Submitted for publication.

# Do We "do"?

Steven A. Sloman

David A. Lagnado

Brown University

Please address correspondence to:

Steven Sloman

Cognitive and Linguistic Sciences
Brown University, Box 1978
Providence, RI 02912
Email: Steven_Sloman@brown.edu.
Phone: 401-863-7595
Fax: 401-863-2255

Sept. '02 - July '03:
Laboratoire de Psychologie Cognitive
Université de Provence
29, avenue Robert Schuman
F-13621 Aix-en-Provence Cedex 1
France

Running head: Undoing effect in causal reasoning

**Abstract**

A normative framework for modeling causal and counterfactual reasoning has been proposed (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). The framework is general, covering both probabilistic and deterministic reasoning, and is built on the premise that reasoning from observation differs fundamentally from reasoning from intervention. Intervention includes actual (e.g., physical) manipulation as well as counterfactual thought (e.g., imagination). The key representational element that affords the distinction is what Pearl calls the *do* operator. The *do* operation represents intervention and has the effect of simplifying a causal model. Construing the *do* operator as a psychological function affords predictions about how people reason when asked counterfactual questions about causal relations. Seven studies are reported that test these predictions for both deterministic and probabilistic causal and conditional (logical) arguments. The results support the proposed representation of causal arguments, especially when the nature of the counterfactual intervention is made explicit. The results also show that conditional relations are construed variously and are highly sensitive to pragmatic context.

Human reasoning is sometimes said to have two principal modes, deductive and inductive. In a sense, these modes have complementary characterizations. Deductive reasoning is easy in principle, difficult in practice; inductive reasoning is difficult in principle, easy in practice. Of course, deductive reasoning faces many obstacles including combinatorial explosion, expressive limitation, and impossibility theorems. Nevertheless, the problem of deciding the validity of a deductive argument is well defined and a variety of automated theorem-proving systems exist. Yet people stumble even with some theoretically simple arguments. For instance, many people fail to determine the validity of arguments of the modus tollens form (see, e.g., Evans, 1982, for a review):

If A then B.
Not B.
Therefore, not A.

In contrast, a prevalent belief is that inductive argument strength cannot be reduced to any kind of formal logic (Hume, 1748; Goodman, 1954) and yet people often come quickly and easily to inductive conclusions that are widely accepted. For example, even very young children would be surprised if the sun didn't rise one morning.

Many authors attribute the human facility with inductive inference to the power of causal reasoning: Our ability to wisely project predicates from one category to another on inductive grounds alone depends on our ability to select the causal relations that support the inference and reason appropriately about them. For example, from the observation that one motorcycle accelerates quickly, one can conclude with some confidence that another motorcycle of the same brand and size will accelerate quickly based on (more or less vague) causal knowledge of motorcycle engines and manufacturing.

Indeed, causal analysis is pervasive. In the law, issues of negligence concern who caused an outcome and the determination of guilt in many countries requires evidence of a causal chain from the accused's intention through their action to the crime at hand. Evidence that might increase the probability of guilt (e.g., an accused's race) is not permitted in court if it doesn't support a causal analysis of the crime. Some legal scholars (Lipton, 1992) claim that legal analyses of causality are in no sense special, that causation in the law derives from everyday thinking about causality. Causal analysis is equally pervasive in science, engineering, politics, indeed in every domain that involves explanation, prediction, and control.

The appeal to causal analysis certainly does not solve all the problems of induction. In fact, Hume (1748) argued that causal induction itself cannot be logically justified. Moreover, causal analysis can be difficult because it depends not only on what happened, but also on what *might* have happened (Mackie, 1974). The claim that an event A caused another event B implies that if A had not occurred, then B would not have occurred (unless of course some other sufficient cause of B also occurred). Likewise, the fact that B would not have occurred if A had not suggests that A is a cause of B.

But the appeal to causal analysis does solve a part of the problem of induction. This is because causal inductions can be made with confidence using a method familiar to all experimental scientists: manipulation of independent variables. Through manipulation, one controls an independent variable, holding other relevant conditions constant, such that changes in its value will determine the value of a dependent variable. This supports an inference about whether the independent variable is a cause of the dependent one: It is if the dependent variable changes after intervention, it isn't if the dependent variable doesn't change. Through manipulation one sets up states to be directly compared, like an

experimental and a control condition, in perfect analogy to the comparison between actual and counterfactual worlds implied by a causal statement. This dependence of causal relations on counterfactuals lies at the heart of a fundamental law of experimental science: Mere observation can only reveal a correlation, not a causal relation. And everyday causal induction has an identical logic; people often must intervene on the world rather than just observing it to draw a causal induction.

If we already have some causal knowledge, then certain causal questions can be answered without actual intervention. Some can be answered through mental intervention; by imagining a counterfactual situation in which a variable is manipulated and determining the effects of that change. People attempt this, for example, whenever they wonder "if only..." (if only I hadn't made that stupid comment... If only my data were different...).

Recent analytic work by Spirtes, Glymour, and Scheines (1993) and by Pearl (2000) presents an even rosier picture. Not only can causal relations be learned through intervention, in some situations merely correlational data suffice. Pearl presents a normative theoretical framework for causal reasoning about both actual and counterfactual events. Central to this framework is the use of directed acyclic graphs to represent both actual and counterfactual causal knowledge. Interpreted as a psychological model, the framework makes predictions about how people reason when asked counterfactual questions about causal relations. The most basic representational distinction in the causal modeling framework is that between observation and action.

**Observation versus Action (Seeing versus Doing)**

**Seeing**. In general, observation can be represented using the tools of conventional probability. The probability of observing an event (say, that a logic gate is working

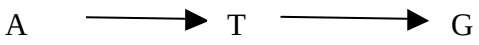
properly) under some circumstance (e.g., the temperature is low) can be represented as the conditional probability that a random variable G, representing the logic gate, is at some level of operation g when temperature T is observed to take some value t:

$$\Pr\{G = g|T = t\} \text{ defined as } \frac{\Pr\{G = g \,\&\, T = t\}}{\Pr\{T = t\}}.$$

Conditional probabilities are symmetric in the sense that, if well-defined, their converses are well-defined too. In fact, given the marginal probabilities of the relevant variables, Bayes' rule tells us how to evaluate the converse:

$$\Pr\{T = t|G = g\} = \Pr\{G = g \mid T = t\}\frac{\Pr\{T = t\}}{\Pr\{G = g\}}. \quad (1)$$

**Doing.** To represent action, Pearl (2000) proposes an operator *do*(•) that controls both the value of a variable that is manipulated as well as the graph that represents causal dependencies. *do*(X=x) has the effect of setting the variable X to the value x and also 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). For example, imagine that you believe that temperature T causally influences the operation of logic gate G, and that altitude A causally influences T. This could be represented in the following causal diagram:

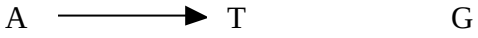



Presumably, changing the operation of the logic gate would not affect temperature (i.e., there's no causal link from G to T). We can decide if this is true by acting on the logic gate to change it to some operational state g and then measure the temperature; i.e., by running an experiment in which the operation of the logic gate is manipulated. We could

not in general determine a causal relation by just observing temperatures under different logic gate conditions, because observation provides merely correlational information. 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):

$$\Pr\{T = t|do(G = g)\}$$

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




The rationale for this is that if I have set G=g, then my intervention renders other potential causes of g irrelevant. I am overriding their effects, so I should not make any inferences about them. Now I can examine the probability distribution of T in the causal graph. But in doing so, I should not take into account the prior probability of g, because I have set its value, making its value certain by virtue of my action. In the causal modeling framework, the absence of a path from one variable to another represents probabilistic independence between each value of those variables. Because the *do* operation removes the link between T and G in the graph, they are rendered probabilistically independent. The result is that:

$$\Pr\{T = t|do(G = g)\} = \Pr\{T = t\}.$$

The *do* operator is used to represent experimental manipulations. It provides a means to talk about causal inference through action. It can also be used to represent *mental* manipulations. It provides a means to make counterfactual inferences by determining the representation of the causal relations relevant to inference if a variable had been set to some counterfactual value.

In the next section of this paper, we report several experiments intended to test whether people are sensitive to the logic of the *do* operator; in particular, whether people disconnect an intervened-on variable from its (normal) causes. In other words, we test the prediction of the do operator that variables manipulated actually or counterfactually should not be treated as diagnostic of their causes. All experiments present participants with a set of premises and then ask them to judge the validity of a particular conclusion based on a supposition. We compare suppositions about observed events to various types of counterfactual suppositions. The causal modeling framework applies to both deterministic and probabilistic causal relations. The first six experiments involve deterministic relations, the final experiment generalizes the conclusions to arguments with probabilistic relations.

## Experiment 1

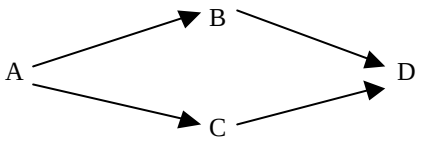
Consider the following set of causal premises in which A, B, C, and D are the only relevant events:

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

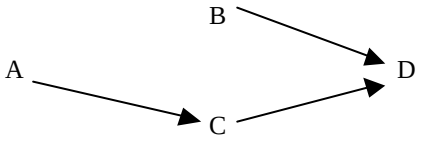
On the basis of these facts, answer the following 2 questions:

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)

Pearl (2000) gives the following analysis of such a system. First, we can graph the causal relations amongst the variables as follows:




You are told 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 4 events have occurred. Thus far the rules of ordinary logic are sufficient to update our model. When asked what would have happened if B had not occurred, however, we should apply the *do* operator, *do*(B = did not occur), with the effect of severing the links to B from its causes:




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

Other theories of propositional reasoning, mental models theory (Johnson-Laird & Byrne, 1991) and any theory based on logic (e.g., Rips, 1994), don't really make predictions in this context because the premises use causal relations and therefore lie outside the propositional domain. The closest they come is to posit that causal relations are interpreted as material conditionals (an assumption made by Goldvarg & Johnson-Laird, 2001). To see if such an interpretation of the causal premises above is valid, we can consider the following conditional premise set:

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 particular prediction about such premises except to say 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. The predictions made by a "material conditional" account will depend on assumptions about how people interpret the questions; i.e., how they modify the original set of premises. To answer question i. people may suppress the statement that D is true, and add the statement that B is false. If they do, the truth of D is indeterminate, because it is not entailed by the falsity of B. Alternatively, people might not suppress D. The answer would then be "yes" because the original premises state that D is true. Such an account yields a less ambiguous answer to question ii. Once people suppose that B is false, they are licensed to infer, by modus tollens, that A is false.

If this "material conditional" construal is extended to the causal premises, it should make identical predictions for corresponding causal premises. In particular, people should respond “no” to the second question, in contrast to the causal modeling prediction which directly contradicts the modus tollens form. The causal modeling framework

states that B's non-occurrence does not imply A's non-occurrence whereas modus tollens requires that, whenever if A then B holds, not B implies not A.

## Method

**Materials**. Three scenarios were used in this experiment, 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 travelling 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 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 amongst 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?

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?

**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. 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." 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). They worked at their own pace and were given as much time as they desired to answer the questions.

## Results and Discussion

Percentages of participants responding "yes" to each question are shown in Table 1. A very different pattern can be observed for the Causal and Conditional statements. The

causal modeling framework correctly predicted the responses to the causal premises, the vast majority of responses were “yes.” The responses to the conditional premises much more variable. For each question in each scenario, the proportion of “yes” responses was significantly higher in the Causal than the Conditional condition (all p’s < .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 the conditional version as causal relations, especially for this problem. The clear physical causality of the robot problem lends itself to causal interpretation.

The predominance of "yes" responses in the causal condition implies that for the majority of participants the supposition that B didn't occur did not influence their beliefs about whether A or D occurred. This is consistent with the idea that these participants mentally severed (undid) 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.

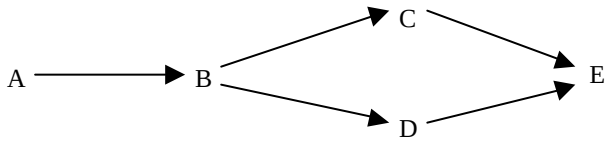
These conclusions are supported by the confidence judgments. Participants were highly confident when answering causal questions (mean of 6.0 on the 1-7 scale). They were appreciably less confident when answering conditional questions (mean of 5.4), $t(236) = 4.77$; s.e. = .13; $p < .0001$.

## Experiment 2

One might argue that the difference between the causal and conditional conditions in Experiment 1 is not a greater tendency to counterfactually decouple variables from their causes in the causal over the conditional context, but instead different pragmatic

implicatures of the two contexts. In particular, the causal context might presuppose the occurrence of A more than the conditional context presupposes the truth of A. Thus, it is (perhaps) more plausible that D would be true in the conditional premise sets even if A were false than that D would have occurred in the causal premises even if A had not. 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 different likelihoods of mentally performing the undoing operation. 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 effect, 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?

The causal modeling framework predicts that a counterfactual assumption about D should disconnect it from B in the causal context so that participants should answer "yes"

to both questions. A parallel conditional version was also used. Participants should only answer "yes" in the conditional context if they interpret the problem causally. Once again the predictions of a material conditional account depend on assumptions about how the questions modify the premises. A plausible assumption is that only statements mentioned in the question are suppressed. Thus in answering question ii., belief about the truth of D and B might be suspended and not-D supposed. However, this leads to a conflict because not-D implies not-B (via modus tollens) but the premises state A and thus imply B (via modus ponens). It is thus unclear whether or not they should infer B. In any case, a material conditional account must predict no difference between the causal and conditional contexts.

**Method**

Twenty Brown University undergraduates received either the causal or conditional versions of the Abstract, Robot, and Politics problems described above, 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 1, 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 was identical to that of Experiment 1.

**Results and Discussion**

The results, shown in Table 2, are comparable to those of Experiment 1 although the proportion of "yes" responses was lower for one of the Robot scenario questions, whether the beam was shining if the solar panel had not been struck (only 55). This difference will be addressed in Experiments 3-7. Overall, the experiment provides further evidence of the undoing effect for causal relations. A difference between causal and conditional premises again obtained for Abstract and Political premises, $z = 2.20$; $p = .01$, and $z = 2.00$, $p = .02$, respectively, but not for Robot ones, $z = 1.18$; n.s. Moreover, 5 of 6 percentages were significantly greater than 50% in the Causal condition (all those greater than or equal to 70). Only 2 of 6 reached significance in the Conditional condition with values of 75 and 80. Both of these questions may well have induced a causal reading. Confidence judgments were again higher for answers to causal questions (mean of 5.89) than for answers to conditional questions (mean of 5.23), $t(38) = 2.30$; s.e. = .27; $p < .05$.

The replication of the undoing effect in this experiment suggests that the earlier results cannot be attributed entirely to different pragmatic implicatures from causal and conditional contexts. Any differences between Experiments 1 and 2, 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 said "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 didn't. Nothing in the

problem explicitly states that the nonoccurrence of D should not be treated as diagnostic of the nonoccurrence of B.

## Experiment 3

The causal modeling framework predicts that the connection between B and D should be mentally undone whenever D is explicitly prevented; when an intervention (mental or physical) outside the model clearly determines the value of D. To simulate such a situation, we repeated Experiment 2, but made the interventional prevention of D explicit. The prediction was that the undoing effect should prove more robust with explicit intervention.

### Method

Different groups of either 18 or 20 Brown University undergraduates saw the same sets of premises in both causal and conditional contexts as in Experiment 2, but were asked different questions, questions that made the external prevention of D explicit. For the abstract causal context, the questions were:

i. If somebody stepped in to prevent D from occurring, would E still have occurred?
ii. If somebody stepped in to prevent D from occurring, would B still have occurred?

For the abstract conditional context, the questions were:

i. If somebody stepped in and changed the value of D to false, would E still be true?
ii. If somebody stepped in and changed the value of D to false, would B still be true?

For the robot and political contexts, the causal and conditional questions were identical to one another, only the paragraphs describing the scenarios differed. The robot questions read:

i. If a lead barrier were placed in the path of Beam B to prevent it from striking the solar panel, would the robot have been activated?
ii. If a lead barrier were placed in the path of Beam B to prevent it from striking the solar panel, would the original (500 unit) beam have been shone through the prism?

The political questions read:

1) If the heavy toll of a natural disaster in Argentina prevents Argentina from declaring war, would Brazil have been aggressive?
2) If the heavy toll of a natural disaster in Argentina prevents Argentina from declaring war, would Brazil have declared war?

**Results and Discussion**

Results are shown in Table 3. The probability of saying "yes" was consistently high in the Causal condition and higher in the explicit prevention context than in its absence (Experiment 2), but not significantly higher, z's < 1 for all 3 scenarios. The differences may not be statistically significant because the probability of saying "yes" was already so high in the causal condition of Experiment 2. In any case, the great majority of participants acted as if explicitly preventing D caused it to have no diagnostic value for its cause (B) and that therefore other effects of the cause (namely E) still held. All percentages were significantly greater than 50% at $p < .001$ except for the Political question about E, $p = .09$. In other words, the effect of explicitly preventing D is well captured by the *do* operator.

An unexpected byproduct of explicit prevention was to increase the proportions of "yes" responses in even the conditional context. Both the Robot problem elicited a mean response of 83, significantly greater than 50% $p < .001$ and a mean of 67, marginally greater than 50%, $p = .07$. The Abstract problem also elicited one response marginally greater than 50%, $p = .07$. The Political problem elicited one response greater than 50%, p < 001. The increase with conditional premises probably occurred because the explicit prevention context made it more likely that the premises would be construed causally.

For example, a question beginning "If a lead barrier were placed in the path of Beam B to prevent it from striking the solar panel," may well have suggested to participants that they should construe the situation in terms of physical causation and reason about the situation using causal logic.

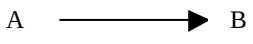
As usual, confidence judgments were higher for answers to causal questions (mean of 5.28) than for answers to conditional questions (mean of 5.10), however in this case the difference did not reach significance, $t < 1$.

## Experiment 4

Experiment 4 attempted to replicate the observations made thus far by showing the undoing effect as well as the enhancement of that effect in an explicit prevention context. Moreover, along with Experiments 5 and 6, it did so using an if-then statement in order to show that a conditional statement can be treated as causal in an appropriate context.

Experiment 4 was also intended to defeat an alternative interpretation of our results. It might be argued that in the causal conditions participants treated the relations in the premises as merely correlational and not causal. This would explain why they responded "yes" to our counterfactual questions: If variables do not cause other variables, then changing the value of one variable should have no effect on the values of other variables; in particular, the counterfactual assumption that one variable did not occur should not change participants' beliefs about the value of any other variable. Of course, this account falters in its failure to explain why participants would believe that events A in Experiment 1 and B in Experiments 2 and 3 occurred in the first place. In any case, the following experiments attempt to put this account fully to rest.

Consider the following scenario that assumes the simplest possible causal graph



and states the relation between A and B using an if-then construction:

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

In the non-explicit prevention condition, participants were shown these statements and then 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?

In the explicit prevention condition, they were 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 the undoing effect, that participants will say "yes" to question i., Component A will continue to operate even if B isn't because A should be disconnected from B by virtue of the counterfactual supposition about B. It also predicts the proportion will be higher in the explicit than non-explicit prevention conditions because the nature of the intervention causing B to be non-operative is less ambiguous. No other framework, logical or otherwise, makes either of these predictions. Unlike previous experiments, and in contrast to the correlational interpretation discussed above, the causal modeling framework predicts that people should respond "no" to the second question regardless of condition. If A is the cause of B, then B should not operate if A does not.

**Method**

The problem was given to the 78 Brown undergraduates who participated in Experiments 2 and 3. Approximately half were given the explicit and half the non-

explicit prevention questions. Half of each group were given the scenario shown above 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.

**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 4. The 68% giving an affirmative answer to the first question in the Non-explicit Prevention condition replicates the undoing effect seen in the previous studies. The even greater percentage (89%, $z = 2.35$; $p < .01$) in the Explicit condition replicates the finding that the undoing effect is greater when the reason that a variable has the specified value is made explicit. Responses to the second question were almost all negative, demonstrating that people clearly understood that the relevant relation was causal. This rules out the alternative explanation for the earlier studies, that participants didn't interpret the relations as causal but merely as correlational. In this experiment, confidence was uniformly high (approximately 6) in all conditions.

## Experiment 5

Like Experiment 4, this experiment contrasted an explicit and non-explicit prevention using the simplest possible causal structure involving only two events. In addition, it included a noncausal condition in which the conditional relation between the two variables was not obviously causal.

The causal scenarios were as follows:

> 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 ten million dollars. Therefore: If John is a Richman, he will have had ten million dollars at some point in his life.

In the Causal, Nonexplicit prevention condition, the following question was asked:

> Imagine that John had never received the ten million dollars, would he have still been a Richman?

In the Causal, Explicit prevention condition, the question made the nature of the counterfactual antecedent more explicit:

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

We predict a larger undoing effect in the Explicit than Nonexplicit conditions, namely that the majority will respond "yes." We also included a noncausal condition that we refer to simply as "Conditional" for which we expected responding to be more variable:

> John is a Richman. This is a name given to all of the people who have had ten million dollars at some point in their life. Therefore: If John is a Richman, he will have had ten million dollars at some point in his life.
>
> Imagine John's wife had prevented him from ever getting ten million dollars, would he have still been a Richman?

**Method**

Thirty different participants were tested in each condition, a mixture of Brown undergraduates and volunteers tested at the local airport. Otherwise the method was the same as previous experiments.

**Results and Discussion**

Percentage "yes" responses in the three conditions appear in Table 5. A highly significant difference obtained across conditions, $F(2,87) = 29.4$, $MSe = 15$, $p < .0001$. Every single participant in the Explicit condition responded "yes," providing strong support for the undoing effect with explicit prevention. In the Nonexplicit prevention condition, less than 50% (37%) of participants responded "yes." This is the first time we have observed such a small percentage in a causal condition. It may be that the pragmatics of the question was such that the negation of the antecedent did read as

diagnostic of its cause. The percentage responding "yes" in the Conditional case was lower (30%), though the same statistically ($t < 1$), suggesting that not getting $10 million was again diagnostic of not being a Richman.

Confidence judgments indicate participants answered with high confidence in the Explicit condition (mean of 6.0). They were not as confident in the Nonexplicit causal or in the Conditional conditions (means of 4.9 and 5.4, respectively), although the effect of condition was not significant, $F(1,87) = 1.85$; $MSe = 2.23$; n.s.

## Experiment 6

This experiment contrasted Causal, Correlational and Conditional conditions using a context of pure physical causality. In the Causal condition, participants were told

> 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.

The causal model underlying this scenario looks like this:


In the correlational condition, the billiard balls did not cause each other to move but were instead all moved by a fourth variable, a common cause:

> Someone is observing three billiard balls that are constantly moving, each on a separate non-adjacent table. They are all being moved by one large magnet that is in the ceiling of the room. The person notices that: If ball 1 moves, then ball 2 moves. If ball 2 moves, then ball 3 moves.

In the conditional condition, the relations at issue are deontic, not causal:

> Someone is being tested on her logical abilities. Her task is to move as many billiard balls as possible, without breaking the following rules: When certain balls move, other balls have to move too. In particular: If ball 1 moves, then ball 2 moves. If ball 2 moves, then ball 3 moves.

In all three conditions, participants were asked the same questions that involved explicit prevention:

1) Suppose that someone held ball 2 so that it could not move, would ball 1 still move?
2) Suppose that someone held ball 2 so that it could not move, would ball 3 still move?

In the Causal condition, the causal modeling framework predicts the undoing effect, i.e., a response of "yes" to the first question and, because Ball 2 is the cause of Ball 3's movement, a response of "no" to the second question. In the correlational condition, holding Ball 2 should have no effect on other balls as they are all effects of a larger cause, the magnet, so participants should respond "yes" to both questions. In the conditional condition, the logically correct answers would be "no" to question 1 (a modus tollens inference) and question 2 would have no necessary response. Of course, the causal modeling framework makes no claim that people will succeed at responding logically and the trouble people have with the modus tollens form suggests we might not see consistent logical responding.

**Method**

The participants from Experiment 5 participated, 30 in each condition. Otherwise the method was the same as previous experiments.

**Results and Discussion**

Choice data are shown in Table 6. The results in the Causal condition were just as predicted. The vast majority responded "yes" to the question about Ball 1 (80%) and "no" to the question about Ball 3 (90%). In the correlational condition, every participant but one responded as predicted to the first question (97%). Surprisingly, only 40% responded "yes" to the second question. One possibility in this condition is that the scenario failed to make the intended causal model clear. In particular, participants may have understood that the magnet influences Ball 3 only by influencing Ball 2. As usual, the responses were highly variable in the Conditional condition. Overall differences were

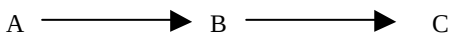
significant for both questions, $F(2,86) = 13.68$; $MSe = 15$; $p < .0001$ for question 1 and $F(2,87) = 3.91$; $MSe = 19$; $p < .05$ for question 2.

As usual, confidence was highest in the Causal condition (mean of 5.9). A one-way analysis of variance across the 3 conditions showed a significant difference, $F(1,88) = 5.71$; $MSe = 2.05$; $p < .05$. The difference is attributable to the Causal condition because confidence judgments did not differ between the Correlational (mean = 5.0) and Conditional (mean = 5.1) conditions, $t < 1$.

## Experiment 7

The causal modeling framework applies to probabilistic arguments as well as deterministic ones. Indeed, the logic of the *do* operator is identical in probabilistic and deterministic contexts and the undoing effect should hold in both. Experiment 7 examines the prediction in a probabilistic context. In accordance with this shift from a deterministic to a probabilistic context, a probability response scale was used. As in most of the previous experiments, causal versions of the arguments were contrasted with conditional versions, and non-explicit prevention questions were contrasted with explicit ones. In addition, in this experiment we looked at two further differences in question format: the contrast between actual and counterfactual intervention, and the contrast between observation and intervention.

Experiment 7 uses the same simple chain structure as Experiment 6:




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.

They were then asked the following probability questions (with a 1-5 response scale):

i. What is the probability that A would have happened if B hadn't happened?
ii. What is the probability that C would have happened if B hadn't happened?

In parallel with previous studies, the causal modeling framework predicts an undoing effect in question (i). That is, when assessing a counterfactual that supposes that B did not occur, participants should mentally sever the link from A to B, and thus not reduce their belief in the occurrence of A. On the probability response scale this would correspond to responses greater than the midpoint (3). In contrast, their responses to question (ii) 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 midpoint of the scale.

The use of a probability response scale also enables us to check whether the undoing effect persists if people have the option to express complete uncertainty (by using the midpoint of the scale). In previous experiments people were given only two response options (yes and no). In each case, they expressed relatively high confidence in their judgments, so it is unlikely that the results would differ if they had been given the opportunity to express uncertainty. Nevertheless, this experiment allows us to verify this directly.

As in all previous experiments except Experiment 4, we contrasted causal to conditional premises. The 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 questions were

i. What is the probability that A would be true if B were false?
ii. What is the probability that C would be true if B were false?

Again we expected that responses would not be systematic with conditional premises. If people use modus tollens then their responses to (i) should be low. However, strictly speaking, modus tollens does not apply with a probabilistic conditional. If people interpret the conditional causally, then of course the predictions are identical to the causal case. The correct response to (ii) is similarly ambiguous. The second premise has no implications when B is false and so people might infer that C remains true, or else they might be confused and just choose to express uncertainty.

As in Experiments 2 versus 3 and Experiments 4 and 5, we contrasted non-explicit prevention to explicit prevention questions. The non-explicit questions were as above. In the causal condition, the explicit prevention questions were:

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?

We predict stronger effects with explicit than non-explicit prevention questions. In the conditional condition, the explicit prevention 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?

Participants may use explicit prevention as a cue that conditional statements should be interpreted causally and thus responses in this condition may prove compatible with causal logic.

The causal modeling framework applies to actual as well as counterfactual intervention. Therefore, we included a Counterfactual condition in this study.

Participants were asked to imagine an intervention rather than being told that an intervention had actually occurred:

i. Imagine a situation where someone intervenes directly on B, preventing it from happening. In that case what is the probability that C would have happened?
ii. Imagine a situation where someone intervenes directly on B, preventing it from happening. In that case what is the probability that A would have happened?

We expected responses in this condition to match those of the Explicit intervention condition.

Finally, the causal modeling framework presupposes a fundamental distinction between observation and intervention. We examined the psychological validity of the distinction by including Observation questions:

i. What is the probability that C would have happened if we observed that B didn't happen?
ii. What is the probability that A would have happened if we observed that B didn't happen?

Unlike previous conditions, these questions explicitly state that B's nonoccurrence was observed, and thus imply that B was not intervened on. Therefore, B should be treated as diagnostic of A and C; in particular, we do not expect the undoing effect. Therefore, the probability of A should be substantially reduced in this condition. Of course, B's nonoccurrence also makes C less likely so the judged probability in question 1 should also be low.

We tested all of these conditions using 3 different scenarios: The abstract scenario above, as well as a scenario concerning physical causality:

When there is gas in the Rocketship's fuel tank, it causes the engine to fire most of the time.
When the engine fires, most of the time it causes the Rocketship to take off.
The Rocketship's fuel tank has gas in it.
The Rocketship takes off.

as well as 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 7 uses a probability response scale, contrasts Causal to Conditional premises, examines 4 varieties of observation/intervention, and uses 3 different scenarios each of a different type, all in the context of probabilistic premises.

## Method

**Design**. All variables were combined factorially: Causal versus Conditional premises x Type of Intervention (Unspecified, Explicit intervention, Counterfactual intervention, Observation) x Scenario (Abstract, Rocketship, 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.

**Participants**. We tested 217 Brown University undergraduates using the same questionnaire format as previous studies. We also tested 160 volunteer participants on the internet using an identical questionnaire. They were obtained by advertising on various websites 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.

**Procedure**. The format of the questionnaire was identical to that of previous experiments except that 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." Also, no confidence judgments were obtained.

**Results and Discussion**

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

When the nature of the intervention was unspecified, little difference was observed between the Causal and Conditional conditions. The undoing effect was 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), $t(41) = 1.5$; s.e. = .16; n.s., and $t < 1$, respectively. Participants were not sure about Event A when told B hadn't happened or that B was false. However, both judgments were higher than corresponding P(C|~B) judgments, $t(41) = 5.09$; s.e. = .17; $p < .0001$ and $t(40) = 3.40$; s.e. = .13; $p < .01$, respectively, suggesting that the negation of B did reduce belief in the occurrence/truth of C to some extent, consistent with a causal reading of the B-C relation.

The pattern in the Observational condition was similar, suggesting that participants treated the negation of B in the Unspecified condition as observational, not interventional. Again, 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, $t(48) = 2.23$; s.e. = .13; $p < .05$ and $t(46) = 1.58$; s.e. = .18; n.s., respectively. Moreover, these were again higher than corresponding P(C|~B) judgments, $t(48) = 3.19$; s.e. = .12; $p < .01$ and $t(46) = 3.28$; s.e. = .13; $p < .01$, respectively. In other words, in the Observational condition, the negation of B was treated as removing any evidence in favor

of A and, to some extent, as evidence against C. Consistent with the causal modeling framework, participants treated observations as correlational evidence and did not exhibit an undoing effect.

Quite a different pattern was observed in the Interventional condition. Here a strong undoing effect occurred, not only in the Causal but in the Conditional cases as well. The mean judged P(A|~B) were appreciably higher than the scale midpoint, 3.9 and 4.1, respectively, $t(48) = 7.75$; s.e. = .12 and $t(47) = 8.32$; s.e. = .13; both p's < .0001. Intervening explicitly to prevent B caused participants to maintain their belief in the occurrence/truth of A. In the Causal case, the nonoccurrence of B suggested to participants that its effect didn't occur either (mean P(C|~B) of 2.3, significantly lower than 3, $t(48) = 4.36$; s.e. = .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 in the Conditional condition, even though B's nonoccurrence did suggest the nonoccurrence of C in the Causal condition.

Judgments after Counterfactual interventions were very similar to judgments in the Interventional condition. Strong undoing effects can be seen for both Causal and Conditional P(A|~B) judgments (means of 3.9 and 4.3, respectively, both greater than 3, $t(40) = 6.44$; s.e. = .14 and $t(50) = 11.05$; s.e. = .12; both p's < .0001). Again, the nonoccurrence of B in the Causal condition lowered the judged probability of C to 2.1, significantly less than 3, $t(40) = 4.81$; s.e. = .18; p < .0001, whereas the falsity of B in the

Conditional condition lowered it to maximal uncertainty (mean of 2.9; $t < 1$). Apparently, participants did not distinguish actual from counterfactual intervention.

The parallel tendencies amongst 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. Indeed, this is a natural interpretation, particularly for the medical and rocketship scenarios.

## General Discussion

These experiments show that the undoing phenomenon is robust and sometimes large. Told that a cause and effect had occurred and then asked to counterfactually assume that the effect had not occurred, people continue to believe in the occurrence of the cause. Undoing was observed when the effect was explicitly prevented by an external agent (Experiments 3-7) and to a lesser extent when the reason for the effect's nonoccurrence was unspecified (Experiments 1, 2, and 4). Undoing was observed for both deterministic (Experiments 1-6) and probabilistic (Experiment 7) arguments. The studies also demonstrate that the causal relations were indeed interpreted as causal by showing that effects were judged not to occur if their sole causes did not (Experiments 4 and 6) and that a relation of the form "A causes B" was not interpreted as a correlation between A and B (Experiment 6). Experiment 7 also showed that participants clearly distinguished between observing the nonoccurrence of an event and an intervention that prevented the event from occurring; undoing obtained after an intervention, but not after an observation. Moreover, the intervention could be either actual or counterfactual (imagined). Finally, Experiments 1-3, 6, and 7 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, a conditional could be interpreted as a causal with enough contextual support (Experiments 1-4, 7). In general, conditionals were not given a consistent interpretation.

The data show that most people obey a rational rule of counterfactual inference, the undoing principle. In every case in which a causal relation existed from A to B, A was known to have occurred and B was explicitly prevented from occurring, the great majority of people judged that A had still occurred. Put this way, undoing seems almost obvious. When reasoning about the consequences of an external intervention or counterfactual supposition of an event, 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 from 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.

Despite the intuitiveness of the undoing principle, its implications are deep and wide-ranging. The most fundamental perhaps is the limit it places on the usefulness of Bayes' rule and its logical correlates for updating belief. 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. And intervention is hardly a rare or special case. Manipulation is an important tool for learning; it is exactly what's required to run the micro-experiments 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 – at least used in the conventional way – will fail as a model of learning.

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 probability is unaffected). It is possible that the frequent observation that people fail to draw valid modus tollens inferences reflects a tendency to interpret apparently logical arguments as causal and "not B" as *do*(B = did not occur).

If this possibility is correct, it would suggest that the interpretation of conditionals varies with the theme of the text that the statements are embedded in. Conditionals embedded in deontic contexts are well known to be interpreted deontically (Manktelow & Over, 1990). Conditionals in other contexts support a variety of different inferences depending on their surrounding context (Almor & Sloman, 1996; cf. Edgington, 1995). The current studies show that when the theme is ambiguous, their interpretation will be highly variable. We 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.

We do not believe that mental model theory can explain our data. Goldvarg and Johnson-Laird (2001) propose 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 is obviously unable to explain the differences we observed between causal and conditional premises. Moreover, because it doesn't allow the possibility "A not B", it is inconsistent with the undoing effect with causal premises.

Goldvarg and Johnson-Laird (2001) do allow however that the set of possibilities can vary with enabling and disabling conditions. To see how this might apply to our problems, consider the simplest case where A causes B, and subjects are asked whether A would still occur if B were prevented from occurring. The statement that B is prevented from occurring presupposes some preventative cause X (e.g., I switch B off). Given X, and the knowledge that X causes not B by virtue of being preventative, 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 that is supposed to represent causal knowledge itself requires causal knowledge to be constructed. The variable X must be invented at the moment at which one learns that B is prevented from occurring. It couldn't exist a priori because that would lead to a combinatorial explosion of models; one would need to represent an enormous number of potentialities: the possibility that Y

enables B even in the presence of disabling condition X, the possibility that X' prevents X, that X'' prevents X', etc. So X must be invented after the intervention, and the set of possible models must then be reconstructed. But if we're reconstructing the possible models, what rules are there to guide us? Why is the model above the only possibility? Without prior causal knowledge another possibility might be

not A X not B

Of course, this possibility does not license the inference to A and so is not consistent with the undoing effect. In sum, mental model reconstruction depends on prior causal models because causal models are the only source of constraints. Pearl (2000) makes an analogous argument against Lewis's (1986) counterfactual analysis of causation. Lewis defines causation in terms of counterfactuals, whereas Pearl argues that it is the causal models that ground (causal) counterfactuals.

Our data support the psychological reality of a central tenet of Pearl's (2000) 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 the undoing principle, Pearl's *do* operator, makes it possible to construct representations that afford valid causal induction and inference – induction of causal relations that support manipulation and control, and inference about the effect of such manipulation, be it from actual physical intervention or merely counterfactual thought about intervention. The *do* operation is precisely what's required to distinguish representations of probability like Bayes' nets from representations of causality.

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, learning and inference can take place by intervening on the representation rather than on the world itself. But none of this can 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 variability 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 can't simply tell me that he's going to replace my knee. I want to know how, what it's going to be replaced with, etc. 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 reasoning. People have a variety of frames available to apply to different problems (Cheng & Holyoak, 1985). Mental models serve particularly well in some domains like 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 amongst events over time.

## References

Almor, A. & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review, 103*, 374-380.

Bara, B.G. & Johnson-Laird, P.N. (1984). *Syllogistic inference*, Cognition, 16.

Edgington, D. (1995). On conditionals. *Mind*, *104*, 235-329.

Evans, J. St. B. T. (1982). *The Psychology of Deductive Reasoning.* London: Routledge & Kegan Paul.

Lipton, P. (1992). Causation outside the law. In H. Gross & R. Harrison (Eds.), *Jurisprudence: Cambridge Essays.* Oxford: Oxford University Press.

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.

Goodman, N. (1955). *Fact, fiction, and forecast.* Harvard University Press.

Hume, D. (1748). *An enquiry concerning human understanding.* London: Millar.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991) *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, D. (1986). *Philosophical papers, vol.2*. Oxford: Oxford Univeristy Press.

Mackie, J.L. (1974). *The cement of the universe.* Oxford: 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*, Chichester: Wiley.

Pearl, J. (2000). *Causality.* Cambridge: Cambridge University Press.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge: The MIT Press.

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

## Acknowledgments

This work was funded by NASA grant NCC2-1217. Some of the results for 2 of the 3 scenarios in Experiments 1-3 along with Experiment 4 were reported at the 2002 Cognitive Science Society conference. We thank Daniel Mochon and Constantinos Hadjichristiditis for their many contributions to this research and Brad Love, Ian Lyons, Peter Desrochers, and Henry Parkin for collecting data. Josh Tenenbaum provided the motivation for Experiment 2.

Table 1. Percentages of participants responding "yes" to two questions about each scenario in both Causal and Conditional conditions of Experiment 1. “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.

| | 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 |

Table 2. Percentages of participants responding "yes" to two questions about each scenario in both Causal and Conditional conditions of Experiment 2. “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.

| | 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 |

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

| | Causal | | Conditional | |
|---|---|---|---|---|
| Scenario | E holds | B holds | E holds | B holds |
| Abstract | 75 | 80 | 50 | 67 |
| Robot | 75 | 75 | 83 | 67 |
| Political | 65 | 90 | 56 | 83 |

Table 4. Percentages of participants responding "yes" to questions in the Rocketship scenario of Experiment 4 given questions with antecedents non-explicitly or explicitly prevented.

| Question | Non-explicit Prevention | Explicit Prevention |
|---|---|---|
| i. if not B, then A? | 68 | 89 |
| ii. if not A, then B? | 2.6 | 5.3 |

Table 5. Percentages of participants responding "yes" to questions in the Richman scenario of Experiment 5 given in Nonexplicit and Explicit Causal and Conditional conditions.

| Nonexplicit Prevention | Explicit Prevention | Conditional |
|---|---|---|
| 37 | 100 | 30 |

Table 6. Percentages of participants responding "yes" to two questions in the three billiard ball problems of Experiment 6. “Ball 1 moves” and “Ball 3 moves” refer to questions 1 and 2 of Experiment 6, respectively.

| | Ball 1 moves | Ball 3 moves |
|---|---|---|
| Causal | 0.80 | 0.10 |
| Correlational | 0.97 | 0.40 |
| Conditional | 0.45 | 0.33 |

Table 7. Mean probability judgments on 1-5 scale for two questions in Experiment 7, four types of intervention and causal and conditional versions, averaged across three scenarios. *P(C|~B)* refers to question i. and *P(A|~B)* to question ii.

| | Causal | | Conditional | |
|---|---|---|---|---|
| | *P(C\|~B)* | *P(A\|~B)* | *P(C\|~B)* | *P(A\|~B)* |
| Unspecified | 2.4 | 3.2 | 2.6 | 3.0 |
| Observational | 2.3 | 2.7 | 2.8 | 3.3 |
| Interventional | 2.3 | 3.9 | 3.0 | 4.1 |
| Counterfactual | 2.1 | 3.9 | 2.9 | 4.3 |