Counterfactual Undoing in Deterministic Causal Reasoning

Steven A. Sloman (Steven_Sloman@brown.edu)
Department of Cognitive & Linguistic Sciences, Box 1978
Brown University, Providence, RI 02912 USA

David A. Lagnado (David_Lagnado@Brown.Edu)
Department of Cognitive and Linguistic Sciences, Box 1978
Brown University, Providence, RI 02912 USA

Abstract

Pearl (2000) offers a formal framework for modeling causal and counterfactual reasoning. By virtue of the way it represents intervention on a causal system, the framework makes predictions about how people reason when asked counterfactual questions about causal relations. Four studies are reported that test the application of the framework to deterministic causal and conditional 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 in different ways.

Introduction

Many questions are decided by causal analysis. In the law, issues of negligence concern who caused an outcome and, at least under common law, the determination of guilt requires evidence of a causal chain leading to a crime. Evidence that might increase the probability of guilt (e.g., an accused's race) is impermissible 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 just as prevalent in science, engineering, politics, indeed in every domain that involves human prediction and control.

Causal analysis is often difficult because it depends not only on what happened, but also on what *might* have happened (Mackie, 1974). Thus the claim that A caused B will often imply that if A had not occurred, then B would not have occurred. Likewise, the fact that B would not have occurred if A had not often suggests that A caused B.

This explains a fundamental law of experimental science: Mere observation can only reveal a correlation, not a causal relation. That's why causal induction requires manipulation, control over an independent variable such that changes in its value will determine the value of the dependent variable whilst holding other relevant conditions constant. Everyday causal induction has these same requirements. Causal inductions in everyday contexts are aided by manipulation of potential causes, by people *intervening* on the world rather than just observing it (the conditions favoring intervention are spelled out in Pearl, 2000; Spirtes, Glymour, & Scheines, 1993).

If we already have some causal knowledge, then certain causal questions can be answered without actual intervention. Some of those questions can be answered through mental intervention, by imagining a counterfactual situation in which a variable is manipulated and determining the effects of 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...).

Pearl (2000) offers a causal modeling framework that covers such counterfactual reasoning. The framework makes predictions about how people reason when asked counterfactual questions about causal relations. Pearl's analysis extends to relations of probabilistic causality but this paper is limited to studies of deterministic arguments. Before describing those studies, we briefly review the relevant aspects of Pearl's analysis.

Observation vs. Causation (Seeing vs. 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:

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

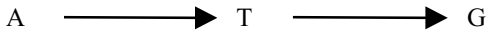
Conditional probabilities are symmetric in the sense that, if well-defined, their converses are well-defined too. In fact, given the marginal probabilities of the relevant variables, Bayes' rule tells us how to evaluate the converse:

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

Doing

To represent action, Pearl proposes an operator *do(•)* that controls both the value of a variable that is manipulated as well as a 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. Because the *do* operation renders T and G 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.

Do we "do"?

Consider the following Causal Argument (1) in which A, B, C, and D are the only relevant events:

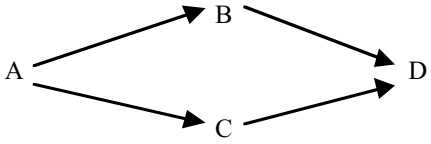
A causes B.
A causes C.
B causes D.
C causes D.
D definitely occurred.

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

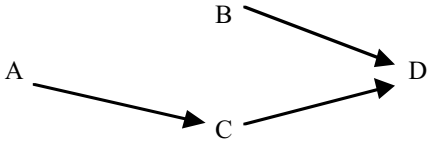
i. If B had not occurred, would D still have occurred?___ (yes or no)

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. Thus, all 4 events have occurred. When asked what would have happened if B had not occurred, 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 already decided 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 argument uses causal relations and therefore lies outside the propositional domain. The closest they can 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 Causal Argument (1) is valid, we can consider Abstract Conditional Argument (1):

If A then B.
If A then C.

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

The corresponding questions were:
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 an argument except to say that, because it does not necessarily concern causal relations, responses could well be different from those for the causal argument. 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, whilst adding 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 these "material conditional" theories make any prediction for the causal arguments, these should correspond to their prediction for comparable conditional arguments.

Experiment 1

Method. 238 University of Texas at Austin undergraduates were given one of the two arguments shown and asked the listed questions.

Results. Responses are shown in Table 1. The predictions of the causal modeling framework were supported for the causal arguments but not for the conditional arguments. 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. Responses to the conditional argument were more variable: no one strategy for interpreting and reasoning with conditional statements dominated.

Table 1: Percentages of participants responding "yes" to Abstract Causal and Conditional Arguments (1).

Question	Causal	Conditional
i. D holds	80%	57%
ii. A holds	79%	36%

These results were replicated with two additional arguments that used an identical causal or logical structure but added semantic content to the problems. For example, one pair of arguments concerned a robot. Here is the causal version of that problem (Robot Causal Argument 1):

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.

i. If Beam B had not struck the solar panel, would the robot have been activated?
ii. If Beam B had not struck the solar panel, would the original (500 unit) beam have been shone through the prism?

The same 238 undergraduates were given either the causal or conditional version of this problem. Their responses are shown in Table 2.

Table 2: Percentages of participants responding "yes" to Robot Causal and Conditional Arguments (1).

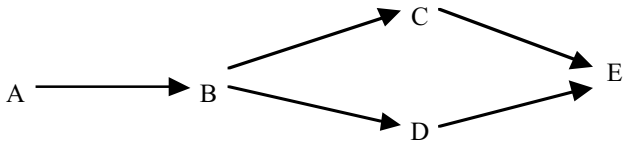
Question	Causal	Conditional
i. robot activated	80%	63%
ii. beam shone	71%	55%

The results are very close to those of the abstract problem except that a higher percentage of participants said "yes" in the conditional version of this problem, $z = 2.83$; $p < .01$. This may have occurred because a larger proportion interpreted the "if-then" connectives of the conditional version as causal relations. The clear physical causality of the robot problem lends itself to causal interpretation.

Experiment 2

One might argue that the difference between the causal and conditional arguments in the previous examples 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 pragmatic implicatures of the two contexts. In particular, perhaps the causal context presupposes the occurrence of A more than the conditional context presupposes the truth of A. It's more plausible that D would be true in the conditional arguments even if A were false than that D would have occurred in the causal arguments even if A had not. If so, then the greater likelihood of saying "yes" in the causal scenarios could be due to these different presuppositions rather than different likelihoods of undoing.

To control for this possibility as well as to replicate the effect, we examined causal and conditional versions of arguments 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 with one of the problems shown, here is the causal version of the abstract problem (Causal Argument 2):

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. 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 and robot problems described above.

Results. The results, shown in Tables 3 and 4, are comparable to those from the earlier problems, although the proportion of "yes" responses tended to be lower in the causal condition, especially for the likelihood of the beam shining if the solar panel had not been struck (only 55% in Table 4).

Table 3: Percentages of participants responding "yes" to Abstract Causal and Conditional Arguments (2).

Question	Causal	Conditional
i. E holds	70%	45%
ii. B holds	74%	50%

Table 4: Percentages of participants responding "yes" to Robot Causal and Conditional Arguments (2).

Question	Causal	Conditional
i. robot activated	90%	75%
ii. beam shone	55%	45%

A difference between causal and conditional arguments again obtained for Abstract arguments, $z = 2.20$; $p = .01$, but not for Robot ones, $z = 1.18$; n.s. The difference for Abstract arguments suggests that the earlier results cannot be attributed entirely to different pragmatic implicatures from causal and conditional contexts. The overall reduction in "yes" responses could be due to either a different participant population, 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 not entirely unreasonable 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 determines the value of D. To simulate such a situation, we repeated Experiment 2, but made the interventional prevention of D explicit.

Method. Participants saw exactly the same sets of premises in both causal and conditional contexts, but were asked different questions, questions that made the external prevention of D explicit (Causal and Conditional Arguments 2EP). 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 context, the questions in the causal and conditional versions were identical (only the paragraphs describing the situation differed):

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?

Responses were obtained from either 18 or 20 Brown undergraduates.

Results. Results are shown in Tables 5 and 6. The probability of saying "yes" was higher in the explicit prevention context than in its absence, but not significantly so, z = 1.16 and 1.39 for Abstract and Robot arguments, respectively. The two may not differ statistically 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 (E) still held. In other words, the effect of explicitly preventing D is well captured by the *do* operator.

Table 5: Percentages of participants responding "yes" to Abstract Causal and Conditional Arguments (2EP), prevention of the antecedent explicit.

Question	Causal	Conditional
i. E holds	75%	50%
ii. B holds	80%	67%

Table 6: Percentages of participants responding "yes" to Robot Causal and Conditional Arguments (2EP).

Question	Causal	Conditional
i. robot activated	75%	83%
ii. beam shone	75%	67%

An unexpected byproduct of explicit prevention was to increase the proportions of "yes" responses in even the conditional context, z = 1.80; p < .05. This probably occurred because the explicit prevention context made it more likely that the arguments 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.

One implication of this observation is 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). The Abstract Conditional Arguments (1) and (2) above show that when the theme is ambiguous, the interpretation will be highly variable. Robot Conditional Argument (2EP) shows that when the theme is causal, conditionals will be interpreted causally.

Experiment 4

The final experiment attempts to replicate the observations made thus far by showing the undoing effect as well as the enhancement of the effect in an explicit prevention context. Moreover, it does so using an if-then statement in order to show that a conditional statement can be treated as causal in an appropriate context.

Method. The following scenario was described to 78 Brown undergraduates:

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

The scenario assumes the simplest possible causal graph:

Notice that the relation between A and B is stated using an if-then construction. Approximately half the participants, in the non-explicit prevention condition, were 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?

The remaining half, in the explicit prevention condition, 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 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 nonoperative is less ambiguous. No other framework, logical or otherwise, makes either of these predictions. Finally, 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.

Results. The results are shown in Table 7. 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 are clearly

understanding that the relevant relation is causal. This rules out an alternative explanation for the earlier studies, that people were treating causes and effects as disconnected because they didn't interpret the relations as causal but merely as correlational.

Table 7: Percentages of participants responding "yes" to questions in the Rocketship scenario 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%

Discussion

These data show that most people obey a rational rule of counterfactual inference, the undoing principle. When reasoning about the consequences of a 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 an experimenter manipulates the brightness of a computer monitor, one should not assume that the monitor needs replacing.

The demonstrations all described a deterministic causal system. The undoing principle also applies to probabilistic causes however.

These 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.

More generally, the findings are consistent in a qualitative sense with the view of cognition assumed by Pearl (2000) following Spirtes, Glymour, and Scheines (1993). Their 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.

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 hip. 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. But the presence of a calculus for causal inference removes any doubt that it's an important one.

Acknowledgments

This work was funded by NASA grant NCC2-1217. We thank Brad Love for his help and Daniel Mochon, Ian Lyons, and Peter Desrochers for collecting data. Josh Tenenbaum provided an important insight.

References

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

Goldvarg, E., & Johnson-Laird, P.N. (2001). Na ve causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565-610.

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

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.