CHAPTER

10

Mental Models and Causation

P. N. Johnson-Laird and Sangeet S. Khemlani

Abstract

The theory of mental models accounts for the meanings of causal relations in daily life. They refer to seven temporally-ordered deterministic relations between possibilities, which include *causes*, *prevents*, and *enables*. Various factors—forces, mechanisms, interventions—can enter into the interpretation of causal assertions, but they are not part of their core meanings. Mental models represent only salient possibilities, and so they are identical for *causes* and *enables*, which may explain failures to distinguish between their meanings. Yet, reasoners deduce different conclusions from them, and distinguish between them in scenarios, such as those in which one event enables a cause to have its effect. Neither causation itself nor the distinction between *causes* and *enables* can be captured in the pure probability calculus. Statistical regularities, however, often underlie the induction of causal relations. The chapter shows how models help to resolve inconsistent causal scenarios and to reverse engineer electrical circuits.

Key Words: abduction, causes, deduction, determinism, explanation, mental models, nonmonotonic reasoning

Hume (1748/1988) remarked that most reasoning about matters of fact depends on causal relations. We accordingly invite readers to make two inferences:

1. Eating protein will cause Evelyn to gain weight.
Evelyn will eat protein.
Will Evelyn gain weight?

2. Marrying Viv on Monday will enable Pat to be happy.
Pat will marry Viv on Monday.
Will Pat be happy?

In a study of several hundred highly intelligent applicants to a selective Italian university, almost all the 132 participants making the first sort of inference responded "yes" (98%), whereas over two-thirds of a separate group of 129 participants making the second sort of inference responded "perhaps yes, perhaps no" (68%) and the remainder in this group responded "yes" (Goldvarg & Johnson-Laird, 2001). These inferences stand in need of an explanation, and one aim of the present chapter is to show that the theory of mental models explains them and causal reasoning in general.

Causation has created controversy for centuries. Some have argued that the notion is irrelevant (Russell, 1912–1913), ill defined (Lindley & Novick, 1981), and inconsistent (Salsburg, 2001, pp. 185–186). Scholars also disagree about its foundations, with whether, for instance, causal relations are objective or subjective, and with whether they hold between actions, events, or states of affairs. Certainly, actions can be causes, such as throwing a switch to cause a light to come on. But, as the proverb says, "for want of a nail the kingdom was lost," and so causes can be negative states of affairs as well. We say no more about these philosophical matters, but we do need to consider the consistency of causal concepts.

Causation is built into so much of language that the concept is hardly inconsistent unless languages themselves are inconsistent (see Solstad & Bott, Chapter 31 in this volume). A causal assertion, such as *eating protein causes Evelyn to lose weight*, can be paraphrased using verbs, such as *makes, gets*, and *forces*. The sentence can also be paraphrased in a conditional assertion: *If Evelyn eats protein, then Evelyn will lose weight*. Many verbs embody causal relations in their meanings. For example, an assertion of the form *x lifts y* can be paraphrased as *x* does something that causes *y* to move upward; an assertion of the form *x offers y to z* can be paraphrased as *x* does something that enables *y* to possess *z*; and an assertion of the form *x hides y from z* can be paraphrased as *x* does something that prevents *z* from seeing *y*. These paraphrases reveal that certain concepts, expressed here in *move, possess*, and *see*, underlie "semantic fields" of verbs, whereas other concepts, expressed in *causes, enables*, and *prevents*, occur in many different semantic fields (Miller & Johnson-Laird, 1976). Indeed, it is easier to frame informative definitions for verbs with a causal meaning than for verbs lacking such a meaning (Johnson-Laird & Quinn, 1976).

Inconsistency arises among beliefs about causation. Some people believe that every event has a cause; some believe that an action or intervention can initiate a causal chain, as when a trial begins in an experiment. And some believe both these propositions (e.g., Mill, 1874). But, they are inconsistent with one another. If every event has a cause, then an action cannot initiate a causal chain, because the action has an earlier cause, and so on, back to the ultimate cause or causes of all chains of events. Beliefs, however, are not part of the meanings of terms. Both *every event has a cause* and *every intervention initiates a causal chain* make sense, and the meaning of *cause* should not rule out either assertion as false. The problem, of course, is to separate beliefs from meanings, and the only guide is usage. We now introduce the theory of mental models.

The origins of the theory go back to Peirce's (1931–1958, Vol. 4) idea that diagrams can present moving pictures of thoughts. The psychologist and physiologist Kenneth Craik (1943) first introduced mental models into psychology. His inspiration was machines such as Kelvin's tidal predictor, and he wrote that if humans build small-scale models of external reality in their heads, they can make sensible predictions from them. Craik's ideas were programmatic and untested, and he supposed that reasoning depends on verbal rules. In contrast, the modern theory began with the idea that reasoning itself is a process of simulation based on mental models (Johnson-Laird, 1983).

The basic principles of the theory of mental models—the "model theory," for short—apply to any domain of reasoning, given an account of the meaning of the essential concepts in the domain (e.g., Khemlani & Johnson-Laird, 2013). The present chapter therefore begins with a theory of the meaning of causal relations, including those exemplified in the preceding inferences, taking pains to distinguish meanings from beliefs. It also distinguishes meanings from their interpretation, which yields mental models of the situations to which meanings refer. Models can be static, or they can unfold in time kinematically in a mental simulation of a sequence of events in a causal chain. Both sorts of models yield inferences, and the chapter considers the three main sorts of reasoning: deduction, induction, and abduction. On occasion, it contrasts the model theory with alternative accounts of causation. The aim is not polemical, but to use the contrast to clarify the theory. The chapter concludes with a summary of outstanding problems.

The Meanings of Causal Relations

Basic Causal Relations

This section presents the model theory of the meanings of everyday causal assertions. It also determines how many different sorts of causal relations exist. Theorists tend not to address this question, and often assume that there is just one—the relation of cause and effect, which may bring about an event or prevent it (Mill, 1874). They argue that *causes* and *enables* do not differ in meaning: causes are abnormal, whereas enablers are normal (Hart & Honoré, 1985), causes are inconstant whereas enablers are constant in the situation (Cheng & Novick, 1991), causes violate a norm whereas enablers do not (Einhorn & Hogarth, 1986), or causes are conversationally relevant whereas enablers are not (Hilton & Erb, 1996; Mackie, 1980). In contrast, the model theory distinguishes the meaning of the two (see also Wolff and Thorstad, Chapter 9 in this volume), and fixes the number of causal relations. A clue is in the mappings in Table 10.1 from quantifiers ranging over possibilities to causation and thence to obligation. A necessary proposition holds in all relevant possibilities, a possible proposition holds in at least one of them, an impossible proposition holds in none of them, and a proposition that is possibly not the case fails to hold in at least one of them. Likewise, if a cause occurs, then its effect is necessary, if an enabling condition occurs then its effect is possible, if a preventive condition occurs then its effect is impossible, and if a condition

Table 10.1 The Set of Mappings over Six Domains from Quantifiers Through Causal Concepts to Modal Notions and on to Deontic Concepts

Quantified Assertions	Causal Verbs	Modal Concepts	Modal Auxiliary Verbs	Deontic Concepts	Deontic Verbs
In *all* possibilities, it occurs.	causes	necessary	will occur	compulsory	obligates
In *some* possibilities, it occurs.	allows/ enables	possible	may occur	permissible	allows/ permits
In *no* possibilities, it occurs.	prevents	impossible	cannot occur	impermissible	prohibits
In *some* possibilities, it does *not* occur.	allows not/ enables not	possible not	may not occur	permissible not	allows not/ permits not

allows an effect *not* to occur then it is possible for it not to occur.

Table 10.1 treats *allows* and *enables* as synonyms, but they have subtle differences in usage. Of course, a single quantifier, such as *some*, together with negation, allows all four cases to be defined (e.g., *all* is equivalent to *it is not the case that some are not*); and the same applies *mutatis mutandis* for each column in Table 10.1 (e.g., *A causes B to occur* is equivalent to *A does not allow B not to occur*). The mappings to deontic concepts are analogous, where deontics embraces what is permissible within morality or within a framework of conventions, such as those governing games or good manners (Bucciarelli & Johnson-Laird, 2005; Bucciarelli, Khemlani, & Johnson-Laird, 2008). Modal auxiliary verbs, such as *must* and *may* in English and analogs in other Indo-European languages, are likewise ambiguous between factual and deontic interpretations. But, the two domains of possibilities and permissibilities cannot be amalgamated into one, because of a radical distinction between them. The failure of a factual necessity to occur renders its description false; whereas the violation of a deontic necessity does not render its description false—people do, alas, violate what is permissible.

Many theories propose a probabilistic meaning for causal relations and for conditionals (see Over, Chapter 18; and Oaksford & Chater, Chapter 19, both in this volume). One difficulty with this proposal is that the difference in meaning between *A causes B* and *A enables B*, which we outline presently, cannot be drawn within the probability calculus: both can yield high conditional probabilities of *B* given *A* (Johnson-Laird, 1999). Indeed, as the founder of Bayesian networks wrote, "Any causal premise that is cast in standard probabilistic expressions . . . can safely be discarded as inadequate" (Pearl, 2009, p. 334). He points out that *cause* has a deterministic meaning and that the probability calculus cannot distinguish between correlation and causation. It cannot express the sentence, "Mud does not cause rain" (Pearl, 2009, p. 412).

One general point on which all parties should agree is that even granted a deterministic meaning, probabilities can enter into causal assertions in various ways (Suppes, 1970). We can be uncertain about a causal relation:

> Eating protein will probably cause Evelyn to gain weight.

And our degree of belief in a deterministic proposition is best thought of as a probability (e.g., Ramsey, 1929/1990). Subtle aspects of experimental procedure can elicit judgments best modeled probabilistically—from references to "most" in the contents of problems to the use of response scales ranging from 0 to 100 (Rehder & Burnett, 2005; and Rehder, Chapters 20 and 21 in this volume). Evidence for a causal relation can itself be statistical. But, Pearl's ultimate wisdom remains: keep causation and statistical considerations separate; introduce special additional apparatus to represent causation within probabilistic systems. This division is now recognized in many different accounts of causation (see, e.g., Waldmann, 1996; and in this volume, Cheng & Lu, Chapter 5; Griffiths, Chapter 7; Lagnado & Gerstenberg, Chapter 29; and Meder & Mayrhofer, Chapter 23).

The consilience of the evidence about reasoning with quantifiers (e.g., Bucciarelli & Johnson-Laird, 1999), modal reasoning (e.g., Bell & Johnson-Laird, 1998), and deontic reasoning (e.g., Bucciarelli &

Johnson-Laird, 2005) implies that each of these domains has deterministic meanings (for a case against probabilistic causal meanings, see Khemlani, Barbey, & Johnson-Laird, 2014). Indeed, if *necessary* were probabilistic, it would allow for exceptions and be indistinguishable from *possible*.

The consensus about causation is that, by default, causes precede their effects. But, a billiard ball causes a simultaneous dent in the cushion on which it rests (Kant, 1781/1934), and the moon causes tides, which in Newtonian mechanics calls for instantaneous action at a distance. Philosophers have even speculated about causes following their effects in time. In daily life, however, the norm is that causes precede their effects, or at least do not occur after them. The sensible option is therefore that a cause does not follow its effect, and the same constraint applies to the prevention or enabling of events.

The correspondences in Table 10.1 imply that possibilities underlie the meanings of causal relations. It follows that there are several sorts of causal relation, and we will now enumerate them and explain why there cannot be any other sorts. An assertion about a specific cause and effect, such as

Eating protein will cause Evelyn to lose weight refers to a key possibility in which Evelyn eats protein and then loses weight. But, what happens if Evelyn does not eat protein? A weak interpretation of *cause* is that the assertion leaves open whether or not Evelyn will lose weight. It could result from some other cause, such as a regimen of rigorous exercise. The meaning of the assertion accordingly refers to a conjunction of three possibilities that are each in the required temporal order:

> Evelyn eats protein and Evelyn loses weight.
> Evelyn doesn't eat protein and Evelyn doesn't lose weight.
> Evelyn doesn't eat protein and Evelyn loses weight.

An assertion of prevention, such as *eating protein prevents Pat from losing weight*, is analogous but refers to the non-occurrence of the effect.

A cause can be the unique way of bringing about an effect. As far as we know, drinking alcohol is the only way to get drunk. Likewise, prevention can be unique. As far as we know, a diet including vitamin C is the only way to prevent scurvy. These stronger senses rule out alternative ways to cause or to prevent effects, and so both sorts of assertion refer only to two possibilities. A unique cause refers to a conjunction of just two possibilities: cause and effect; and no cause and no effect.

An enabling relation such as *eating protein enables Evelyn to lose weight* doesn't mean that protein necessarily leads Evelyn to lose weight: it may or may not happen, depending on the occurrence or non-occurrence of a cause, such as eating less of other foods. What happens if Evelyn does not eat protein? A weak interpretation is again that Evelyn may or may not lose weight. All four temporally ordered possibilities can therefore occur. They are equivalent to the weak enabling condition for a non-occurrence of an effect, *eating protein enables Evelyn* not *to lose weight*. A stronger and more frequent interpretation of the affirmative assertion is that eating protein is a unique and necessary condition for Evelyn to lose weight. The assertion's meaning is therefore a conjunction of these three temporally ordered possibilities:

> Evelyn eats protein and Evelyn loses weight.
> Evelyn eats protein and Evelyn doesn't lose weight.
> Evelyn doesn't eat protein and Evelyn doesn't lose weight.

The negative assertion, *eating protein enables Evelyn not to lose weight*, is analogous but refers instead to the non-occurrence of the effect.

We have now outlined seven distinct causal relations. Their meanings, according to the model theory, refer to different conjunctions of possibilities, which each embodies a temporal order. These meanings are summarized in Table 10.2, and they exhaust all possible causal relations. If *A* or *not-A* can occur, and *B* or *not-B* can occur, there are four possible contingencies of *A* and *B* and their respective negations, and there are 16 possible subsets of these contingencies. Four of these subsets consist in a single possibility, such as *A & not-B*, which is a categorical assertion of a conjunction. Four of these subsets consist in a conjunction of two possibilities, such as *A & B* and *not-A & B*, which corresponds to a categorical assertion, as in this case in which *B* holds whether or not *A* holds. And one subset is the empty set corresponding to a self-contradiction. The remaining seven of the 16 subsets of possibilities yield the distinct causal relations in Table 10.2: the strong and weak senses of *causes*, the strong and weak senses of *prevents*, the strong senses of *enables* and *enables_not*, and their identical weak senses. Granted that the meanings of causal relations refer to conjunctions of possibilities, no other causal relations can exist.

The conjunctions of possibilities in Table 10.2 are analogous to truth tables in sentential logic. Some critics have therefore argued that the model theory's

Table 10.2 The Core Meanings of the Seven Possible Causal Relations in Terms of the Conjunctions of Temporally Ordered Possibilities to Which They Refer

	The Seven Conjunctions of Possibilities Yielding Distinct Causal Relations						
	a b **not-a not-b** **not-a b**	**a b** **not-a not-b**	**a b** **a not-b** **not-a not-b** **not-a b**	**a b** **a not-b** **not-a not-b**	**a not-b** **not-a not-b** **not-a b**	**a not-b** **not-a b**	**a b** **a not-b** **not-a b**
A causes B	Weak	Strong					
A allows B			Weak	Strong			
A prevents B					Weak	Strong	
A allows not B			Weak				Strong

Note: Strong interpretations correspond to unique causes, preventers, and enablers; weak interpretations allow for others.

account of *A causes B* in its weak sense is equivalent to material implication in logic (e.g., Ali, Chater, & Oaksford, 2011; Kuhnmünch & Beller, 2005; Sloman, Barbey, & Hotaling, 2009). Other critics have made the same claim about the model theory's treatment of conditionals (see, e.g., Evans & Over, 2004). For readers unfamiliar with material implication, it is equivalent to *not-A or B, or both*, and so on this account *A causes B* is true provided that *A* is false or *B* is true. That's clearly wrong, and so, according to these critics, the model theory is therefore wrong, too. In fact, their argument is flawed, because it fails to distinguish between possibilities and truth values. Two truth values such as *true* and *false* are inconsistent with one another. In contrast, the possibility of *A* is entirely consistent with the possibility of *not-A*. Likewise, the conjunction of *possibly* (*A & B)* and *possibly* (*not-A & not-B*) is consistent. Indeed, the model theory postulates that the meaning of *A causes B* in its strong sense refers to this conjunction of possibilities. Its weak sense adds a third possibility to the conjunction: *not-A & B* (see Table 10.2). The mere falsity of *A* does not establish that this conjunction of three possibilities is true, nor does the mere truth of *B*. The same argument applies to conditionals, *if A then B* (cf. Johnson-Laird & Byrne, 2002). Hence, according to the model theory, the meaning of *A causes B* in its weak sense, and the meaning of a basic conditional, *if A then B*, both differ from the meaning of material implication in logic.

Counterfactual Conditionals and Causation

When the facts are known, it is appropriate to say, for instance, *eating protein caused Evelyn to lose weight*. But, as Hume (1748/1988, p. 115) recognized, it is equally appropriate to assert a "counterfactual" conditional: *If Evelyn hadn't eaten protein, then Evelyn might not have lost weight*. The appropriate counterfactual depends both on the nature of the causal relation and on the nature of the facts. Suppose, in contrast, that the causal relation still holds but the facts are that Evelyn didn't eat protein and didn't lose weight. A different counterfactual captures the causal relation:

> If Evelyn had eaten protein, then Evelyn would have lost weight.

The meanings of counterfactuals are straightforward. The assertion *A causes B* in its weak sense refers to a conjunction of three possibilities (Table 10.2), but it also means that *A and not-B* is impossible. Hence, given this sense and that the facts are *A and B*, the two other possibilities are counterfactual, where we define a *counterfactual* possibility as a contingency that was once possible but that didn't in fact happen (Byrne, 2005; Johnson-Laird & Byrne, 2002).

The meaning of a counterfactual conditional therefore depends on three principles. First, the negations of its two clauses refer to the facts but with a caveat, namely, when "still" occurs in the *then*-clause, the clause itself describes the facts. Second, its two clauses refer to the main counterfactual possibility. Third, any other counterfactual possibility depends on the modal auxiliary that occurs in its *then*-clause, which has the same force as its present tense meanings: *would* means the event necessarily occurs, and *might* and *could* mean the event possibly occurs. But, their negations differ in scope: *couldn't* means that the event is not possible, and *mightn't* means the event possibly does not occur (see Table 10.1). Other cognate modals express the same meanings.

As an example, consider again the counterfactual conditional: *If Evelyn had eaten protein, Evelyn would have lost weight*. The first principle determines the facts about Evelyn:

Didn't eat protein   Didn't lose weight [facts]

The second principle determines the main counterfactual possibility about her:

Ate protein   Lost weight [counterfactual possibility]

The third principle tells us that eating protein necessarily leads Evelyn to lose weight—there is no alternative in that case. The only counterfactual possibility that remains is therefore that Evelyn didn't eat protein but for some unknown reason nevertheless lost weight. A corollary for interpretation is that, as the preceding examples illustrate, models have to keep track of the status of a contingency—as a fact, a possibility, a counterfactual possibility, or an impossibility (Byrne, 2005; Johnson-Laird & Byrne, 2002).

Table 10.3 puts the three principles into practice. It presents the set of counterfactuals expressing the seven causal relations, in their strong and weak senses, depending on the facts of the matter. In general, the meaning of causal counterfactuals resists analysis unless one considers the possibilities to which they refer. The experimental evidence suggests that counterfactuals tend to elicit mental models of the facts and only the main counterfactual alternative to them (Byrne, 2005). It is tempting to identify causation and counterfactuals. But, like regular conditionals, counterfactuals need not express causal relations (e.g., *If the number hadn't been divisible by 2 without remainder then it wouldn't have been even*).

Are There Other Components in the Meanings of Causal Relations?

Experiments have shown that most participants list as possible the deterministic cases corresponding to the strong meanings of *causes, prevents, enables*, and *enables_not* (i.e., the participants tend to minimize the number of possibilities to which assertions refer; Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Goldvarg-Steingold, 2007). Is anything more at stake in the meanings of causal relations, that is, anything more than a conjunction of temporally ordered possibilities? The question concerns the meaning of causal assertions in everyday life as opposed to one's degree of belief in them or to how one might establish their truth (or falsity). For the latter, observational evidence can help, but experimentation is the final arbiter because one has to determine what happens given the putative cause, and what happens without it. If a conditional expresses the correct temporal relations between physical states, then it has a potential causal meaning, for example:

If the rooster crows, then the sun rises.

But, an adroit experiment will show that the preceding conditional is false, and hence that the rooster's crowing is not the cause of sunrise.

As skeptics sometimes claim, to say that *A causes B* at a minimum refers to an additional relation that goes beyond a universal succession of *A* followed by *B* (pace Hume, 1748/1988, p. 115). Table 10.2 presents such a relation. When *A* occurs, the only possibility is that *B* occurs, and so the relation is a necessary one (see Kant, 1781/1934). Other theorists have proposed various additional elements of meaning. But, before we consider them, we need to clarify the nature of our argument. In the model theory, the interpretation of any assertion is open to a process of *modulation* in which knowledge of meanings, referents, and context can eliminate models or add information to them over and above those conveyed by literal meanings (see, e.g., Johnson-Laird & Byrne, 2002). It can add, for instance, a temporal relation between the clauses of a conditional, such as *if he passed the exam, then he did study hard* (see, e.g., Juhos, Quelhas, & Johnson-Laird, 2012). The case against additional elements of meaning that we are going to make in what follows is against their addition to the core meanings of temporally ordered possibilities (see Table 10.2). But, modulation can incorporate these additional elements in the process of interpretation. For instance, Pearl's (2009) central assumption about causation is encapsulated in this principle: *Y is the cause of Z if we can change Z by manipulating Y*. But, we can manipulate a number so that it is or isn't divisible by 2 without remainder, and the manipulation changes whether or not the number is even. This condition, however, doesn't cause the number to be even. It necessitates its evenness. So, the criterion of manipulation lumps together mathematical necessity and causal necessity. It is also equivalent to a recursive definition in which "cause" is referred to in its own definition:

Y causes Z $=_{\text{def}}$ A manipulation of Y causes Z to change.

And recursive definitions need a condition that allows the recursion to bottom out, otherwise they lead to infinite loops (cf. Woodward, 2003). The model theory provides such a condition for any

Table 10.3 The Set of Counterfactual Conditionals Expressing Causal Relations Depending on the Facts of the Matter

			The Facts			
Causal Relation	**Strength**	**Meaning**	**not-a not-b**	**not-a b**	**a b**	**a not-b**
A causes B	Weak	a b not-a b not-a not-b	If A had happened, then B would have happened.	If A had happened, then B still would have happened.	If A hadn't happened, then B mightn't have happened.	The causal relation rules out this fact.
	Strong	a b not-a not-b	If A had happened, then B would have happened.	The causal relation rules out this fact.	If A hadn't happened, then B couldn't have happened.	The causal relation rules out this fact.
A prevents B	Weak	a not-b not-a b not-a not-b	If A had happened, then B still couldn't have happened.	If A had happened, then B couldn't have happened.	The causal relation rules out this fact.	If A hadn't happened, then B still mightn't have happened.
	Strong	a not-b not-a b	The causal relation rules out this fact.	If A had happened, then B couldn't have happened.	The causal relation rules out this fact.	If A hadn't happened, then B would have happened.
A allows B	Weak	a b a not-b not-a b not-a not-b	If A had happened, then B might have happened.	If A had happened, then B still might have happened.	If A hadn't happened, then B still might have happened.	If A hadn't happened, then B still mightn't have happened.
	Strong	a b a not-b not-a not-b	If A had happened, then B might have happened.	The causal relation rules out this fact.	If A hadn't happened, then B couldn't have happened.	If A hadn't happened, then B couldn't have happened.
A allows not-B.	Strong	a b a not-b not-a b	The causal relation rules out this fact.	If A had happened, then B still might have happened.	If A hadn't happened, then B still would have happened.	If A hadn't happened, then B would have happened.

Note: From left to right, the columns show the causal relation, its interpretation as weak or strong, the possibilities to which it refers, and the four sorts of fact. Each entry presents an appropriate counterfactual conditional for the causal relation and the facts; the other possibilities in the causal relation become counterfactual given the facts. The weak interpretation of *A allows not-B* is the same as for *A allows B*.

causal relation (see Table 10.2), and it allows modulation to incorporate manipulation into the interpretation of causation.

Some theorists have argued that part of the meaning of *A enables B to happen* is the existence of another factor that, when it holds, causes the effect (see Sloman et al., 2009). However, the truth of an enabling assertion doesn't establish the necessary existence of a cause, for example:

> The vapor enabled an explosion to occur, but luckily no cause occurred, and so there wasn't an explosion.

If part of the meaning of "enabled" was the existence of a cause, then the previous assertion would be self-contradictory. Modulation, however, can certainly add the existence of a cause.

When a rolling billiard ball collides with another stationary one, observers see the physical contact and perceive that one ball caused the other to move (Michotte, 1946/1963; White, Chapter 14 in this volume). Some theorists have accordingly argued that physical contact or contiguity is part of the meaning of causal assertions (e.g., Geminiani, Carassa, & Bara, 1996). But, consider these claims:

> Lax monetary policy enabled the explosion in credit to occur in the early 2000s.
> The explosion in credit caused the 2008 financial crash.

It would be supererogatory even to try to establish a chain of physical contacts here. Conversely, the meaning of causation hardly legislates for the falsity of action at a distance: the meaning of *cause* doesn't show that Newton's physics is false. Hence, the meanings of elements interrelated by causal claims can modulate interpretation to add physical contact.

Reasoners know that the wind exerts a force that can blow trees down, that electricity has the power to turn an engine, that workers (or robots) on a production belt are a means for making automobiles, that the mechanism in a radio converts electromagnetic waves into sounds, that explanatory principles account for inflation, and that scientific laws underlie the claim that the moon causes tides. Theorists have accordingly invoked as part of causal meanings: force (Wolff and Thorstad, Chapter 9 in this volume), power (Cheng & Lu, Chapter 5 in this volume), means of production (Harré & Madden, 1975), mechanisms and manipulations (Pearl, 2009), explanatory principles (Hart & Honoré, 1985), and scientific laws. Some of these factors, as Hume (1739/1978) argued, are impossible to define without referring to causation itself. Yet, a potent reason to infer a causal relation is relevant knowledge of any of these factors. If you are explaining the mechanism of a sewing machine to a child who persists in asking *how?*, there will come a point—perhaps after you've explained that a catch on the rotating bobbin pulls a loop of the upper thread around it—when you can no longer provide a mechanism. A mechanism is a hierarchy of causal relations (Miyake, 1986), and each relation may have its own underlying mechanism, but the recursion has to bottom out, or otherwise there is an infinite regress. There must be at least one causal relation for which there is no mechanism. The meaning of your final causal claim about the bobbin therefore need not refer to any mechanism. It follows that the core meanings of causal assertions need not refer to mechanisms.

In sum, knowledge of any of the factors that theorists invoke—force, power, means of production, interventions, explanatory principles, and mechanisms—can modulate the models based on core causal meanings, which do not embody them (see Khemlani, Barbey, & Johnson-Laird, 2014, for an integration of force and the model theory). Hence, the various putative elements of meaning beyond those of temporally ordered possibilities can play a role in the interpretation of causal claims, but they are not part of the core meanings of such claims, on pain of circularity or infinite regress.

Mental Models of Causal Assertions

The model theory distinguishes between the meanings of assertions, and the mental models of the possibilities to which these meanings refer. In the theory's computational implementation, the parsing of a sentence yields a representation of its meaning, and this representation is used to build or to update the mental models of the situation under description (Khemlani & Johnson-Laird, 2012). Each mental model represents a distinct possibility (i.e., it represents what is common to the different ways in which it may occur). The more models that individuals have to represent, the greater the load on working memory, and the more difficult reasoning becomes—a result that is highly robust and that, as far as we know, has no counterexamples in the experimental literature (see, e.g., Johnson-Laird, 2006). Indeed, once reasoners have to deal with more than one or two mental models, their task becomes very difficult, as experiments have shown (e.g., Bauer & Johnson-Laird, 1993; García-Madruga et al., 2001).

Table 10.2, which we presented earlier, shows the full set of possibilities to which the seven causal relations refer. In contrast, mental models are based on a principle of *truth*. They represent what is true in a possibility rather than what is false. The assertion

A will cause B to occur

has only two mental models:

a   b

  . . .

The first mental model represents the possibility in which *A* occurs no later than *B*, and the second mental model has no explicit content, as denoted by the ellipsis, but stands in for other possibilities in which *A* does not occur. Hence, mental models represent the salient possibility in which the antecedent event and its causal consequence both hold. An enabling assertion

A will enable B to occur

has exactly the same mental models as the preceding ones. And a preventive assertion:

A will prevent B from occurring

has the mental models:

a   not-b

  . . .

where *not-b* denotes B not occurring. The assertion *A will enable B not to occur* has these mental models as well. Of course, mental models are not letters or words, which we use here for convenience. They can be static spatiotemporal representations of the world, or kinematic simulations in which events follow one after the other (Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). Yet, just two sorts of sets of mental models represent all seven causal relations in Table 10.2, because mental models, which embody the principle of truth, do not distinguish between *causes* and *enables*, or between *prevents* and *enables not to occur*, in either their strong or weak senses.

For easy tasks, such as listing the possibilities to which an assertion refers, individuals can use the meaning of an assertion to flesh out mental models into *fully explicit* models of all the possibilities to which the assertion refers (as in Table 10.2). Even so, individuals begin their lists with the possibilities corresponding to mental models (Johnson-Laird & Goldvarg-Steingold, 2007). Only if they construct fully explicit models can they distinguish between causal and enabling assertions. A common misconception of the theory is that fully explicit models are used in all inferential tasks (pace Kuhnmünch & Beller, 2005). In fact, mental models are the foundation of intuitions and most inferences.

Deductions from Causal Relations

Reasoning starts with perceptions, descriptions, or memories. We refer to "the premises" in order to include any of these sources, and we distinguish among three principal sorts of reasoning: the *deduction* of valid conclusions; the *induction* of conclusions, such as generalizations that go beyond the given information; and a special sort of induction, known as *abduction*, which yields explanations. In what follows, we outline the model theory for each of these sorts of causal inference, starting with deduction.

Naïve individuals tend to reason based on mental models, and to draw conclusions that hold in the set of mental models of the premises. In logic, an inference is valid if the truth of a conclusion follows from the truth of the premises (Jeffrey, 1981). But, in the model theory, for a premise to imply that a conclusion is true, the premise has to imply each of the possibilities to which the conclusion refers. In logic, inferences of this sort are valid:

A.
Therefore, A or B, or both

because the disjunction is true if one or both of its clauses are true. But, the inference is unacceptable according to the model theory, because the premise doesn't imply the truth of one of the possibilities to which the conclusion refers: *not-A and B*. Analogous principles hold for inferring probabilities (see, e.g., Khemlani, Lotstein, & Johnson-Laird, 2015).

Individuals are able to deduce the consequences of causal chains. In one experiment (Goldvarg & Johnson-Laird, 2001), the first premise interrelated two events, *A* and *B*, using a causal relation, and the second premise likewise interrelated *B* and *C*. The participants' task was to say what, if anything, followed from each pair of premises. The experiment examined all 16 possible pairs of relations based on *causes, prevents, allows*, and *allows_not*. The contents of the problems were abstract entities familiar to the participants (e.g., *obedience causes motivation to increase*), but which could plausibly occur in any sort of problem. One sort was of the form

A causes B.
B prevents C.
What, if anything, follows?

The premises yield the mental models, as shown in a computer program implementing the theory:

```
a    b    not-c
       . . .
```

and, as the mental models predict, all participants in the experiment concluded that *A prevents C*. The same conclusion follows from fully explicit models representing all the possibilities to which the premises refer, though reasoners are most unlikely to consider all of them. In contrast, these premises

A prevents B.
B causes C.

yield the mental models:

```
a    not-b
          b        c
     . . .
```

All but one of the participants drew the conclusion that these mental models predict *A prevents C*. But, the six fully explicit models of these premises show that all four contingencies between *a* and *c*, and their respective negations, are possible. Hence, all that follows is that *A allows C* and *A allows_not C*. In general, the results bore out the predictions based on mental models of the premises, rather than fully explicit models (see Barbey & Wolff, 2007, for a replication). To what extent performance also reflects an "atmosphere" effect in which participants draw conclusions biased by the verbs in the premises calls for further research (see Sloman et al., 2009).

A crucial test for mental models is the occurrence of so-called "illusory" inferences. These are fallacious inferences that occur because mental models embody the principle of truth, and so they do not represent what is false. Illusions occur in all the domains of reasoning for which they have been tested, including reasoning based on disjunctions and conditionals, probabilistic reasoning, modal reasoning, reasoning about consistency, and quantified reasoning (for a review, see Johnson-Laird & Khemlani, 2013). They are a crucial test because no other theory of reasoning predicts them. Here is a typical instance of an illusory inference in causal reasoning:

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

The rubric to the problem is equivalent to an exclusive disjunction of the first two premises, and so, as the program shows, they yield the following mental models of Viv's state:

marry	relax	(first premise is true)
not-marry	relax	(second premise is true)

The third assertion eliminates the second model, and so it seems that Viv will relax. But, when one premise is true, the other premise is false. If the first premise is false, then Viv won't relax even though Viv marries Pat. If the second premise is false, then Viv won't relax even though Viv doesn't marry Pat. Either way, on a weak interpretation of *cause*, Viv won't relax. On a strong interpretation of *cause*, the premises imply nothing whatsoever about whether Viv will relax. It is therefore an illusion that Viv will relax. Nearly everyone in an experiment made illusory inferences, but they made correct inferences from control premises (Goldvarg & Johnson-Laird, 2001).

Causes Versus Enabling Conditions

Consider the first inference at the start of the chapter:

> Eating protein will cause Evelyn to gain weight.
> Evelyn will eat protein.
> Will Evelyn gain weight?

The mental models of Evelyn's state from the causal premise are as follows:

Eating protein	Gain weight
. . .	

The categorical premise that Evelyn will eat protein eliminates the second implicit model, and so it follows that Evelyn will gain weight. Only 2% of participants failed to draw this conclusion. The same conclusion follows from mental models when the first premise states an enabling condition:

> Eating protein will allow Evelyn to gain weight.

Only those individuals who flesh out their models of the enabling assertion to represent the alternative possibility:

Eating protein	Not gain weight

will infer that Evelyn may or may not gain weight. Many people (32%), but not all, are able to envisage this alternative possibility in which Evelyn eats protein but does not gain weight (Goldvarg & Johnson-Laird, 2001). A similar study used "when" instead of "if" (e.g., *when magnetism occurs, magnetism causes ionization*), and yielded similar results (Sloman et al., 2009).

Readers should try to identify the cause and the enabler in the following scenario:

> If you take the drug *Coldgon*, then, given that you stay in bed, you will recover from the common cold in one day. However, if you don't stay in bed, then you won't recover from the common cold in one day, even if you take this drug.

Reasoners are most unlikely to envisage all the possibilities to which this description refers, but they should be able to think of the most salient ones, which are represented in mental models:

Take drug	stay in bed	recover
Take drug	not stay in bed	not recover
	. . .	

Reasoners should therefore realize that staying in bed is the catalyst that enables the drug to cause the one-day cure.

An experiment compared scenarios such as the preceding with those in which the causal roles were swapped around, for example:

> If you stay in bed, then given that you take the drug *Coldgon*, you will recover from the common cold in one day. However, if you don't take this drug, then you won't recover from the common cold in one day, even if you stay in bed.

Eight scenarios ranged over various domains—physical, physiological, mechanical, socioeconomic, and psychological—and counterbalanced the order of mention of cause and enabler. The participants read just one version of each scenario. They identified the predicted causes and enablers on 85% of trials, and each of them did so more often than not, and each scenario bore out the difference (Goldvarg & Johnson-Laird, 2001). Cheng and Novick (1991) showed that their participants could distinguish between causes and enablers in similar sorts of everyday scenarios, but, for reasons pertaining to their probabilistic theory, their scenarios described enabling conditions that were constant throughout the events in the scenario, such as the presence of gravity, whereas causes were not constant, such as a boy pushing a girl. But, the present study swapped the roles of causes and enablers from one scenario to another, and neither was constant. In the preceding example, a person might or might not stay in bed, and might or might not take *Coldgon*. Hence, constancy is not crucial for individuals to identify an enabler, and inconstancy is not crucial for them to identify a cause.

Linguistic cues, such as "if" versus "given that," might have signaled the distinction between causes and enablers (Kuhnmünch & Beller, 2005). But, when these cues were rigorously counterbalanced or eliminated altogether, individuals still reliably distinguished between causes and enablers (see Frosch & Byrne, 2006). Likewise, when scenarios contained only a cause, or only an enabler, and used the same linguistic cue to introduce both, individuals still reliably identified them (Frosch, Johnson-Laird, & Cowley, 2007). This follow-up study contrasted causes and enablers within six scenarios about wrongdoing, such as:

> Mary threw a lighted cigarette into a bush. Just as the cigarette was going out, Laura deliberately threw petrol on it. The resulting fire burnt down her neighbor's house.

The participants again distinguished between those individuals whose actions caused criminal events, such as Laura, and those who enabled them to occur, such as Mary. Moreover, they judged causers to be more responsible than enablers, liable for longer prison sentences, and liable to pay greater damages. It is regrettable that neither English nor American law makes the distinction between causers and enablers (Johnson-Laird, 1999)—a legacy of Mill's (1874) views, as embodied in judicial theory (see Hart & Honoré, 1985; Lagnado & Gerstenberg, Chapter 29 in this volume).

According to the model theory, a single instance of *A* and *not-B* refutes *A causes B* in either its strong or its weak sense (see Table 10.1). The refutation of an enabling relation is more problematic. In its strong sense, it is necessary to show that the effect can occur in the *absence* of the enabler; in its weak sense, only temporal order is at issue. A further difficulty is that both causes and enablers have the same mental models. Frosch and Johnson-Laird (2011) invited their participants to select which sort of evidence, *A and not-B* or *not-A and B*, provides more decisive evidence against each of eight causal and eight enabling assertions, such as

> Regular exercise of this sort causes a person to build muscle.

and:

> Regular exercise of this sort enables a person to build muscle.

Every single participant chose *A and not-B* more often than *not-A and B*, but, as the theory predicts, they chose *not-A and B* reliably more often as a refutation for *enables* (25% of occasions) than for *causes* (10%), even though it refutes the strong meaning of *causes* too. They had an analogous bias in judging whether a single observation sufficed to refute a claim.

The general conclusion from these studies is that individuals distinguish between causes and enabling conditions in deductions, in inferring the role of actors in scenarios, and in assessing refutations of causal claims. In each of these cases, the distinction follows from the model theory's deterministic account of the meanings of *causes* and *enables* (see Table 10.2). It is not at all clear that theories that do not base the distinction between these relations on different sets of possibilities can explain these results (cf. Ali et al., 2011; Sloman et al., 2009). The model theory makes further predictions about ternary causal relations, such as:

> Staying in bed enables Coldgon to cause you to recover in a day.

Ternary relations of the sort *A enables B to cause C* are distinct from a conjunction of *A enables C* and *B causes C*, and so challenge the representational power of probabilistic networks, whose binary links have no natural way to represent them.

Inductions of Causal Relations

Learning is often a matter of inducing causal relations from observations of the relative frequencies in the covariations of contingencies (see, e.g., Perales & Shanks, 2007; Lu et al., 2008; and see Perales, Catena, Cándido, & Maldonado, Chapter 3 in this volume). Conditioning and reinforcement learning also concern causation (see Le Pelley, Griffiths, & Beesley, Chapter 2 in this volume). Probabilistic inductions at one level can feed into those at a higher or more abstract level in a hierarchical Bayesian network (e.g., Tenenbaum, Griffiths, & Kemp, 2006). Once the network is established, it can assign values to conditional probabilities that interrelate variables at one level or another (see Griffiths, Chapter 7 in this volume). Yet, causal relations are deterministic, and it is our ignorance and uncertainty that force us to treat them as probabilistic (Pearl, 2009, Ch. 1).

Inductive reasoning can yield deterministic causal relations. For instance, Robert Boyle carried out experiments in which he varied the pressure of a gas, such as air, and discovered that the pressure of a given quantity of gas at a fixed temperature is inversely proportional to the volume it occupies. This well-known law is deterministic, and so it is ironic that its ultimate explanation is the statistical kinetic theory of gases. Inductions of causal relations are also the intellectual backbone of medicine (see Lombrozo & Vasilyeva, Chapter 22 in this volume). A typical example is the discovery of the pathology and communication of cholera. When it first arrived in Britain in the nineteenth century, doctors induced that they were dealing with a single disease with a single pathology, not a set of alternative diseases, because of its common symptoms and prognosis. The induction reflected the heuristic that similar causes have similar effects (Hume, 1748/1988, p. 80). How the disease was communicated from one person to another was more mysterious. The arrival of an infected person in a particular place often led to an outbreak there. Doctors induced that the illness was either infectious or contagious. Sometimes, however, the disease could leap distances of several miles. Doctors induced that it could be conveyed through the air. The prevalence of cholera in slums with their stinking air seemed to corroborate this "miasmal" hypothesis. The doctor who discovered the true mode of the disease's communication, John Snow, was an expert on anesthesia, and his familiarity with Boyle's law and the other gas laws enabled him to infer the impossibility of the miasmal account (Johnson-Laird, 2006, Ch. 27). His bias toward parsimony led him to induce a common cause. Infected individuals could transmit some sort of particle of the disease, even perhaps an animalcule, to others who were in contact with their fecal matter. If these particles got into the water supply, they could then be transmitted over larger distances. Snow constructed a causal chain that explained both the pathology of the disease and its communication. And he made many observations that corroborated the idea. He then turned to a series of brilliant natural experiments. He found streets in London supplied with water from two companies, one that drew its water from the Thames downstream from the main sewer outflows and one that drew it upstream from them. As he predicted, 20 times more deaths from the disease occurred in those households supplied from the downstream company than in those supplied from the upstream company. Frequencies accordingly entered into his tests of the theory, but not into its mechanism.

As the preceding account suggests, inductions are easy. There was no shortage of hypotheses about what caused cholera to spread from person to person: infection, contagion, miasma. Knowledge can lead to an induction from a single observation—a claim supported by considerable evidence (see, e.g., Johnson-Laird, 2006, Ch. 13; White, 2014). One source of such inferences is knowledge of a potential mechanism (see Johnson & Ahn, Chapter 8 in this volume), which itself may take the form of a model—a point that we elucidate later. Likewise, "magical" thinking, which underlies common beliefs in all societies, is a result of induction and the Humean heuristic that similar causes have similar effects (Johnson-Laird, 2006, Ch. 5). The hard task is to use observation and experiment to eliminate erroneous inductions. It is simple to refute the strong claim:

> The rooster's crowing causes the sun to rise.

The observation that the sun also rises when the rooster does not crow suffices. The weaker claim that the rooster's crowing suffices for the sun to rise but other putative causes exist too, calls for an experiment in which the rooster is made to crow, say, at midnight. General causal claims, however, are notoriously difficult to refute. That is the business of experimental sciences.

At the center of the model theory is the idea that the process of interpretation builds models. In induction, modulation increases information. One way in which it does so is to add knowledge to a model. For instance, it sets up causal relations between events in the model in so-called bridging inferences (Clark, 1975), that is, inferences that build a bridge from an assertion to its appropriate antecedent. An experiment showed the potency of such inferences (Khemlani & Johnson-Laird, 2015). In one condition, the experiment presented sets of assertions for which the participants could induce a causal chain, for example:

David put a book on the shelf.
The shelf collapsed.
The vase broke.

In a control condition, the experiment presented sets of assertions for which the participants could not readily infer a causal chain, for example:

Robert heard a creak in the hall closet.
The faucet dripped.
The lawn sprinklers started.

When a subsequent assertion contradicted the first assertion in a set, the consequences were quite different between the two conditions. In the causal condition, the contradictory assertion:

David didn't put a book on the shelf

led to a decline in the participants' ratings of the strength of their beliefs in each of the subsequent assertions: only 30% of them now believed that the vase broke. In the control condition, the contradictory assertion:

Robert did not hear a creak in the hall closet

had no reliable effect on the participants' strength of belief in the subsequent assertions. All of them continued to believe that the lawn sprinklers started. This difference in the propagation of doubt is attributable to the causal interpretation of the first sort of scenario, and the near impossibility of a causal interpretation for the second scenario.

The model theory assumes that knowledge and beliefs can themselves be represented in models, and so the essence of modulation, which occurs in bridging inferences, is to make a conjunction of two sets of models: one set represents the possibilities to which assertions refer, and the other set represents possibilities in knowledge. A simple example of the process occurs when knowledge modulates the core interpretation of conditionals by blocking the construction of models (see Johnson-Laird & Byrne, 2002). A slightly different case is likely to have occurred in Snow's thinking about cholera. The received view was that cholera was transmitted in various ways—by infection or contagion when there was physical contact with a victim or by a miasma in other cases:

Physical contact	contagion	transmission
Physical contact	infection	transmission
No physical contact	miasma	transmission

Snow's knowledge of the gas laws yielded two negative cases:

Physical contact	infection	no transmission
No physical contact	miasma	no transmission

In deductive reasoning, the conjunction of two inconsistent models, such as the models in these sets concerning infection and miasma, results in the empty model (akin to the empty set), which represents contradictions. But, when one model is based on knowledge, it takes precedence over a model based on premises (Johnson-Laird, Girotto, & Legrenzi, 2004). Precedence in the conjunction of the two sets of models above yields models in which no transmission occurs by infection or miasma, and only one mechanism transmits the disease:

Physical contact contagion transmission

Snow knew, however, that the disease could also be transmitted over distances. Induction could not yield its mode of transmission. An explanation called for a more powerful sort of inference, abduction, to which we know turn. In Snow's case, it led to an explanation based on the transmission of "particles" of the disease through the water supply. This idea was never accepted in his lifetime, but he had inferred the disease's mode of transmission without any knowledge of germs, and his "particles" were later identified as the bacterium *Vibrio cholerae*.

Abductions of Causal Explanations

A fundamental aspect of human reasoning is abduction: the creation of explanations. Like inductions, they increase information, but unlike inductions, they also introduce new concepts that are not part of their premises. Abduction, in turn, depends on understanding, and according to the model theory, if you understand, say, inflation, the way a computer works, DNA, or a divorce, then you have a mental model of them. It may be rich in detail or simple—much as a clock functions as a model of the earth's rotation (Johnson-Laird, 1983, p. 2). Abductions usually concern causation. Investigators have studied them in applied domains, such as medical diagnosis (see Meder & Mayrhofer, Chapter 23 in this volume). To illustrate the role of models in abductions, we consider two cases: the resolution of causal inconsistencies and the reverse engineering of electrical circuits.

Explanations of Inconsistencies

When you are surprised in daily life, something has usually happened contrary to your beliefs or their consequences. You believe that a friend has gone to fetch the car to pick you up, and that if so, your friend should be back in no more than five minutes. When your friend fails to return within 20 minutes, this fact refutes the consequences of your beliefs. A large literature exists in philosophy and artificial intelligence on how you then ought to modify or withdraw your conclusion and revise your beliefs—a process that is known as "non-monotonic" or "defeasible" reasoning (see, e.g., Brewka, Dix, & Konolige, 1997). What is more important in daily life, however, is to explain the origins of the inconsistency—why your friend hasn't returned—because such an explanation is vital to your decision about what to do. But, where do explanations come from?

The answer has to be from knowledge (see Lombrozo and Vasilyeva, Chapter 22 in this volume). Some explanations are recalled, but many are novel: they are created from knowledge of causal relations, that is, models in long-term memory of what causes, enables, and prevents various events. This knowledge can be used to construct a simulation of a causal chain. A computer program implements the process (see Johnson-Laird et al., 2004). To understand it, readers should try to answer the following question:

> If someone pulled the trigger, then the pistol fired.
> Someone pulled the trigger. But the pistol did not fire.
> Why not?

The program constructs a model of the possibility described in the first two assertions:

trigger pulled   pistol fired

But, as it detects, the third assertion is inconsistent with this model. The conditional expresses a useful idealization, and the program builds a model of the facts, and its counterfactual possibilities (cf. Pearl, 2009, Ch. 7):

trigger pulled   not(pistol fires)   [the facts]
trigger pulled   pistol fires   [counterfactual possibilities]
. . .

The program has a knowledge base consisting of fully explicit models of several ways in which a pistol may fail to fire (i.e., preventive conditions such as *something jammed the pistol, there were no bullets in the pistol, its safety catch was on*). The model of the preceding facts triggers one such model, which the program chooses arbitrarily if the evidence leaves open more than one option, and the model takes precedence over the facts to create a possibility, for example:

not(bullets in pistol) trigger pulled not(pistol fires)

The new proposition, not(bullets in pistol), elicits a cause from another set of models in the knowledge base, for example, *if a person empties the bullets from the pistol, then there are no bullets in the pistol*. In this way, the program constructs a novel causal chain. The resulting possibility explains the inconsistency: a person intervened to empty the pistol of bullets. And the counterfactual possibilities yield the claim:

> If the person hadn't emptied the pistol, then it would have had bullets, and it would have fired.

The fact that the pistol did not fire has been used to create an explanation from knowledge, which in turn transforms the generalization into a counterfactual claim. Intervention is sometimes said to demand its own logic (Sloman, 2005, p. 82; see also Glymour, Spirtes, & Scheines, 2000; Pearl, 2000), but the standard machinery of modulation copes with precedence given to models based on knowledge in case of inconsistencies (Johnson-Laird, 2006, p. 313). This same machinery handles the "non-monotonic" withdrawal of conclusions and modification of beliefs.

The theory predicts that explanations consisting of a causal chain, such as a cause and effect, should be highly plausible. They should be rated as more probable than explanations consisting of the cause alone, or the effect alone. An experiment corroborated this prediction in a study of 20 different inconsistent scenarios (see Johnson-Laird et al., 2004). The participants rated the probability of various putative explanations, and they tended to rank the cause-and-effect explanations as the most probable. Hence, individuals do not always accommodate a new fact with a minimal change to their existing beliefs (see also Walsh & Johnson-Laird, 2009). The acceptance of a conjunction of a cause and effect calls for a greater change than the acceptance of just the cause or the effect. Another study showed that individuals also rate explanations as more probable than minimal revisions to either the conditional or the categorical premise to restore consistency (Khemlani & Johnson-Laird, 2011). Contrary to a common view, which William James (1907, p. 59) first propounded, the most plausible explanation is not always minimal.

Causal Abduction and Reverse Engineering

The lighting in the halls of many houses has an ingenious causal control, using a switch on the ground floor and a switch on the upper floor. If the lights are on, either switch can turn them off; if the lights are off, either switch can turn them on. The reader should jot down a diagram of the wiring required for this happy arrangement. The problem is an instance of "reverse engineering": to abduce a causal mechanism underlying a system of a known functionality. A study of the reverse engineering of such circuits revealed a useful distinction between two levels of knowledge—global and local (Lee & Johnson-Laird, 2013). A simple switch closes to make a circuit and opens to break the circuit, but a more complicated switch is needed for the lighting problem. It has two positions, and in one position it closes one circuit, and in the other position it both breaks this circuit and closes a separate circuit. It can also be used merely to make or break a single circuit. Figure 10.1 is a diagram showing the two positions of such a switch.

An experimental study examined naïve individuals' ability to reverse-engineer three sorts of circuits containing two switches: a circuit in which the light comes on only when both switches are on (a conjunction), a circuit in which the light comes on when one or both switches are on (an inclusive disjunction), and the hall circuit in which the light comes on when one switch or else the other is on (an exclusive disjunction). Each problem was presented in a table showing the four possible joint positions of the two switches and whether the light was on or off in each case. The participants knew nothing about wiring circuits in series or in parallel, but the experimenter described how the switch in Figure 10.1 worked, and explained that electricity "flows" when a circuit is completed from one terminal on a battery (or power source) to the other. The task was to design correct circuits for the three sorts of problem in which, as the participants knew, there was already one direct connection from the power to the light.

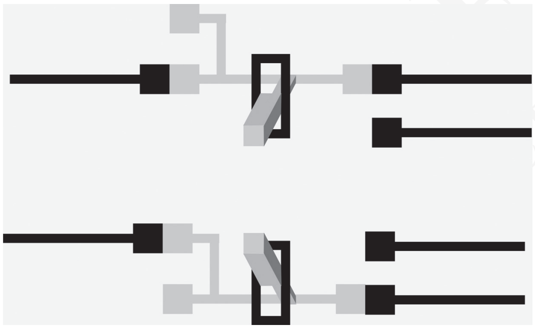

Figure 10.1 A diagram of the two different positions of a switch making or breaking two alternative circuits.

Figure 10.2 shows simple solutions for the three circuits. The experimenter video-recorded how people wired up actual switches, and in other experiments how they drew a succession of circuit diagrams to try to solve a problem, or else diagrams of pipes and faucets for three isomorphic problems about the flow of water to turn a turbine. Most participants focused either on getting the circuit to deliver one correct output at a time (i.e., a single causal possibility), taking into account the positions of both switches, but a few tried to get one switch at a time to work correctly. The difficulty of reverse engineering should depend on the number of possible configurations, determined by the number of variable components (the switches), the number of their settings that yield positive outputs (the light comes on), and the interdependence of the components in controlling the outputs. Only the exclusive disjunction depends on the joint positions of the two switches both to turn the light on and to turn it off. The results showed that both the number of settings with positive outcomes

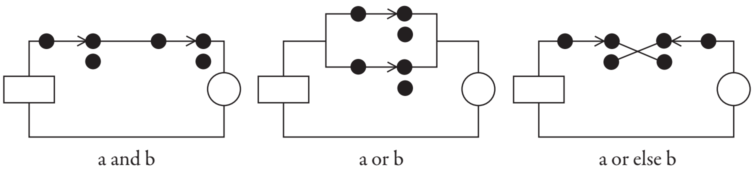

Figure 10.2 Minimal circuits for *a and b, a or b*, and *a or else b*. The rectangle and the circle represent the battery and the bulb, respectively, and each black dot represents a terminal of a switch of the sort shown in Figure 10.1. The light is on in the circuits shown for *and* and *or*, but off in the circuit for *or else*.

Table 10.4 Mean Number of Times in 20,000 Trials in Which the Program Reverse Engineered *and, or*, and *or else* Switch Circuits, Depending on the Constraints on Its Generative Process

Type of Problem	Local Constraints	Global Constraints	Local and Global Constraints
a and b	41	3316	4881
a or b, or both	95	619	1359
a or else b, but not both	0	1	6

Source: Based on Lee & Johnson-Laird (2013).

and interdependence increased the difficulty of the task, and so conjunctions were easier than disjunctions, which in turn were easier than exclusive disjunctions.

A computer program implementing abduction solves the problems. It explores models of the circuits by making arbitrary wirings under the control of local or global knowledge, or both (see Lee & Johnson-Laird, 2013). The program has access to five local constraints, which govern individual components in the model: a single wire should not connect a terminal on a switch or light to itself or to any other terminal on the same component, it should not be duplicated or be the converse of an existing wire, and it should not connect the power directly to the light (because of the pre-existing connection between them). The program also had access to six global constraints, which govern the model as a whole: the circuit should yield the given output for each switch position, it should contain at least six wires, it should connect the battery to at least one switch, it should connect the light to at least one switch, and each switch should have a wire on its single terminal and another wire on at least one of its double terminals. Table 10.4 shows the results of 20,000 computer simulations in each of several conditions depending on the constraints governing its performance. As it shows, global constraints are more efficient than local constraints, but the two combined increase performance to a level comparable to that of the human participants. Like them, the program almost always fails with an exclusive disjunction. Yet, in a rare instance, it did discover a novel circuit for the exclusive disjunction, which Figure 10.3 presents, and which is a two-dimensional solution unlike the one in Figure 10.2 in which one wire crosses over another.

The program uses abduction to produce circuits, and deduction to test their causal consequences. This procedure is common in the creation of explanations. Reasoners also used both abduction and deduction to create informal algorithms for rearranging the order of cars in a train on a track that included a siding (Khemlani et al., 2013).

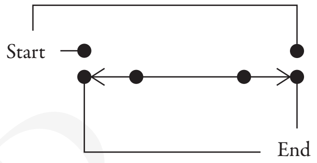

Figure 10.3 A novel two-dimensional circuit for *a or else b* that the computer program discovered. *Start* corresponds to the battery, and *End* to the bulb, with their other two terminals connected directly. Either position in which one switch is up and the other is down causes the current to flow.

Conclusions

The theory of mental models accounts for the meaning of causal relations, their mental representation, and reasoning from them. It proposes meanings that are deterministic. *A causes B* means that given *A* the occurrence of B is necessary; *A allows B* means that given *A* the occurrence of *B* is possible; and *A prevents B* means that given *A* the occurrence of *B* is impossible. If these relations were probabilistic, then *necessity* would tolerate exceptions and be equivalent to *possibility*, and *causes* would be equivalent to *enables*. The consilience of evidence corroborates deterministic meanings. For instance, the inference at the head of this chapter:

> Eating protein will cause Evelyn to gain weight.
> Evelyn will eat protein.
> Will Evelyn gain weight?

elicited an almost unanimous response of "yes," which is incompatible with a probabilistic interpretation of causation. Likewise, other studies, which we have described in this chapter, bore out deterministic predictions (e.g., individuals treat a single counterexample as refuting causation). Of course, causal claims can be explicitly probabilistic, as in this example from Suppes (1970, p. 7):

> Due to your own laziness, you will in all likelihood fail this course.

Likewise, generic assertions, whether they are about causes:

> Asbestos causes mesothelioma

or not:

> Asbestos is in ceiling tiles

can tolerate exceptions (Khemlani, Barbey, & Johnson-Laird, 2014). But, just as inferences differ between conditionals with and without probabilistic qualifications (Goodwin, 2014), they are likely to do so for causal relations. Skeptics may say that fully explicit models of real causal relations should contain myriad hidden variables (preventers, enablers, alternative causes) in complex structures, and so they ought to be much more complex than the models in this chapter. We agree. And we refer readers to the models of causal relations in real legal cases (Johnson-Laird, 1999). They soon overtake the reasoning ability of naïve individuals—just as comparable estimates of probabilities do (Khemlani, Lotstein, & Johnson-Laird, 2015). Other theories of causation postulate other elements in its meaning, such as forces, mechanisms, and interventions. The model theory accommodates these elements, but not in the meanings of causal relations. They are incorporated into the model as a result of modulation—the process that integrates models of knowledge and models of discourse, with the former taking precedence over the latter in case of conflicts. Modulation in the process of interpretation incorporates knowledge of these other elements into models.

The model theory is sometimes wrongly classified as concerned solely with binary truth values. In fact, as this chapter has aimed to show, it is rooted in possibilities. They readily extend to yield extensional probabilities based on proportions of possibilities or their frequencies (Johnson-Laird et al., 1999), and non-extensional probabilities based on evidence (Khemlani, Lotstein, & Johnson-Laird, 2015). And possibilities yield seven, and only seven, distinct causal relations: strong and weak meanings of *causes, prevents*, and *allows* and *allows_not*, with the weak meanings of the latter two relations being identical. The only proviso in their meanings is that their antecedents cannot occur after their effects. The mental models of these relations represent the situations they refer to, and they are identical for *causes* and *allows* unless individuals flesh out their models with explicit models of other possibilities. This identity is reflected in experimental results—individuals often infer an effect from the statement of either a cause or an enabling condition. It is also reflected in a long tradition that the difference between the two relations lies, not in their meanings, but in other factors such as normality, constancy, and relevance—a tradition that still lives in common law. Mental models suffice for many inferences. The principle that they represent only what is possible given the truth of the premises yields systematic illusory inferences. But, only fully explicit models elucidate ternary relations of the sort:

> Staying in bed enables *Coldgon* to cause your recovery from a cold in one day.

Such relations cannot be reduced to a conjunction of causing and enabling.

Inductions of causal relations rely on knowledge, especially those inductions—known as *abductions*—that yield explanations. In daily life, abductions rely on knowledge of causes and their effects. The model theory explains the process in terms of modulation, which also explains how individuals cope with inconsistencies: models of knowledge take precedence over other sorts of model. Hence, abduction leads to explanations that resolve inconsistencies, to the nonmonotonic withdrawal of conclusions, and to the revision of beliefs. We have illustrated this role of models and their role in reverse engineering. The latter sort of abduction depends on both knowledge of local constraints governing the components in a model, and knowledge of global constraints on models as a whole.

In sum, causal relations refer to conjunctions of temporally ordered possibilities. Human reasoners envisage these possibilities in mental models, which highlight only the salient cases. They use their knowledge to modulate these representations, and they infer the consequences of the resulting models.

Future Directions

Psychological research into causation is burgeoning, and so we describe here three directions of research most pertinent to the model theory.

1. Reasoning in certain domains depends on the use of a kinematic model that unfolds in time to represent a succession of events (Khemlani et al., 2013). Such mental simulations should also underlie causal reasoning, but the hypothesis has yet to be tested in experiments.

2. “The reason for Viv divorcing Pat was infidelity.” Are reasons merely causes of another sort? Many philosophers have supposed so (see, e.g., Dretske, 1989), but to the best of our knowledge no empirical research has examined this idea. Perhaps some reasons are causes of intentions rather than direct causes of actions (Miller & Johnson-Laird, 1976).

3. The treatment of causal relations as probabilistic has been very fruitful. But, the evidence that we have considered supports deterministic meanings for causation, and the use of probabilities as a way to treat human ignorance—a Bayesian approach that we have defended for the probabilities of unique events (Khemlani, Lotstein, & Johnson-Laird, 2015). A major task for the field is to reach a consensus about how to incorporate probabilities into causal reasoning in a way that distinguishes between causes and enabling conditions.

References

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*, 403–418.

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

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*, 372–378.

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

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning.* Stanford, CA: CSLI.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*, 247–303.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology*, *50*, 159–193.

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision Making*, *3*, 121–139.

Byrne, R. M. J. (2005). *The rational imagination*. Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

Clark, H. H. (1975). Bridging. In R. C. Schank & B. L. Nash-Webber (Eds.), *Theoretical issues in natural language processing* (pp. 169–174). New York: Association for Computing Machinery.

Craik, K. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.

Dretske, F. (1989). Reasons and causes. *Philosophical Perspectives*, *3*, 1–15.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.

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

Frosch, C. A., & Byrne, R. M. J. (2006). Priming causal conditionals. In R. Sun (Ed.). *Proceedings of 28th annual conference of the Cognitive Science Society* (p. 2485). Mahwah, NJ: Lawrence Erlbaum Associates.

Frosch, C. A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*, 280–291.

Frosch, C. A., Johnson-Laird, P. N., & Cowley, M. (2007). It’s not my fault, your Honor, I’m only the enabler. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual meeting of the Cognitive Science Society* (p. 1755). Mahwah, NJ: Lawrence Erlbaum Associates.

García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, F., & Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology*, *54A*, 613–632.

Geminiani, G. C., Carassa, A., & Bara, B. G. (1996). Causality by contact. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 275–303). Hove, East Sussex: Psychology Press.

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

Goodwin, G. P. (2014). Is the basic conditional probabilistic? *Journal of Experimental Psychology: General*, *143*, 1214–1241.

Harré, R., and Madden, E. H. (1975). *Causal powers*. Oxford: Blackwell.

Hart, H. L. A., and Honoré, A. M. (1959/1985). *Causation in the law* (2nd ed.). Oxford: Clarendon Press.

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

Hume, D. (1739/1978). *A treatise on human nature*. (L. A. Selby-Bigge, Ed.) (2nd ed.). Oxford: Oxford University Press.

Hume, D. (1748/1988). *An enquiry concerning human understanding*. (A. Flew, Ed.) La Salle, IL: Open Court. James, W. (1907). *Pragmatism*. New York: Longmans, Green.

Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York: McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1999). Causation, mental models, and the law. *Brooklyn Law Review*, *65*, 67–103.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford: Oxford University Press.

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

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640–661.

Johnson-Laird, P. N., & Goldvarg-Steingold, E. (2007). Models of cause and effect. In W. Schaeken et al. (Eds.), *The mental models theory of reasoning* (pp. 167–189). Mahwah, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., & Khemlani, S. S. (2013). Toward a unified theory of reasoning. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 59, pp. 1–42). New York: Elsevier.

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

Johnson-Laird, P. N., & Quinn, J. G. (1976). To define true meaning. *Nature*, *264*, 635–636.

Juhos, C., Quelhas, C., & Johnson-Laird, P. N. (2012). Temporal and spatial relations in sentential reasoning. *Cognition*, *122*, 393–404.

Kant, I. (1781/1934). *Critique of pure reason*. New York: Dutton.

Khemlani, S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning: Mental computations, and brain mechanisms. *Frontiers in Human Neuroscience*, *8*, 1–15.

Khemlani, S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*, 2276–2288.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument and Computation*, *4*, 4–20.

Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2015). Naive probability: Model-based estimates of unique events. *Cognitive Science*, *39*, 1216–1258.

Khemlani, S., & Johnson-Laird, P. N. (2015). Domino effects in causal contradictions. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, *110*, 16766–16771.

Kuhnmünch, G., & Beller, S. (2005). Causes and enabling conditions: Through mental models of linguistic cues? *Cognitive Science*, *29*, 1077–1090.

Lee, N. Y. L, & Johnson-Laird, P. N. (2013). A theory of reverse engineering and its application to Boolean systems. *Journal of Cognitive Psychology*, *25*, 365–389.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *Annals of Statistics*, *9*, 45–58.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955–984.

Mackie, J. L. (1980). *The cement of the universe: A study in causation* (2nd ed.). Oxford: Oxford University Press

Michotte, A. (1946/1963). *The perception of causality*. London: Methuen. Mill, J. S. (1874). *A system of logic, ratiocinative and inductive* (8th ed.). New York: Harper.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, *10*, 151–177.

Pearl, J. (2009). *Causality* (2nd ed.). New York: Cambridge University Press.

Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce*. (C. Hartshorne, P. Weiss, & A. Burks, Eds.). Cambridge, MA: Harvard University Press.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*, 577–596.

Ramsey, F. P. (1929/1990). Probability and partial belief. In D. H. Mellor (Ed.), *F. P. Ramsey: Philosophical papers*. Cambridge: Cambridge University Press.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314.

Russell, B. A. W. (1912–1913). On the notion of cause. *Proceedings of the Aristotelian Society*, *13*, 1–26.

Salsburg, D. (2001). *The lady tasting tea.* New York: W. H. Freeman.

Sloman, S. (2005). *Causal models*. New York: Oxford University Press.

Sloman, S., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of *cause, enable*, and *prevent*. *Cognitive Science*, *33*, 21–50.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Waldmann, M. R. (1996). Knowledge-based causal induction. *Psychology of Learning and Motivation*, *34*, 47–88.

Walsh, C. R., & Johnson-Laird, P. N. (2009). Changing your mind. *Memory & Cognition*, *37*, 624–631.

White, P. A. (2014). Singular cues to causality and their use in human causal judgment. *Cognitive Science*, *38*, 38–75.

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