Reasoning with Causal Relations

Yevgeniya Goldvarg (goldvarg@phoenix.princeton.edu)
Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08544

Philip N. Johnson-Laird (phil@clarity.princeton.edu)
Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08544

Abstract

The mental model theory postulates that reasoners build models of the situations described in premises, and that each model represents a possibility. The present paper proposes that causal relations, such as A causes B and A allows B, have meanings that concern only possibilities and a temporal constraint that B cannot precede A. This theory predicts that causes and enabling conditions differ in meanings, contrary to a long tradition in philosophy and psychology that they are logically indistinguishable. It also predicts that individuals should reason about causation on the basis of mental models rather than on fully explicit models. Three experiments corroborated these predictions.

Introduction

People reason about causal relations in order to predict what will happen. For example, given the following inference:

A prevents B.
B causes C.
What follows?

most people respond: A prevents C, but this conclusion is invalid. Our goal in the present paper is to present a theory of causation that predicts this phenomenon and that gives a general account of both the meaning of causal relations and of how people reason from them. The theory is based on mental models. Our plan in what follows is, first, to outline the theory of mental models and its extension to causal relations; second, to explain how enabling conditions and causes differ; and, third, to describe how mental models underlie causal deductions.

The Mental Model Theory of Causal Relations

The mental model theory postulates that reasoning is a semantic process, which depends on understanding the meaning of premises and using this meaning to envisage the situations that are possible given the truth of the premises (Johnson-Laird and Byrne, 1991). A mental model accordingly represents a possibility, and its structure and content capture what is common to the different ways in which the possibility may occur. A conclusion is necessary - it must be true – if it holds in all the models of the premises. It is possible – it may be true – if it holds in at least one model of the premises (Bell and Johnson-Laird, 1998). And its probability – its likelihood of being true – depends on the proportion of models of the premises in which it is true (Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni, 1998). The principle of truth is a fundamental assumption of the theory: in order to minimize the load on working memory, mental models normally represent only what is true. This principle applies at two levels: individuals represent only true possibilities; and within those possibilities they represent the literal propositions in the premises only when they are true. Consider the following exclusive disjunction, for example:

There is a not a circle or else there is a triangle.

The mental models of the disjunction represent only the true possibilities, and within them, they represent only their true components:

¬o
•

where '¬' denotes negation, 'o' denotes a model of the circle, '•' denotes a model of the triangle, and each row denotes a model of a separate possibility. Hence, the first model does not represent explicitly that it is false that there is a triangle in this case; and the second model does not represent explicitly that it is false that there is not a circle in this case. Reasoners make a 'mental footnote' to keep track of this false information, but these footnotes are soon likely to be forgotten. Indeed, the failure to cope with falsity gives rise to illusory inferences about modal conclusions, i.e. inferences that nearly everyone makes, but that are wrong (Johnson-Laird & Goldvarg, 1997). Only fully explicit models of what is possible given the exclusive disjunction represent the false components in each model:

¬o ¬•
o •

where a false affirmative (there is a triangle) is represented by a true negation, and a false negative (there is not a circle) is represented by a true affirmative.

The theory deals with causal relations in a natural way. Its first assumption is that causal relations concern physical

possibilities and place a temporal constraint on them. The relation A causes B means that there are three possibilities:

a b
¬ a b
¬ a ¬ b

and that B cannot precede A in time. Likewise, A allows B means that there are three possibilities:

a b
a ¬ b
¬ a ¬ b

and that B cannot precede A in time. These possibilities correspond to fully explicit models of the premises, but the theory postulates that logically-untrained individuals normally represent only those possibilities in which A and B are true. Thus, their mental models of A causes B are as follows:

a b
. . .

where the ellipsis represents those possibilities in which A is false. Table 1 presents the mental models and the fully explicit models of the four principal causal relations.

Table 1: The mental models and the fully explicit models of the four main causal relations.

Causal Relation	Mental Models	Fully Explicit Models
A causes B	a b . . .	a b ¬ a b ¬ a ¬ b
A allows B	a b . .	a b a ¬ b ¬ a ¬ b
A prevents B	a ¬ b . . .	a ¬ b ¬ a b ¬ a ¬ b
A allows not-B	a ¬ b . . .	a ¬ b a b ¬ a b

The fully explicit models of A causes B correspond to A being sufficient for B. Similarly, the fully explicit models of A allows B correspond to A being necessary for B. In both cases, however, the causal models also embody a temporal constraint that is not essential for necessary or sufficient conditions. There are, of course, many other ways to describe each of the causal relations. If both A causes B and A allows B, then a strong causal relation holds between them corresponding to the following models:

a b
¬ a ¬ b

Likewise, there is a strong relation that combines A prevents B and A allows not-B:

a ¬ b
¬ a b

Experiment 1:

Causal Possibilities

Our first experiment was designed to test the theory embodied in Table 1. The participants' task was to list the true possibilities and the false possibilities for five causal assertions. They were the three relations: A causes B, A allows B, and A prevents B; and two different ways of paraphrasing cause: A prevents not-B, and not-A allows not-B. We devised five lexical contents that were rotated over the relations for different participants, so that each participant saw each content just once, but in the experiment as a whole each content occurred equally often with the five relations. The contents are illustrated by the following examples:

Exercise that is excessive causes the development of angina.

Use of solar energy prevents the occurrence of global warming.

Twenty Princeton undergraduates carried out the experiment. They were told to list all possibilities that would make an assertion true and all possibilities that would make it false. They could list them in any order. The model theory makes three main predictions. First, as Table 1 implies, the participants should list both true cases and false cases. Second, they should start with the possibilities corresponding to the mental models of the relations. Third, they should tend to confuse allows with causes, because they have the same mental models.

Experiment 2:

Enabling Conditions and Causes

Many philosophers and psychologists have argued that there is no logical distinction between the meaning of enabling conditions (as expressed by allows) and causes (see e.g. Mill, 1843; Mackie, 1980; Hart and Honoré, 1985). What then does distinguish them? A wide variety of answers is to be found in the literature: enabling conditions occur early but causes immediately precipitate the effect (Mill, 1843); enabling conditions are common but causes are rare (Hart and Honoré, 1985); enabling conditions are the norm but causes violate the norm (see e.g. Kahneman and Tversky, 1982; Kahneman and Miller, 1986; Einhorn and Hogarth, 1986); enabling conditions are constant but causes are inconstant (Cheng and Novick, 1991); enabling conditions are irrelevant to explanations but causes are relevant(e.g. Mackie, 1980; Turnbull and Slugoski, 1988; Hilton and Erb , 1996). And there are still other views (see Hesslow, 1988, for a review).

All of these hypotheses could be true, yet, as Experiment 1 showed, people can draw a distinction between the meaning of enabling conditions and causes. The two *are* logically distinct. We propose that causal interpretation depends crucially on how people conceive the circumstances of events, that is, on the particular states that they consider to be possible, whether real, hypothetical, or counterfactual. Consider, for example, the following scenario:

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

The circumstances described here correspond to the fully explicit possibilities:

Sunlight	Fertilizer	Grow-well
Sunlight	¬ Fertilizer	Grow-well
Sunlight	¬ Fertilizer	¬ Grow-well
¬ Sunlight	Fertilizer	¬ Grow-well
¬ Sunlight	¬ Fertilizer	¬ Grow-well

In these circumstances, sunlight is an enabling condition for the flowers to grow well:

Sunlight	Grow-well
Sunlight	¬ Grow-well
¬ Sunlight	¬ Grow-well

In contrast, all four possible contingencies occur concerning the fertilizer and the flowers growing well. But, the sunlight enables the fertilizer to cause the flowers to grow well.

Now, consider the following scenario:

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

These circumstances correspond to the possibilities:

Sunlight	Fertilizer	Grow-well
¬ Sunlight	Fertilizer	Grow-well
¬ Sunlight	Fertilizer	¬ Grow-well
Sunlight	¬ Fertilizer	¬ Grow-well
¬ Sunlight	¬ Fertilizer	¬ Grow-well

In these circumstances, the respective causal roles have been swapped around: the fertilizer enables the sunlight to cause the flowers to grow well. In both these cases, the cause and the enabling condition have different meanings and neither of them is constant in the circumstances .

Experiment 2 tested whether 20 Princeton Undergraduates could distinguish causes and enabling conditions. Cheng and Novick (1991) showed that individuals could distinguish them in cases where the enabling conditions were constant and the causes were inconstant. Our aim was to show that they could do so when neither enabling conditions nor causes were constant. We prepared eight pairs of scenarios such as the examples above, in which there were two precursors to the effect and their respective roles as enabling condition and cause were counterbalanced in the two scenarios. We also prepared versions of the pairs of scenarios in which we reversed the order of mention of cause and enabling condition. Thus for the previous examples, there were scenarios as follows:

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

and:

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

Each participant encountered just one version of a particular scenario, but an equal number of the four sorts of scenario in the experiment as a whole. The order of presentation was randomized for each participant. The task was to identify the enabling condition and the cause in each scenario, and the experiment included two filler items – one in which there were two joint causes and one in which there were no causes. The participants correctly identified the enabling conditions and causes on 85% of trials, and every participant was correct more often than not ($p = 0.5^{20}$). Hence, individuals can distinguish enabling conditions from causes even in scenarios where neither is constant.

Experiment 3:

Mental Models in Causal Deductions

The previous experiments corroborate the model theory of naïve causation, but they may be consistent with other accounts. Rips (1994, p. 336), for example, argues that it should be possible to frame an account of causal reasoning based on formal rules. Is there any critical test to show that individuals are using models in reasoning about causal relations? The answer is that mental models predict the occurrence of certain systematic errors in reasoning. Consider, first, an inference of the following form:

A causes B.
B prevents C.

What follows?

The premises yield the following mental models (seeTable 1):

a b ¬ c
. . .

Reasoners should therefore conclude:

A prevents C.

The fully explicit models of the premises also support this conclusion, and so it is valid. Second, consider an inference of the following form:

A prevents B.
B causes C.
What follows?

The premises yield the following mental models:

a ¬ b
b c
. . .

Reasoners should therefore conclude:

A prevents C

because A occurs in a model that does not yield C, and C occurs in a model without A. In this case, however, the conclusion is invalid. The fully explicit models of the premises are as follows:

a	¬ b	c
a	¬ b	¬ c
¬ a	b	c
¬ a	¬ b	c
¬ a	¬ b	¬ c

Table 2: The 16 causal inferences in Experiment 3, and the conclusions that the mental models of the premises support (invalid conclusions are asterisked). The table also shows the number of participants (n = 20) who drew the predicted conclusions.

	FIRST PREMISE			
	A causes B	A allows B	A prevents B	Not-A causes B
SECOND PREMISE B causes C	A causes C: 20	*A allows C: 18	*A prevents C: 19	Not-A causes C: 20
B allows C	*A allows C: 19	A allows C: 19	A prevents C: 20	*Not-A allows C: 20
B prevents C	A prevents C: 20	*A allows not-C: 14	*A prevents C: 15	Not-A prevents C: 20
Not-B causes C	*A allows not-C: 8	A allows not-C 12	A causes C: 17	*Not-A prevents C:15

As these models show, there is no causal relation between A and C.

Experiment 3 examined all possible inferences in which the premises were laid out in the following figure:

A - B.
B - C.

and each of the four causal relations was systematically inserted in order to yield the 16 distinct inferences shown in Table 2. The model theory predicts that reasoners should draw a conclusion in all 16 cases, but as the Table shows, half of these conclusions are valid, and half of them are invalid.

Twenty Princeton undergraduates carried out the 16 inferences in random orders, and A, B, and C, referred to abstract topics, e.g. obedience allows motivation to increase; increased motivation causes eccentricity. These contents were rotated over the set of inferences so that each content occurred equally often with each sort of inference in the experiment as a whole. The participants drew the predicted conclusions whether they were valid (93% of conclusions as predicted) or invalid (80% of conclusions as predicted). Each participant drew more predicted than unpredicted conclusions ($p = 0.5^{20}$) and 15 out of the 16 inferences (Sign test, $p < 0.00$).

General Discussion

The experiments corroborated the model theory's account of causal relations. Individuals envisage the possibilities corresponding to causal relations. They can distinguish between enabling conditions and causes, even though, according to the model theory, neither relation necessarily depends on the constant presence of an antecedent (pace Cheng, 1997). Likewise, as the theory predicts, reasoners drew systematically invalid conclusions supported by the mental models of the premises. We have carried out other studies that also support the theory. We conclude that individuals represent the meanings of causal relations in mental models. These models denote possibilities and constrain the order of antecedents and effects: effects do not precede their antecedents. The interpretation of causal relations depends on how individuals represent the circumstances, i.e., the models that they envisage of enabling conditions, causes, and effects. Because mental models do not make the possibilities fully explicit, reasoners should be misled in the case of certain inferences. Experiment 3 corroborated this prediction.

Is there any alternative explanation of our results? One possibility is that people use formal rules of inference to make causal deductions. No such account currently exists, and it is hard to see that it could account, say, for the results of Experiment 1, which depend on a semantic interpretation of causal verbs. Another class of psychological theories postulates that cause is a probabilistic notion (see e.g. Suppes, 1970; Cheng, 1997):

A cause B $=_{def}$ P(B | A) > P(B | not A)

In our view, probabilities enter into the induction of causal relations from empirical observations, but not their meaning. Thus, for example, naïve reasoners distinguish between the assertion:

Smoking causes lung cancer

and:

Smoking causes lung cancer with a certain probability.

The latter claim would be superfluous if causal claims tacitly embodied probabilities. Likewise, the definition above applies equally to causes and enabling conditions. In fact, cause may appear to be probabilistic because of the role of enabling conditions. The circumstance often allow cases in which the cause does not lead to the effect, because the enabling condition fails to hold.

Is there anything more to the meaning of causal relations apart from possibilities and the temporal constraint? We have yet to discern any such missing element.

Acknowledgments

We are grateful to the following colleagues for their helpful advice: Victoria Bell, Zachary Estes, Hansjoerg Neth, Mary Newsome, Sergio Moreno Rios, Vladimir Sloutsky, Jean-Baptiste van der Henst, and Yingrui Yang.

References

Bell, V., and Johnson-Laird, P.N. (1998) A model theory of modal reasoning. Cognitive Science, 22, 25-51.

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

Cheng, P.W., and Novick, L.R. (1991) Causes versus enabling conditions. Cognition, 40, 83-120.

Einhorn, H.J., and Hogarth, R.M. (1986) Judging probable cause. Psychological Bulletin, 99, 3-19.

Hart, H.L.A., and Honoré, A.M. (1985) Causation in the Law. Second Edition. Oxford: Clarendon Press.

Hesslow, G. (1988) The problem of causal selection. In Hilton, D.J. (Ed.) Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality. pp. 11-32. Brighton, Sussex: Harvester Press.

Hilton, D.J., and Erb, H-P. (1996) Mental models and causal explanation: Judgements of probable cause and explanatory relevance. Thinking & Reasoning, 2, 273-308.

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

Johnson-Laird, P.N., and Goldvarg, Y. (1997). How to make the impossible seem possible. Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, 354-357.

Johnson-Laird, P.N., Legrenzi, P., Girotto, P., Legrenzi, M.S., and Caverni, J-P. (1998) Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88.

Kahneman, D., and Miller, D.T. (1986) Norm theory: Comparing reality to its alternative. Psychological Review, 93, 75-88.

Kahneman, D., and Tversky, A. (1982) The simulation heuristic. In Kahneman, D., Slovic, P., and Tversky, A. (Eds.) Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press.

Mackie, J.L. (1980) The Cement of the Universe: A Study in Causation. Second edition. Oxford: Oxford University Press.

Mill, J.S. (1843/1973) A System of Logic Ratiocinative and Inductive. Toronto, Ontario: University of Toronto Press.

Rips, L.J. (1994) The Psychology of Proof. Cambridge, MA: MIT Press.

Suppes, P. (1970). A Probabilistic Theory of Causality. Amsterdam: North-Holland.

Turnbull, W., and Slugoski, B.R. (1988) Conversational and linguistic processes in causal attribution. In Hilton, D. (Ed.) Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality. pp. 66-93. Brighton, Sussex: Harvester Press.