic are sufficient to update our model. But in Step 2, action, we represent what would have happened if B had not occurred. We apply the *do* operator, *do*(B = did not occur), with the effect of severing the links to B from its causes:



Therefore, in Step 3, prediction, we should not draw any inferences about A from the absence of B. So the answer to Question (ii) is "yes" because we had already determined that A occurred, and we have no reason to change our minds. The answer to Question (i) is also "yes" because A occurred, and we know A causes C, which is sufficient for D.

To compare causal and logical frames, we again include a conditional condition that uses a conditional premise set. Consider

If A then B.
If A then C.
If B then D.
If C then D.
D is true.

Along with the questions:

(i) If B were false, would D still be true? ___ (yes or no)
(ii) If B were false, would A be true? ___ (yes or no)

The causal modeling framework makes no prediction about such premises, except that, because they do not necessarily concern causal relations, responses could well be different from those for the causal premises. Of course, if the context supports a causal interpretation, then they should elicit the same behavior as the causal set. For example, we include a problem in the following section with clear causal context (the robot problem). It would not be surprising if the undoing effect arose in that condition.

4.1. Method

4.1.1. Materials

Three scenarios were used, each with a causal and a conditional version. One scenario (abstract) used the premise sets just shown, involving causal or conditional relations between letters with no real semantic content. Two additional scenarios with identical causal or logical structure and clear semantic content were also used. One pair of premise sets concerned a robot. The causal version of that problem read:

A certain robot is activated by 100 (or more) units of light energy. A 500-unit beam of light is shone through a prism, which splits the beam into two parts of equal energy, Beam A and Beam B, each now traveling in a new direction. Beam A strikes a solar panel connected to the robot with some 250 units of energy, causing the robot's activation. Beam B simultaneously strikes another solar panel also connected to the robot. Beam B also contains around 250 units of light energy, enough to cause activation. Not surprisingly, the robot has been activated.

(1) If Beam B had not struck the solar panel, would the robot have been activated?

(2) If Beam B had not struck the solar panel, would the original (500-unit) beam have been shone through the prism?

The conditional version was parallel, except that causal statements were replaced by if … then … statements:

A certain robot is activated by 100 (or more) units of light energy. If a 500-unit beam of light is split into two equal beams by a prism, one of these beams, Beam A, will strike a solar panel connected to the robot with some 250 units of energy. If the 500-unit beam of

light is split into two equal beams by a prism, the second of these beams, Beam B, will strike a second solar panel connected to the robot with some 250 units of energy. If Beam A strikes the first solar panel, the robot will be activated. If Beam B strikes the second solar panel, the robot will be activated. The robot is activated.

(1) If Beam B had not struck the solar panel, would the original (500-unit) beam have passed through the prism?

(2) If Beam B had not struck the solar panel, would the robot have been activated?

The third scenario involved political antagonisms among three states. Here is the causal version:

Germany's undue aggression has caused France to declare war. Germany's undue aggression has caused England to declare war. France's declaration causes Germany to declare war. England's declaration causes Germany to declare war. And so, Germany declares war.

(1) If England had not declared war, would Germany have declared war?

(2) If England had not declared war, would Germany have been aggressive?

And here is the conditional version:

If Germany is unduly aggressive, then France will declare war. If Germany is unduly aggressive, then England will declare war. If France declares war, Germany will declare war. If England declares war, Germany will declare war. Germany has declared war.

(1) If England had not declared war, would Germany have declared war?

(2) If England had not declared war, would Germany have been aggressive?

4.1.2. Participants and procedure

238 University of Texas at Austin undergraduates were shown all three scenarios in questionnaire format, 118 the causal versions and 120 the conditional versions. Scenario order was counterbalanced across participants. Participants circled either "Yes" or "No" to answer each question and were then asked to rate their confidence in their decision on a scale from 1 (*completely unsure*) to 7 (*completely certain*). Otherwise, the procedure was identical to previous experiments.

4.2. Results and discussion

Percentages of participants responding "yes" to each question are shown in Table 3. A very different pattern can be observed for the causal and conditional statements. The undoing hypothesis correctly predicted the responses to the causal premises, the great majority being "yes." The responses to the conditional premises were more variable. For each question in each scenario, the proportion of "yes" responses was significantly higher for causal than for conditional versions (all ps $< .01$ by z test). Moreover, all of the causal but only one of the conditional percentages was greater than chance (50%; $p < .001$), the exception being whether D would hold in the robot scenario. Some participants may have interpreted the "if–then" connectives of

Table 3
Percentages of participants responding "yes" to two questions about each scenario in both Causal and Conditional conditions of Experiment 3

	Causal		Conditional	
Scenario	D Holds	A Holds	D Holds	A Holds
Abstract	80	79	57	36
Robot	80	71	63	55
Political	75	90	54	47

Note. "D holds" and "A holds" refer to questions about variables D and A respectively in the Abstract scenario and corresponding questions for the Robot and Political scenarios.

the conditional version as causal, especially for this problem. The clear physical causality of the robot suggests a causal interpretation.

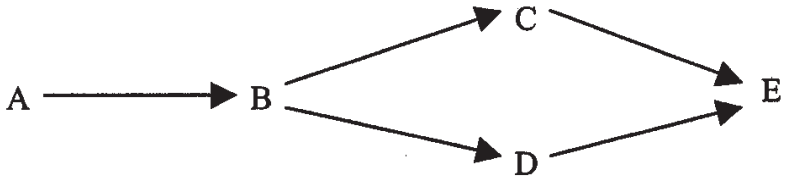
The predominance of "yes" responses in the causal condition implies that for the majority of participants the supposition that B did not occur did not influence their beliefs about whether A or D occurred. This is consistent with the idea that these participants mentally severed the causal link between A and B and thus did not draw new conclusions about A or about the effects of A from a counterfactual assumption about B. The response variability for the conditional premises suggests that no one strategy dominated for interpreting and reasoning with conditional statements. These conclusions are supported by the confidence judgments. Participants were highly confident when answering causal questions (mean of 6.0 on the 1–7 scale). They were appreciably less confident when answering conditional questions (mean of 5.4), $t(236) = 4.77$, $SE = .13$, $p < .0001$.

The order of scenarios had no effect on the thematic causal problems. However, the abstract scenario did show a small effect. The undoing effect was greater with the abstract scenario when it followed the thematic problems (85% "yes" responses) than when it came first (69%; $z = 2.00$, $p << .05$). A small number of participants may have been encouraged by the thematic cases to read the abstract counterfactual as interventional. No such effect occurred with conditional problems.

5. Experiment 4

One might argue that the difference between the causal and conditional versions in Experiment 3 is not due to a greater tendency to counterfactually decouple variables from their causes in the causal over the conditional context, but instead to different presuppositions of the two contexts. In particular, the causal context presupposes the occurrence of A more than the conditional context presupposes the truth of A. If so, then the greater likelihood of saying "yes" to the A question in the causal scenarios could be due to these different presuppositions rather than to differential likelihoods of treating B as nondiagnostic of A. And if people consider A more likely, then they might also be expected to be more likely to confirm the occurrence of D.

To control for this possibility as well as to replicate the findings of Experiment 3, we examined causal and conditional versions of premises with the following structure:



Participants were told not only that the final effect, E, had occurred, but also that the initial cause, A, had too. This should eliminate any difference in presupposition of the initial variable because its value is made explicit. To illustrate, here is the causal version of the abstract problem:

A causes B.
B causes C.
B causes D.
C causes E.
D causes E.
A definitely occurred.
E definitely occurred.
(i) If D did not occur, would E still have occurred?
(ii) If D did not occur, would B still have occurred?

To examine the robustness of the undoing effect, the mood of the antecedent ("did not occur") had slightly less counterfactual emphasis than in the previous experiment ("had not occurred").

Following essentially the same logic spelled out to derive Experiment 3's predictions, Pearl's (2000) theory predicts that a counterfactual assumption about D should disconnect it from B in the causal context, rendering it nondiagnostic of B so that participants should answer "yes" to both questions. Again, a parallel conditional version was also used. Participants should only answer "yes" in the conditional context if they interpret the problem causally. Note that, like previous experiments, a material conditional account of the meaning of *cause* must predict no difference between the causal and conditional contexts.

5.1. Method

Two groups of 20 Brown University undergraduates each received either the causal or conditional versions of the abstract, robot, and politics problems described previously, but modified so that the occurrence/truth of the variable corresponding to B in the example was disambiguated by adding a fifth variable. Because of concerns about the clarity of the political problem in Experiment 3, it was revised for this experiment. Here is the causal version:

Brazil's undue aggressiveness is a consequence of its political instability. Brazil's undue aggression causes Chile to declare war. Brazil's undue aggression causes Argentina to declare war. Chile's declaration causes Brazil to declare war. Argentina's declaration causes Brazil to declare war. Brazil is in fact politically unstable. Brazil declares war.

Otherwise, the method and materials were identical to that of Experiment 3.

5.2. Results

The results, shown in Table 4, are comparable to those of Experiment 3, although the proportion of "yes" responses was lower for one of the robot questions, whether the beam was shining if the solar panel had not been struck (only 55). Overall, the experiment provides further evidence of undoing for causal relations. Five of six percentages were significantly greater than 50% in the causal condition (all those greater than or equal to 70). Only two of six reached significance in the conditional case, with values of 75 and 80. A difference between causal and conditional premises was again observed for abstract and political premises, $z = 2.20$, $p = .01$, and $z = 2.00$, $p = .02$, respectively, but not for robot ones, $z = 1.18$, *ns*. Confidence judgments were again higher for answers to causal questions (mean of 5.89) than for answers to conditional questions (mean of 5.23), $t(38) = 2.30$, $SE = .27$, $p < .05$.

5.3. Discussion

The replication of the undoing effect in this experiment suggests that the earlier results should not be attributed to different pragmatic implicatures from causal and conditional contexts. Any differences between Experiments 3 and 4, especially the absence of the undoing effect for the one robot question, could be due to a different participant population, a smaller sample size in this study, some proportion of participants failing to establish an accurate causal model with these more complicated scenarios, or participants not implementing the undoing operation in the expected way (i.e., not mentally disconnecting B from D). Failure to undo is plausible for these problems because D's nonoccurrence is not definitively counterfactual. The question read, "If D did not occur," which does not state why D did not occur; the reason is left ambiguous. One possibility is that D did not occur because B did not. Nothing in the problem explicitly states that the nonoccurrence of D should not be treated as diagnostic of the nonoccurrence of B, although it remains the most likely interpretation, especially because the consequent clauses used the word "still" (e.g., "would E still have occurred?"), which clearly suggests a counterfactual interpretation.

Table 4
Percentages of participants responding "yes" to two questions about each scenario in both Causal and Conditional conditions of Experiment 4

	Causal		Conditional	
Scenario	E Holds	B Holds	E Holds	B Holds
Abstract	70	74	45	50
Robot	90	55	75	45
Political	75	90	45	80

Note. "E holds" and "B holds" refer to questions about variables E and B respectively in the Abstract scenario and corresponding questions for the Robot and Political scenarios.

6. Experiment 5

This experiment examines directly the relation between interventions and counterfactuals. We compare the degree of undoing in a case of explicit prevention to a case where the reason for the negation of the effect is vague; it could be interpreted in terms of either prevention or observation. Experiment 5 also uses the simplest possible causal model, A causes B, with no other relevant variables. If the undoing effect is mediated by a deeply rooted psychological mechanism, then it should arise when people are asked about it directly.

The following scenario states the relation between A and B using an if–then construction:

> All rocket ships have two components, A and B. Component A causes Component B to operate. In other words, if A, then B.

The scenario can be represented with the causal graph

In an explicit prevention condition we asked

(i) Suppose Component B were prevented from operating, would Component A still operate?
(ii) Suppose Component A were prevented from operating, would Component B still operate?

The causal modeling framework predicts an undoing effect, that participants will say "yes" to question (i) about Component A because A should be disconnected from B by virtue of the counterfactual prevention of B. It also predicts a "no" response to the second question. If A is the cause of B, then B should not operate if A does not.

Will we also find undoing when B is negated with a counterfactual supposition without explicit prevention? If so, will the effect be as strong as it is with prevention? To find out, we introduced a nonexplicit counterfactual condition differing only in the questions asked:

(i) Suppose Component B were not operating, would Component A still operate?
(ii) Suppose Component A were not operating, would Component B still operate?

Based on Experiments 3 and 4, we expected to find an undoing effect—the same pattern as with explicit prevention—even with this very simple causal model. The question is how the magnitude of the effect will vary across conditions.

6.1. Method

Approximately half of 78 Brown undergraduates were given the counterfactual and the rest the explicit prevention questions. Half of each group were given the previously shown scenario and half were shown an identical scenario except that the roles of components A and B were reversed. Otherwise, the method was identical to that of previous experiments. Participants were never shown the causal graph.

6.2. Results and discussion

The results, averaged over the "If A then B" and "If B then A" groups (and described in terms of the former), are shown in Table 5. The 68% giving an affirmative answer to the first question in the counterfactual condition replicates the undoing effects seen in the previous studies. The even greater percentage (89%, $z = 2.35$, $p < .01$) in the explicit prevention condition shows that the undoing effect is even greater when the interventional nature of the antecedent is made explicit. Responses to the second question were almost all negative, demonstrating that people clearly understood that the relevant relation was causal. In this experiment, confidence was uniformly high (approximately 6) in all conditions.

The undoing effect here is the finding that people responded "yes" to the question "Suppose Component B were not operating, would Component A still operate?" on the assumption that *still operate* in this context means *to be operational.* But the predicate permits an unintended reading. Perhaps people understood it to mean *to be potentially operational,* for example, to not be broken. In that case, a "yes" response would be appropriate, not because of the *do* operator, but because the mere fact that B is operating would not entail that A is broken; A would still have the potential to operate whether or not it is able to at the moment. This alternative interpretation is defeated, however, by responses to the second question. If participants interpreted this question to be asking whether B would be in working order if A were not operating, then they should have answered "yes." The fact that 97% of them said "no" implies that they understood *to operate* as originally intended, as *operational at the moment.* In any case, this is tested further in Experiment 6.

7. Experiment 6

This experiment is like Experiment 5 except that (i), instead of predicating the rocket ship components ambiguously with the verb *to operate,* unambiguously asks whether components *are moving;* and (ii), instead of contrasting an explicit prevention to a counterfactual question, contrasts an explicit prevention to an explicit observation question. The causal modeling framework expects the undoing effect with prevention, but not observation (as in Experiment 2). Therefore, if the effect does occur less often with observation questions, that would suggest that the explicit prevention effect is a result of interventional logic and not any other source of bias to treat the effect as nondiagnostic.

Table 5
Percentages of participants responding "yes" to questions in the Rocketship scenario of Experiment 5 given Counterfactual and Explicit Prevention questions

Question	Counterfactual	Explicit Prevention
i. if not B, then A?	68	89
ii. if not A, then B?	2.6	5.3

7.1. Method

Participants were obtained by advertising on a campus-based electronic newspaper. For doing this and an unrelated questionnaire, they were given a small chance to win a $40 prize. The explicit prevention problem was changed to the following:

> All rocket ships have two components, A and B. Movement of Component A causes Component B to move. In other words, if A, then B. Both are moving.
>
> (i) Suppose Component A were prevented from moving, would Component B still be moving?
> (ii) Suppose Component B were prevented from moving, would Component A still be moving?

The explicit observation scenario was identical, but the following questions were asked:

> (i) Suppose Component A were observed to not be moving, would Component B still be moving?
> (ii) Suppose Component B were observed to not be moving, would Component A still be moving?

Twenty-six participants were given the explicit prevention, and 23 explicit observation questions. All scenarios stipulated that A was the cause of B. Question (i) preceded Question (ii) for approximately half the participants and vice versa for the other half. Otherwise, the method was identical to Experiment 5.

7.2. Results and discussion

The explicit prevention condition replicated Experiment 5 (see Table 6) in that a strong undoing effect occurred (85% "yes"), but a greater proportion of people said that B would move even if A were prevented, although still a minority (19%). In the explicit observation condition, only a few participants showed undoing (22%), significantly fewer than explicit prevention ($z = 4.42$, $p < .001$). The presence of undoing with explicit prevention implies that the undoing effect in Experiment 5 is not due to the ambiguity of the predicate. The clear difference between prevention and observation suggests that the effect is not due to a general bias to treat effects as nondiagnostic, but to treat them as nondiagnostic only when they have been intervened on.

Table 6
Percentages of participants responding "yes" to questions in the Rocketship scenario of Experiment 6 given Explicit Prevention versus Explicit Observation questions

Question	Explicit Prevention	Explicit Observation
i. if not B, then A?	85	22
ii. if not A, then B?	19	30

8. General discussion

Six experiments have shown effects of undoing, that intervention on an effect reduces its dependence on its normal causes and therefore renders it less diagnostic of whether the cause could change (Experiments 1 and 5) or whether it would change from a current value (Experiments 2, 3, 4, and 6). When the intervention explicitly involved prevention (Experiments 1, 2, 5, and 6), effects were large and robust. People understand the logic of intervention. When given a counterfactual in the form of a subjunctive conditional (Experiment 3), an indicative conditional with a subjunctive consequent (Experiment 4), or a supposition (Experiment 5), a majority of people still behaved in accordance with interventional logic, although not as many as in the explicit cases. The evidence suggests that intervention on an effect is understood to render the effect independent of its causes.

Undoing effects were observed with a range of causal models, in a range of problem contexts, and with both deterministic (Experiments 1, 3–6) and probabilistic (Experiment 2) arguments. The studies also validated that the causal relations were interpreted as causal by showing that effects were judged not to occur—or to occur at their base rate—if their sole causes did not (Experiments 1, 2, 5, and 6). Experiments 2 and 6 showed that undoing obtained after an intervention that prevented an event from occurring, but not after merely observing the nonoccurrence of an event. Finally, Experiments 1, 3, and 4 showed that a causal statement (A causes B) is not necessarily reasoned about in the same way as a conditional statement (if A then B). However, conditionals might have been interpreted as causal in certain contexts. In general, conditionals were not given a consistent interpretation.

The data clearly show that counterfactual statements are sometimes interpreted as causal interventions, but the conditions favoring this interpretation are still not entirely clear. One necessary condition is that the counterfactual concerns entities that are understood to be causally related. Hence, conditional statements concerning entities with known causal relations are more prone to undoing (Experiments 2, 3, and 4). This is not sufficient, however. The counterfactual must be marked as interventional because a counterfactual can refer to an observation, ("If the patient had lived longer, then my diagnosis would have been different"; see Experiments 2 and 5).

Most people obey to some extent a rational rule of counterfactual inference, the undoing principle. In every case in which a causal relation existed from A to B, and B was counterfactually prevented, the majority of people judged that A could or would still occur. Put this way, undoing seems obvious. When reasoning about the consequences of an external intervention via a counterfactual supposition, most people do not change their beliefs about the state of the normal causes of the event. They reason as if the mentally changed event is disconnected and therefore not diagnostic of its causes. This is a rational principle of inference because an effect is indeed not diagnostic of its causes whenever the effect is not being generated by those causes but instead by mental or physical intervention by an agent outside the normal causal system. To illustrate, when a drug is used to relax a patient, one should not assume that the reasons for the patient's anxiety are no longer present. Indeed, the need to use a drug to relax a patient may increase belief that the patient's anxiety has some deeper source (Hilton, personal communication, October 1, 2003).

We now consider various formal explanations of our results, including theories of mental logic, mental models, noninterventional and interventional Bayesian probability.

8.1. Mental logic

To our knowledge, the psychology literature does not contain a logical-rule theory of reasoning with causal relations (although Rips & Marcus, 1977, discussed conditional reasoning in the context of suppositions). Braine and O'Brien (1991) did offer a theory of how English speakers reason with the word *if.* A modest extension of the theory makes predictions about our nonprobabilistic conditional arguments. The theory posits a lexical entry that includes two inference schemas, Modus ponens and a schema for conditional proof. It also posits a reasoning program and pragmatic principles. Rather than describe the theory in detail, we derive predictions for Experiment 1.

Consider the question:

(1) Imagine that Ball 2 could not move, would Ball 1 still move?

with the response options "It could," "It could not," and "I don't know." We assume that the question is equivalent to "If Ball 2 could not move, would Ball 1 still move?"—otherwise the theory makes no prediction. Given the premise "If A then B," Braine and O'Brien (1991) used their theory to derive the contrapositive "If not B then not A." The theory does not distinguish reasoning about counterfactual and indicative conditionals. Therefore, if we substitute "Ball 1 would move" for A and "Ball 2 can move" for B, the contrapositive implies that Ball 1 would not move. But the fact that Ball 1 *would* not move does not imply that Ball 1 *could* not move, and we can assume that the scenarios all presuppose that the balls can move initially. So the theory is consistent with the modal causal conditional answer that Ball 1 could move. However, only 45% gave this response for the logical conditional. Therefore, the theory fails by not distinguishing causal from other kinds of conditionals. The theory does allow pragmatic principles to affect inference, but none of the principles offered by Braine and O'Brien (1991) would change this particular prediction.

We interpret the remaining question

(2) Imagine that Ball 2 could not move, would Ball 3 still move?

as "If Ball 2 could not move, would Ball 3 still move?" Response options again concerned whether Ball 3 could move. There is no way to derive that Ball 3 *could* not move because it could move even if it happens not to. So the theory is unable to explain why people conclude that Ball 3 could not move in all three conditions.

In Experiment 3, participants were asked

QA: If B were false, would A be true?

One of our premises was If A then B. Braine and O'Brien's (1991) derivation of the contrapositive implies that the answer to QA should be "no." A majority of participants gave this answer to the abstract problem, but only half did so to the robot and political problems. The analog of this question in Experiment 4 is

QB: If D were false, would B still be true?

The same analysis applies except for the presence of the premises "If A then B" and "A is true" implies that B must hold via Modus Ponens. This is inconsistent with the supposition not B and, therefore, on Braine and O'Brien's theory, leads to the conclusion "Nothing follows." Indeed participants were split, at least with the first two scenarios (see Table 4).

Experiment 3 also asked participants

QD: If B were false, would D still be true?

The theory states that if D is true on the supposition that B is false, then the conditional proof schema implies the truth of "If not B, then D." One of the premises states that D is indeed true, a proposition unaffected by the falsity of B. Therefore, the theory predicts an answer of "yes" to QD. The data are supportive inasmuch as the proportions of participants giving this response was greater for QD than for QA. However, the proportions are small (see Table 1). For the corresponding question in Experiment 4, the model also predicts a "yes" response: E should be judged true because E was stated to be true. Moreover, because A is true, a path of transitive modus ponens inferences leads from A to E. This prediction fails for two of three scenarios.

For Experiments 5 and 6, we can derive predictions as in Experiment 1 except that questions did not ask whether events *could* happen, but whether they *would* happen. As a result, the derivation of the contrapositive conclusion applies directly, and the theory wrongly predicts that participants would respond "no," Component A would not still operate. The theory correctly predicts the "no" response for Component B.

In sum, the mental logic theory proposed by Braine and O'Brien (1991) is not wholly consistent with the data from our conditional conditions. We know of no theory of mental logic that makes predictions in our causal conditions.

8.2. Mental model theory

Byrne's (2002) mental model theory of counterfactual thinking has no clear application to these experiments. Goldvarg and Johnson-Laird (2001) proposed that the statement, "A causes B" refers to the same set of possibilities as "if A then B" along with a temporal constraint (B does not precede A). They represent the set of possibilities as a list of mental models:

A B
not A B
not A not B

Because it equates the set of possibilities associated with causal and conditional relations, this proposal predicts identical responses for the two and is therefore unable to explain the differences we observed between causal and logical conditional scenarios. Moreover, because it does not allow the possibility "A not B," it is inconsistent with the undoing effect with causal premises.

Because mental model theory assumes only a single kind of conditional, we derive a prediction from mental model theory for Experiment 1, making the assumption that the question "Imagine that Ball 2 could not move, would Ball 1 still move?" can be translated to "If Ball 2

did not move, did Ball 1 still move?" To answer, the following fully explicit set of models is required:

Ball 1 moves	Ball 2 moves
not Ball 1 moves	Ball 2 moves
not Ball 1 moves	not Ball 2 moves

The only model in which Ball 2 is not moving is the last one, and Ball 1 is not moving in it, so the model predicts the response "no, Ball 1 cannot move." This is not what was observed, especially in the causal conditions. A similar analysis applies to Experiments 5 and 6.

The problem faced by mental model theory is that it is incomplete in the constraints it imposes on how possibilities are generated. One constraint follows from the undoing effect: Counterfactual interventions on variables do not determine the value of the causes of those variables. Because mental model theory is extensional and therefore has no means to represent causal structure, it has no way to represent this constraint.

To answer the second question, the theory assumes that people appeal to the models

Ball 2 moves	Ball 3 moves
not Ball 2 moves	Ball 3 moves
not Ball 2 moves	not Ball 3 moves

In the models in which Ball 2 does not move, Ball 3 might or might not be moving. This implies that people should not know the answer. But few people responded that way.

In Experiment 2, participants were told that A happened, so the mental models associated with the premises should exclude all those that do not include A, and the theory therefore predicts that people should respond "Nothing follows." But in fact, in the interventional condition, people inferred that the probability of A was high.

Mental model theory has no mechanism for distinguishing subjunctives such as "if B were not to occur" from simple declaratives such as "if B does not occur." Therefore, the best this implementation of mental model theory can do to make a prediction on our causal problems in Experiment 3 is to treat the premises as material conditionals. Phil Johnson-Laird (Johnson-Laird, personal communication, October, 2000) was kind enough to run his mental model program on our premises. The program concluded, like our participants, that D occurred (because the premises state that D occurred and no other statement implies otherwise). However, unlike our participants, the program also concluded that A did not occur (because A is sufficient for B and B did not occur).

The premises in Experiment 4 entail more models than those of Experiment 3, and this should result in more errors. Otherwise, the theory would predict that E occurred for the same reasons that it predicted D in Experiment 3. However, it is not clear how to handle the question about B. The model would be faced with an inconsistency resulting from, on one hand, the premises "If A then B" and "A is true," whose models should combine to produce B. On the other hand, the assumption in the question that D is false, combined with "If B then D," should lead to the conclusion not B. According to mental model theory, the contradictory conclusions B and not B would lead to the prediction that nothing follows. However, the backward inference to not B is less likely according to the theory than the forward inference to B. Therefore, the

theory might predict more "yes" than "no" responses. The one strong prediction is that the proportions of "yes" responses should be somewhere between 50% and 100% and should be consistent across all conditions. Although it is true that the proportions are all close to the predicted range (45% to 90%), they are not consistent.

Goldvarg and Johnson-Laird (2001) allowed that the set of mental possibilities can vary with enabling and disabling conditions. Could this explain any of our data? Consider the simplest case where A causes B, and the question, "Would A still occur if B were prevented from occurring?" The statement that B is prevented presupposes some preventative cause X (e.g., I switch B off). Given X, and the knowledge that X causes not B through prevention, people might allow A. That is, they might add to the list of possibilities the mental model

A X not B

which licenses the inference from not B to A.

The problem with this move is that the mental model intended to explain causal knowledge itself requires causal knowledge to be constructed. The variable X must be invented after learning that B has been prevented. It could not exist a priori because that would lead to a combinatorial explosion of models; one would need to represent the possibility that Y enables B even in the presence of disabling condition X, the possibility that X' prevents X, that X'' prevents X', and so forth. But if we are reconstructing the possible models after positing X, why is the previously mentioned model the only possibility? Without prior causal knowledge, another possibility might be a possibility that is not consistent with the undoing effect:

not A X not B

In sum, mental model reconstruction depends on prior causal models because those are the source of the relevant constraints. Contrary to Goldvarg and Johnson-Laird (2001), causal predicates cannot be reduced to a list of possible states, but require knowledge about causal mechanism. Pearl (2000) made an analogous argument against Lewis's (1986) counterfactual analysis of causation. Lewis defined *causation* in terms of counterfactuals, whereas Pearl argued that it is the causal models that ground (causal) counterfactuals.

8.3. Conditional probability account: Noninterventional Bayesian probability

Perhaps the causal relations in all of these experiments are not understood deterministically, but probabilistically. On this view, one might ask whether the results could be fit, using a more common probabilistic model that conditions on actual events and does not use the *do* operator, a framework we will refer to as *noninterventional Bayesian probability.* The simple answer is that our questions cannot even be asked in such a framework because it provides no way to represent counterfactuals, such as "Imagine that Ball 2 could not move." It cannot distinguish it from an observational statement, such as "if Ball 2 did not move."

The closest we can come in this framework to a model of Question 1 in Experiment 1 is P(Ball 1 moves|Ball 2 does not move). On all reasonable parameter choices, this will be low. However, there are degenerate parameter sets that will make it high. If P(Ball 1 moves) is set sufficiently high, then it will be high whether Ball 2 moves or not. But if it is set to 1, the model

fails entirely because P(Ball 1 moves|Ball 2 does not move) = 0. Consider extreme probabilities between 0 and 1, say P(Ball 2 moves|Ball 1 moves) = .9 and P(Ball 2 moves|Ball 1 does not move) = .1. We plot the value of P(Ball 1 moves|Ball 2 does not move) given these parameters for a variety of values of P(Ball 1 moves) in Fig. 1. The vast majority of parameter values lead to a low probability that Ball 1 can move. P(Ball 1 moves) must be greater than .9 before the relevant conditional probability is greater than .5.

This model has more success with Question 2, "Imagine that Ball 2 could not move, would Ball 3 still move?" because the question does not require distinguishing counterfactuals from observations; both give the same answer. When Ball 2 causes Ball 3's movement, it is reasonable to assume that the probability that Ball 3 can move, given that Ball 2 cannot, is low, and this is what people said in the causal conditions. The model makes no distinct claims in the logical conditional condition.

The same analysis applies to Experiment 2 and a similar one to Experiments 5 and 6. Consider a noninterventional analysis of the question from Experiment 3, "If B had not occurred, would A have occurred?" If we neglect the counterfactual nature of the question and assume it is asking for the probability of A, given that B did not occur, and given D (stated in an earlier premise), the closest expression in this framework to the question is P(A|D,~B).[2,3] Complete specification of a Bayesian model of binary variables with the requisite, diamond-shaped structure requires nine parameters: P(A), P(B|A), P(C|A), P(B|~A), P(C|~A), P(D|B,C), P(D|~B,C), P(D|B,~C), and P(D|~B,~C). The great majority of people said "yes" in the causal

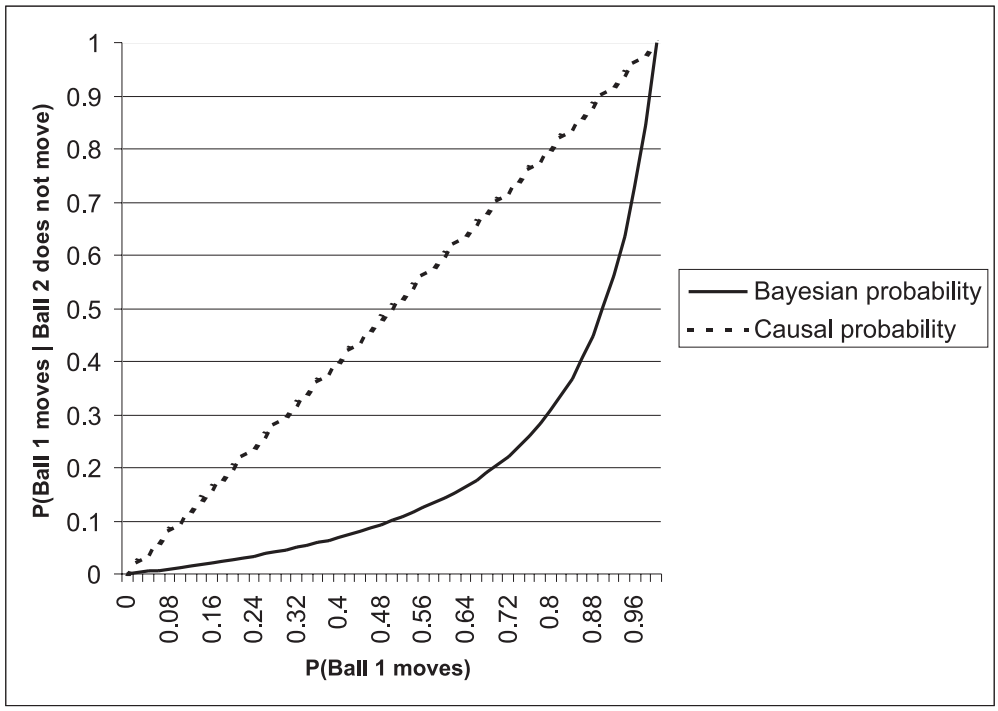


Fig. 1. Predicted response to question "If Ball 2 could not move, would Ball 1 still move?" in Experiment 1 by noninterventional Bayesian and causal probability models.

condition (see Table 3). Fig. 2 shows the value of P(A|D,~B) for four choices of the parameters. In every case, the prior probability of the root node A is set to .5, and D is assumed to be a noisy-OR gate where B or C cause D independently with no other causes (P(D|~B,~C) = 0. The choice labeled *fully deterministic* is the extreme case in which all conditional probabilities are either 0 or 1. The model fails to make a prediction in this case because the conditional probability is undefined (indicated by a "?" in Fig. 2). Adding a little noise to the probabilities by making them close to 0 and 1 ("extreme probabilities") allows the model to make a prediction, but not the right one (.5). Making the probabilities less extreme ("nonextreme probabilities") does not help; the prediction is still equivocal (.5). The model makes a correct prediction if it is assumed to be semideterministic (an event's occurrence makes its effect possible, but an event's nonoccurrence guarantees the absence of its effect). In that case, the predicted conditional probability is 1. The model also makes a correct (high) prediction if all conditional probability parameters are low (this occurs because a very high probability is required to explain why E occurred if all events are unlikely). The model again fails if parameters are high (.30).

Consider the other question, "If B had not occurred, would D still have occurred?" which 80% of participants answered "yes." The framework cannot distinguish "D has occurred" from "D *still* has occurred." Lacking a veridical representation, one could suspend judgment about D and ask for P(D|~B). The values of this conditional probability are presented in Fig. 2 for the same parameter values as the previous model. The figure makes apparent that the model makes the wrong (low) prediction for all parameter sets. The model does make the correct (high) prediction if all parameters are set high. But as we have already seen, doing so would make the model of the other question, P(A|D,~B), inconsistent with the data.

A different nonveridical representation would assume participants incorporate the premises, including that D occurred, and interpret the question as "B is false. Is D true?" This sug-

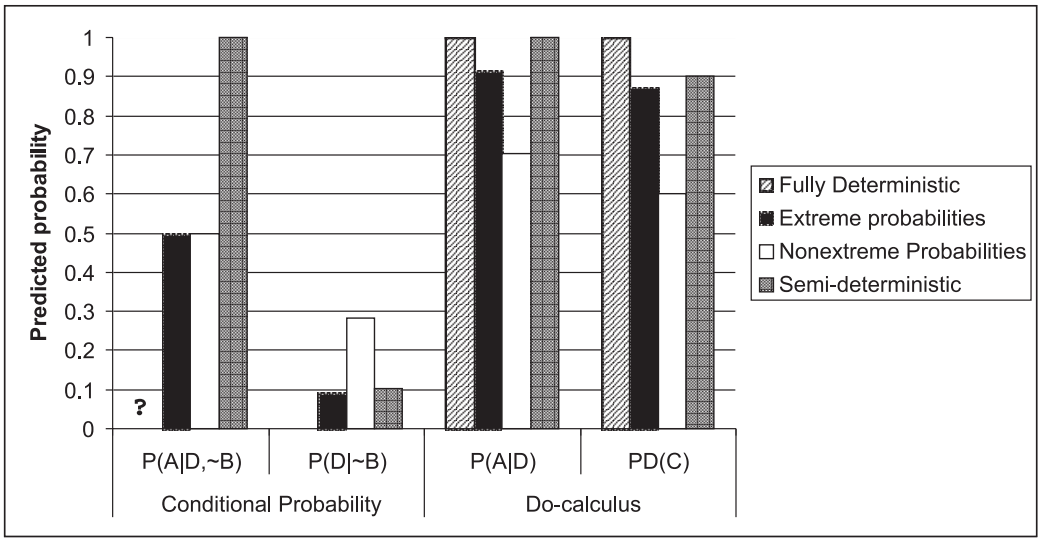


Fig. 2. Noninterventional Bayesian conditional probability and *do*-calculus interventional probability models of two questions from Experiment 3 for 4 parameter sets. Parameter values are shown below figure. "?" indicates the model is unable to generate a prediction.

gests the model P(D|D,~B) = 1. But such a representation ignores both the semantics of the conditional question and the subjunctive that makes it counterfactual. On this analysis, one could never consider a counterfactual that denies an event that occurred, such as "Would I have passed the exam if I had studied?"—a question which makes most sense on the assumption that I failed.

In sum a noninterventional Bayesian account of Experiment 3 is not only incoherent, but unlikely. An analogous argument reveals the limitations of such an analysis for Experiment 4. We omit the details here due to space considerations.

8.4. Causal probability account: Causal Bayesian analysis

The causal modeling framework provides a natural and distinct representation of Experiment 1's Question 1, "If Ball 2 could not move, would Ball 1 still move?" namely,

P(Ball 1 moves|*do*(Ball 2 does not move)) = P(Ball 1 moves)

because Ball 1 and Ball 2's movements are independent under the *do* operation. Values of this interventional probability are plotted in Fig. 1. It predicts a greater likelihood of claiming that Ball 1 can still move than the noninterventional probabilistic model for every possible assumption about P(Ball 1 moves), except the degenerate cases in which Ball 1 always or never moves. The model makes the same correct prediction for the second question that the noninterventional model makes. This theory also makes no specific claims about the logical conditional condition. Altogether, the theory offers the simplest and most comprehensive account of the data in the causal conditions. Similar analyses apply to Experiments 2, 5, and 6.

For Experiment 3, an alternative probabilistic model that generalizes to the deterministic case can be derived using Pearl's (2000) three-step theory of counterfactual inference: (i) Abduction: We update the probability of the root node A given the fact D. D is diagnostic of one or both of B and C, which are in turn diagnostic of A. So the probability of A is raised. (ii) Action: The link from A to B is removed to represent *do*(~B). (iii) Prediction: To find out if A would have occurred if B had not, we note that the counterfactual assumption about B has no effect on A because the link between B and A has been removed. Therefore, the probability of A remains what it was after Step (i), P(A|D). The predictions of this representation are presented in Fig. 2 for the same parameter sets as previously shown. Notice that this model predicts a high value for all four parameter sets shown. The only parameter set where the model does not make the correct prediction (not shown) is when the conditional probability parameters are all set high. The model fails then because the prior probability of A is not high, P(A) = .5, and the occurrence of D is not diagnostic because D has high probability whether or not the other events occur. Therefore D does not change the probability of A much, P(A|D) = .53. But participants are not likely to assume that all variables are probable regardless of their causes. If they did, they would also assume that A was likely, which would again result in a correct prediction.

The question, "If B had not occurred, would D still have occurred?" can be modeled by assuming a logic parallel to the deterministic case. Steps (i) and (ii) are the same as for the previous question. For Step (iii), note that D would occur only if C occurs because it is an OR gate, and we have set B to *not occur*. C will occur with some probability if A occurs and with a lower

probability if A does not occur, P(C|A)P(A|D) + P(C|~A)P(~A|D). This is the probability of C in the distribution in which A has been updated by knowledge of D (denoted $P_D(C)$). Again, this model is consistent with the data for all parameter sets. The only case we have observed in which this model makes the wrong prediction is when all conditional probability parameters are set low because then D is never expected to occur.

For Experiment 4, the causal models we investigated—parallel to those of Experiment 3—are P(B|A,E) and $P_{A,E}(D)$, where $P_{A,E}$ represents the probability distribution in which B has been updated by knowledge of A and E. We omit the details. In sum, the observed data fall out of the causal Bayesian framework without additional assumptions for almost all parameter sets.

8.5. Implications

The implications of the logic of intervention are deep and wide-ranging. The most fundamental perhaps is the precondition it places on the application of Bayes' rule and its logical correlates for updating belief. Beliefs should only be updated after taking into account the status of the evidence (interventional versus observational) and making the necessary changes to the causal model. Bayes' rule is by far the most prevalent tool for adjusting belief in a hypothesis based on new evidence. A situation frequently modeled using Bayes' rule instantiates the hypothesis as a cause and the evidence as an effect. For example, the hypotheses might be the possible causes of a plane crash, and the evidence might be the effects of the crash found on the ground. The evidence is used to make diagnostic inferences about the causes. This is fine when the evidence is observed, but not if any manipulation by an external agent has occurred. The probability of a cause given a manipulated effect (i.e., given a *do* operation on the effect) cannot be determined using simple Bayesian inversion from the probabilities of the effect, given its causes. Intervention is hardly a rare or special case. Manipulation is an important tool for learning; it is exactly what is required to run the microexperiments necessary to learn about the causal relations that structure the world. Whenever we use this learning tool, as a baby does when manipulating objects, Bayes' rule will fail as a model of learning, just as it failed as a model of inference across our five experiments, unless probabilities are conditioned on interventions (as done by, e.g., Heckerman, 1995; Spirtes et al., 1993).

The *do* operator also clearly distinguishes representations of logical validity from representations of causality. This is seen most directly by comparing the modus tollens structure (If A then B, not B, therefore not A) to its corresponding causal *do* structure (A causes B, B is prevented, therefore A's truth is unaffected). People may on occasion fail to draw valid modus tollens inferences as a result of interpreting a logical argument as causal and "not B" as *do*(B = did not occur).

If this possibility is correct, it supports our supposition that the interpretation of conditionals varies with the theme of the text that the statements are embedded in (a conclusion already well documented, e.g., Almor & Sloman, 1996; Braine & O'Brien, 1991; Cheng & Holyoak, 1985; Edgington, 1995; Johnson-Laird & Byrne, 2002). Conditionals embedded in deontic contexts are known to be reasoned about in a way consis-

tent with deontic logic (Manktelow & Over, 1990). Causal conditionals must also be distinguished from definitions. Consider the conditional

If John is a Richman, he will have had $10 million at some point in his life.

This can either be stated in a context that makes it causal,

John is a Richman. The Richmen is a group of successful people who get elected based on merit and then get rewarded. All of their members are given $10 million.

or in a context that makes it definitional:

John is a Richman. This is a name given to all of the people who have had $10 million at some point in their life.

In the causal context, we would expect to observe the undoing effect, in the definitional case we would not. This is just what we have found. When we asked people

Imagine John's wife had prevented him from ever getting $10 million, would he have still been a Richman?

One hundred percent of people given the causal context said "yes," whereas only 30% of those given the definitional context did. Models of mental logic and mental model theory fail to explain our results in part because they fail to make these distinctions.

Our studies also found that people consistently expressed more confidence when answering causal over conditional questions. This supports our assertion that causal problems are more natural and that conditional ones lend themselves to more variable construal.

Our data support the psychological reality of a central tenet of the causal modeling framework. The principle is so central because it serves to distinguish causal relations from other relations, such as mere probabilistic ones. The presence of a formal operator that enforces undoing, Pearl's (2000) *do* operator, makes it possible to construct representations, affording valid causal induction, that support manipulation and control, and inference about the effect of manipulation, be it from actual physical intervention or counterfactual thought about intervention. The *do* operation is precisely what is required to distinguish representations of causality from representations of probability, possibility, and truth.

Overall, the findings provide qualitative support for the causal modeling framework (cf. Glymour, 2001). The causal modeling analysis starts with the assumption that people construe the world as a set of autonomous causal mechanisms and that thought and action follow from that construal. The problems of prediction, control, and understanding can therefore be reduced to the problems of learning and inference in a network that represents causal mechanisms veridically. Once a veridical representation of causal mechanisms has been established, inference can take place by intervening on the representation rather than on the world itself. But this cannot be achieved without a suitable representation of intervention. The *do* operator is intended to allow such a representation, and the studies reported herein provide some evidence that people are able to use it correctly to reason.

Representing intervention is not always as easy as forcing a variable to some value and cutting the variable off from its causes. Indeed, most of the data reported here show some variabil-

ity in people's responses. People are not generally satisfied to simply implement a *do* operation. People often want to know precisely how an intervention is taking place. A surgeon cannot simply tell me that he is going to replace my knee. I want to know how, what it is going to be replaced with, and so forth. After all, knowing the details is the only way for me to know with any precision how to intervene on my representation, which variables to *do,* and thus what can be safely learned and inferred.

Causal reasoning is not the only mode of inference. People have a variety of frames available to apply to different problems (Cheng & Holyoak, 1985). Mental models serve particularly well in some domains such as syllogistic reasoning (Bara & Johnson-Laird, 1984), and sometimes reasoning is associative (see Sloman, 1996). The presence of a calculus for causal inference, however, provides a means to think about how people learn and reason about the interactions among events over time. They do so by temporarily considering other possible worlds that they construct by disconnecting events from their normal causes in the actual world.

Notes

1. We thank Reviewer no. 3 for this example.
2. We use a capital letter to stand for an event that occurred, we add "~" if the event did not occur, and we use a corresponding boldface letter for the variable ranging over these two values.
3. P(A|D, ~B) = P(A & D & ~B)/P(D & ~B) by the definition of conditional probability. If we treat the graph as a Bayesian network, the joint probability P(**A**,**B**,**C**,**D**) = P(**A**)•P(**B**|**A**)•P(**C**|**A**)•P(**D**|**B**,**C**). Both numerator and denominator of the conditional probability can then be calculated by summing terms of the joint distribution.

Acknowledgments

This work was funded by NASA grant NCC2-1217. Some of the results for 2 of the 3 scenarios in Experiments 3 and 4 and Experiment 5 were reported at the 2002 Cognitive Science Society conference. We thank Phil Johnson-Laird, Clark Glymour, Denis Hilton, Jean-François Bonnefon, David Over, an anonymous reviewer, and especially Josh Tenenbaum for valuable discussions of this work and comments on prior drafts. Daniel Mochon and Constantinos Hadjichristiditis made many contributions to the research, and Brad Love, Ian Lyons, Peter Desrochers, Henry Parkin, Heloise Joly, and Clare Walsh helped to collect data.

References

Almor, A., & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review, 103,* 374–380.
Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bacon, F. (1620). *Novum organum,* Chicago: Open Court.
Bara, B. G., & Johnson-Laird, P. N. (1984). Syllogistic inference. *Cognition, 16,* 1–61.
Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: Lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98,*182–203.
Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences, 6,* 426–431.
Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416.
Edgington, D. (1995). On conditionals. *Mind, 104,* 235–329.
De Finetti, F. (1980). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokler (Eds.), *Studies in subjective probability* (2nd ed., pp. 53–118). Huntington, NY: Krieger.
Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.
Goldvarg, E., & Johnson-Laird, P. N. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25,* 565–610.
Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32.
Heckerman, D. (1995). A tutorial on learning with Bayesian networks. *Microsoft Technical Report MSR-TR-95-06.* Redmond, WA: Microsoft Research.
Johnson-Laird, P. N., & Byrne, R. M. J. (1991) *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Johnson-Laird, P. N., & Byrne, R. M. J. (2002) Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109,* 646–678.
Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30,* 856–876.
Lewis, D. (1986). *Philosophical papers* (Vol. 2). Oxford, England: Oxford University Press.
Manktelow, K. I., & Over, D. E. (1990). Deontic thought and the selection task. In K. J. Gilhooly, M. Keane, R. H. Logie, & G. Erdos (Eds), *Lines of thinking* (Vol. 1, pp. 153–164). Chichester, England: Wiley.
Meek, C., & Glymour, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science, 45,* 1001–1021
Nozick, R. (1995). *The nature of rationality.* Princeton, NJ: Princeton University Press.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Francisco: Morgan Kaufmann.
Pearl, J. (2000). *Causality.* Cambridge, England: Cambridge University Press.
Reichenbach, H. (1956). *The direction of time.* Berkeley: University of California Press.
Rips, L. J., & Marcus, S. L. (1977). Suppositions and the analysis of conditional sentences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Savage, L. J. (1954). The foundations of statistics. New York: Wiley.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.
Sloman, S. A., & Lagnado, D. (2004). Casual invariance in reasoning and learning. In B. Ross (Ed.), The psychology of learning and motivation, Vol. 44 (pp. 287–325). San Diego, CA: Academic Press.
Sobel, D. (2003). *Watch it, do it, or watch it done.* Manuscript submitted for publication.
Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer-Verlag.
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489.
Von Mises, R. (1957). *Probability, statistics, and truth.* London: Allen & Unwin.
Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, England: Oxford University Press.

Appendix A. French Language Materials Used in Experiment 1

Causal scenario

Il y a trois boules de billard sur une table, qui se comportent de la manière suivante:

Le déplacement de la boule 1 cause le déplacement de la boule 2. Le déplacement de la boule 2 cause le déplacement de la boule 3.

Causal conditional scenario

Il y a trois boules de billard sur une table, qui se comporte de la manière suivante:

Si la boule 1 se déplace, alors la boule 2 se déplace. Si la boule 2 se déplace, alors la boule 3 se déplace.

Logical conditional scenario

Une personne fait la preuve de ses capacités logiques. Elle fait se déplacer les boules sans violer les règles suivantes: Si la boule 1 se déplace, alors la boule 2 se déplace. Si la boule 2 se déplace, alors la boule 3 se déplace.

Questions (same in all conditions)

(1) Imaginez que la boule 2 ne puisse pas bouger. Est-ce-que la boule 1 bougerait quand même?
Encerclez une des trois options:

Elle le pourrait Elle ne le pourrait pas Je ne sais pas

(2) Imaginez que la boule 2 ne puisse pas bouger. Est-ce-que la boule 3 bougerait quand même?
Encerclez une des trois options:

Elle le pourrait Elle ne le pourrait pas Je ne sais pas