19

MENTAL MODELS AND REASONING

Philip N. Johnson-Laird, Geoffrey P. Goodwin, and Sangeet S. Khemlani

Introduction

The theory of mental models has a long history going back to the logic diagrams of C.S. Peirce in the nineteenth century. But it was the psychologist and physiologist Kenneth Craik who first introduced mental models into psychology. Individuals build a model of the world in their minds, so that they can simulate future events and thereby make prescient decisions (Craik, 1943). But reasoning, he thought, depends on verbal rules. He died tragically young, and had no chance to test these ideas. The current "model" theory began with the hypothesis that reasoning too depends on simulations using mental models (Johnson-Laird, 1980).

Reasoning is a systematic process that starts with semantic information in a set of premises, and transfers it to a conclusion. Semantic information increases with the number of possibilities that an assertion eliminates, and so it is inversely related to an assertion's probability (Johnson-Laird, 1983, Ch. 2; Adams, 1998). And semantic information yields a taxonomy of reasoning. Deductions do not increase semantic information even if they concern probabilities, but inductions do increase it. Simple inductions, such as generalizations, rule out more possibilities than the premises do. Abductions, which are a special case of induction, introduce concepts that are not in the premises in order to create explanations (see Koslowski, this volume). The present chapter illustrates how the model theory elucidates these three major sorts of reasoning: deduction, abduction, and induction.

Deductions yield *valid* inferences in which true premises are bound to yield true conclusions. Psychologists sometimes suggest that deductions play less of a role in daily life than probabilities. However, the maintenance of consistent beliefs is necessary for sensible decisions, because if beliefs are inconsistent then at least one of them is false, and to act on a false belief is a recipe for disaster (see Collins & Hahn, this volume). To test consistency, however, is to use a deductive procedure. Indeed, some systems of logic work on the principle that a deduction is valid because its premises are inconsistent with the denial of its conclusion (Jeffrey, 1981). And deductions can be about uncertainties, possibilities, and probabilities. So they are crucial in everyday life.

Abductions of explanations serve diverse purposes. They can resolve inconsistencies. You know that your friends have gone shopping, and that if so they'll be back in fifteen minutes. When they haven't returned in two hours, you worry. You inferred their imminent return, and their absence conflicts with your conclusion. In orthodox logic, you don't have to withdraw

your conclusion. Logic is *monotonic*: new facts imply new conclusions, never the need to retract old ones, not even those that facts contradict. Everyday reasoning, however, is not monotonic, and so theorists have devised various "nonmonotonic" logics that allow facts to undermine conclusions (see Stenning & Varga, this volume, and Oaksford & Chater, this volume). The model theory takes a different tack: your main concern is not to withdraw your conclusion, but to determine what's happened to your friends. You need an explanation of their delay so that you can decide what to do. Abduction generates explanations, and, as we show, it can lead to the withdrawal of conclusions. Abduction also underlies the development of computer programs – they are, in effect, explanations of how to get from a start to an end. And mental simulations underlie their abduction.

Simple inductions turn facts into general rules. You keep your money in a bank, because it will be safe. Your model of banks has gone beyond observation to generalization (for an account of the process, see Johnson-Laird, 1993). But canny inductions yield probabilities. You keep your money in a bank, because it is almost certain to be safe. You might even be prepared to infer a numerical value, say, 95%. The economist John Maynard Keynes tried to meld probability and deduction. That goal, too, is one to which the model theory aspires. In what follows, we review case studies illustrating the role of models in the three sorts of reasoning: deductions that draw conclusions from conditionals, abductions that resolve inconsistencies and that create programs, and inductions that yield numerical probabilities.

Deductions from conditionals

The model theory

The basic model theory is simple. People think about possibilities. They represent possibilities in mental models. And a model is a model because its structure corresponds to the structure of what it represents. So, a sequence of events can be simulated in a kinematic model that unfolds in time. An inference from premises is valid if its conclusion holds in all of their models. Intuitive reasoning copes only with a single mental model at a time, because it has no access to working memory, whereas deliberative reasoning can access working memory, and so it can construct alternative models and carry our more powerful computations (Johnson-Laird, 1983, Ch. 6). Like many "dual process" accounts (e.g., Kahneman, 2011), the model theory postulates one system (1) for intuitions and another system (2) for deliberations. In contrast, Evans (this volume) distinguishes, not two different systems, but two different sorts of process. Unlike other accounts, however, the model theory's two systems and the interactions between them are embodied in a computer program, *mReasoner* (Khemlani & Johnson-Laird, 2013). To illustrate the theory, we describe deductions from conditional assertions, because *if* is central to everyday reasoning.

Consider the following inference:

If the car started, then the battery has power.
The car started.
What follows?

The conclusion is immediate: *the battery has power*. In contrast, consider this inference:

If the car started, then the battery has power.
The battery doesn't have power.
What follows?

The inference is harder, and those individuals who rely only on intuition respond: "nothing follows". The conditional yields two *mental models*, which we represent in the following diagram:

car started . . . battery has power

The first model represents the possibility in which both clauses in the conditional are true. For convenience, the diagram omits a symbol for conjunction, and uses English phrases, whereas actual models are representations of the world. The second mental model, shown as an ellipsis, has no explicit content, but stands in for the other possibilities in which the car didn't start. In the first inference, the premise that the car started picks out the explicit model, which yields the conclusion: *the battery has power*. But, in the second inference, the premise that the battery doesn't have power eliminates the explicit model, and nothing follows from the implicit model, because it has no content. Which explains the intuition that nothing follows. However, those who deliberate can flesh out mental models into *fully explicit models* of all the possibilities to which the conditional refers. These models enable them to draw a valid conclusion. Individuals can also list these possibilities (e.g., Barrouillet, Grosset, & Lecas, 2000). For a basic conditional (*If A then C*), where knowledge does not affect interpretation, they list them in the following order:

car started	battery has power	(A C: the explicit mental model)
not (car started)	not (battery has power)	(not-A not-C)
not (car started)	battery has power	(not-A C)

They also list the following case as not possible:

car started	not (battery has power)	(A not-C)

The premise that the battery doesn't have power rules out the first and third possibilities above, and the second possibility yields the valid conclusion:

The car didn't start.

Table 19.1 summarizes these models, and those for other sorts of conditionals, which we discuss.

Table 19.1 The sets of possibilities to which various sorts of conditional refer, depending on whether the conditionals are affirmed or denied. These possibilities correspond to fully explicit models of the conditionals, and the empty cells are cases that are not possible according to the conditionals. The two scopes of denial yield identical possibilities for biconditionals and relevance conditionals.

Basic conditionals e.g., *If the car started, then the battery has power.*			Biconditionals e.g., *If and only if the car started, then the battery has power.*		Relevance conditionals e.g., *If it's raining, then here's an umbrella.*	
Affirmed	Denied with large scope: *Not if A then C*	Denied with small scope: *If A then not C*	Affirmed	Denied	Affirmed	Denied
A C not-A C not-A not-C	A not-C	A not-C not-A C not-A not-C	A C not-A not-C	A not-C not-A C	A C not-A C	A not-C not-A not-C

Every psychological theory of deduction explains the difference between the easy conditional inference (aka "modus ponens") and the difficult one (aka "modus tollens"). But, several results corroborate the model theory and challenge other accounts. One result is that the difficult inference becomes easier with biconditionals, for example:

If, and only if, the car started, then the battery has power.

The model theory predicts the effect (Johnson-Laird, Byrne, & Schaeken, 1992), because biconditionals have only two fully explicit models (see Table 19.1):

car started	battery has power	(A C)
not (car started)	not (battery has power)	(not-A not-C)

Another predicted result is that the difficult inference becomes easier when the categorical premise is presented first:

The battery doesn't have power.
If the car started, then the battery has power.

The categorical premise blocks the construction of the explicit mental model of the conditional premise, and so reasoners are more likely to consider the fully explicit model:

not (car started)	not (battery has power)

And it shows that the car didn't start (Girotto, Mazzocco, & Tasso, 1997). Theories based on formal rules predict neither of these effects (cf. Rips, 1994). In a review, Oberauer (2006) showed that the model theory also gives a better account of conditional reasoning than a probabilistic theory (Oaksford, Chater, & Larkin, 2000).

The verification of conditionals

Let us turn to the circumstances in which conditionals are true, and to a result that seems contrary to the model theory. It comes from the first study to investigate the verification of conditionals. The participants had to judge the impact of single pieces of evidence on the truth or falsity of such conditionals as:

If there is an A on the left, then there is a 2 on the right.

On each trial, the participants weighed the evidence of a single card, such as: *A 2*. They tended to make these judgments (Johnson-Laird & Tagart, 1969):

A 2: shows that the conditional is true.
A 3: shows that the conditional is false.
B 2: is irrelevant to the conditional's truth value.
B 3: is irrelevant to the conditional's truth value.

Evidence that a conditional's *if*-clause is false seems irrelevant to whether the conditional is true or false. In contrast, when individuals list what is possible given a conditional, they list these cases as possible (e.g., Barrouillet et al., 2000). And so there is a discrepancy, because a case that is possible according to a conditional ought to show that the conditional is true. Some theorists argue that the judgments of irrelevance confirm that conditionals are void – they have no truth value – when

their *if*-clauses are false. This hypothesis, which goes back to de Finetti (1937/1964), is one of the chief assumptions of the "new paradigm", which is a family of theories that seek to replace logic with probability (see, e.g., Elqayam, this volume; Oaksford & Chater, this volume; Over & Cruz, this volume; and, for a review, Johnson-Laird, Khemlani, & Goodwin, 2015). The hypothesis implies that if a conditional is true, then its *if*-clause must be true too, because otherwise the conditional would have no truth value. But, consider the implications. This conditional is true:

If God exists, atheism is wrong.

And so its *if*-clause is true too: therefore, God exists. The inference is valid and a very short proof of God's existence. But if the proof isn't convincing, it must be because a true conditional can have a false *if*-clause. Hence, Johnson-Laird and Tagart's verification task is misleading (Schroyens, 2010).

Indeed, in a different task, participants do evaluate conditionals as true when their *if*-clauses are false. Participants assessed whether sets of assertions containing conditionals could all be true at the same time (Goodwin & Johnson-Laird, 2015). The context of the following assertions made clear that they concerned a particular animal:

If the animal has brown fur then it is a bear.	(If A then C)
The animal does not have brown fur.	(Not-A)
The animal is a bear.	(C)
Could all of these assertions be true at the same time?	

The experiment examined all four possible sorts of categorical evidence, and even when, as in this example, the conditional's *if*-clause was false, the participants judged that the assertions could all be true (at around 80%).

Readers might judge that two assertions, *A* and *B*, could both be true at the same time, and yet *A* is irrelevant to the truth of *B*. For example, "Viv is tall" can be true at the same time as "Viv is amusing", but each assertion is irrelevant to the truth of the other. The preceding results, however, are different, because the assertions are a conditional and its two clauses. Individuals judged that a conditional ("If Viv is tall, then she is amusing") can be true at the same time as an assertion that its *if*-clause is false ("Viv is not tall"). It follows that the conditional can be true when its *if*-clause is false. Otherwise, the participants would judge that the two assertions couldn't both be true at the same time. This result therefore undermines the idea that a conditional has no truth value when its *if*-clause is false.

Suppose you observe:

It's raining and it's hot.

Your observation establishes the falsity of the conditional: *if it's raining, then it's not hot*, and of its equivalent: *if it's hot, then it's not raining*. But, by itself, it does *not* establish the truth of the affirmative conditional:

If it's raining, then it's hot.

Your observation fails to distinguish between this conditional and one that is not equivalent to it:

If it's hot, then it's raining.

A conditional refers to a set of cases, and they each have to be possible for the conditional to be true. If your observation shows that one of them is a matter of fact:

raining hot

then the other cases:

not raining not hot
not raining hot

become counterfactual possibilities (Byrne, 2005; Byrne, this volume). If they are true counterfactuals, then the conditional is true. Thus, if a single categorical observation is to verify a conditional (as in the Johnson-Laird & Tagart study), then the context needs to establish the appropriate possibilities prior to the observation. In a recent study, Goodwin and Johnson-Laird (2015) framed the verification task so that the first assertion made the disjunctive possibilities explicit, e.g.:

John is deciding whether to fire one of his two employees, Charlie or Annie, and possibly both of them.

In the end, John fired Annie.
A colleague of John's had predicted: John will fire Charlie, if not Annie.

Was the colleague's prediction true, false, or irrelevant, in light of what happened?

Participants should build these mental models of the people whom John fired according to the conditional and its disjunctive context:

Charlie not Annie
Annie
Charlie Annie

The results were that most participants judged the conditional prediction to be true even in those cases, such as the one above, in which its *if*-clause was false. Of course, these findings may not generalize beyond this study. Yet, like the proof of God's existence, they are a mystery if conditionals cannot be true when their *if*-clauses are false.

The modulation and denial of conditionals

Knowledge can *modulate* the interpretation of conditionals and other connectives (Johnson-Laird & Byrne, 2002). It can introduce temporal relations between events, e.g., *if he passed the exam, then he studied hard*, in which the studying preceded the exam. It can also block the construction of a fully explicit model of a conditional. This blocking yields two main alternatives to a basic conditional, *If A, then C*. One is a biconditional interpretation when knowledge blocks the possibility of *not-A* and *C* (see Table 19.1), e.g.:

If it's raining, then it's pouring

because it can't pour without raining. The other is a relevance interpretation that blocks the possibility of *not-A* and *not-C* (see Table 19.1), e.g.:

If it's raining, then I have an umbrella.

The *if*-clause states a condition that is no more than relevant to the truth of the *then*-clause. Modulation cannot block the mental model, *A* and *C*, at least when another possibility holds. But, ironic conditionals can block it and refer to only a single possibility, for example:

If you're right, then I'll eat my hat.

That is to say, you're not right and I won't eat my hat (Johnson-Laird & Byrne, 2002).

Some conditionals contain a modal verb so that they refer to a possible consequence, e.g.:

If it is raining, then there may be a flood.

They usually mean that *A enables B to occur*, and so *not-A and B* is impossible. Modulation can affect their interpretation too, and therefore the inferences that participants draw from them (Quelhas, Johnson-Laird, & Juhos, 2010). It also affects the interpretation of the temporal relations between events, including the participants' uses of tense in drawing conclusions (Juhos, Quelhas, & Johnson-Laird, 2012). Another effect of modulation is to establish the truth values of some conditionals a priori. For instance, knowledge of the meaning of "atheism" entails the falsity of *If God exists, then atheism is right*.

The denial of a conditional, *if A then C*, transforms the case of *A & not-C* into a possibility and the case of *A & C* into an impossibility. One complication, however, is that the *if*-clause of a conditional is subsidiary to its main *then*-clause (Khemlani, Orenes, & Johnson-Laird, 2012, 2014). Hence, as experiments show (Khemlani, Orenes, & Johnson-Laird, 2014), the denial of a conditional:

It's not the case that if it's raining, then it's hot

is sometimes taken to be equivalent to:

It's raining and it's not hot

and sometimes taken to be equivalent to:

If it's raining, then it's not hot.

In the first denial, negation has the whole of the conditional within its scope, and in the second denial, negation has only the *then*-clause within its scope. The two scopes yield identical possibilities for biconditionals and relevance conditionals. Table 19.1 shows these denials.

Conditionals' conjunctions of possibilities

An analogy exists between the fully explicit models of possibilities for basic conditionals and a connective in logic known as *material implication*. The proposition that *A* materially implies *C* is true or false depending only on the truth or falsity of its two clauses. It is true in just the three cases that are possible for basic conditionals (see the left-most column of Table 19.1), and it is false in the one case that is impossible (*A & not-C*). So, the only way in which a material implication can be false is in case *A* is true and *C* is false. On this account, the false conditional:

If God exists then atheism is right

has an *if*-clause that is true, i.e., God exists. It's an absurd consequence. Likewise, material implication yields paradoxical inferences, such as:

Donald Trump will not be the next U.S. president.
Therefore, if Donald Trump is the next U.S. president, then ISIS takes over Tokyo.

The premise establishes the falsity of the conditional's *if*-clause, and that suffices for the conditional to be true granted that it's a material implication. So what does the model theory imply?

It accepts the analogy: basic conditionals refer to the three possibilities that are true for a material implication. But, analogy is not identity. Truth values are alternatives, and so an assertion and its negation cannot both be true: *it will rain and it won't rain* is a self-contradiction, because one clause is true and the other is false. But possibilities are conjunctive, and so an assertion and its negation can both be possible: *possibly it will rain and possibly it won't rain*. A basic conditional is true a priori in case all and only its three fully explicit models refer to possibilities (see Table 19.1). The falsity of a conditional, such as *if God exists, then atheism is right* does not imply that its *if*-clause is true, but that the case in which both clauses are true is impossible, and that the case in which its *if*-clause is true and its *then*-clause is false is possible. Likewise, the falsity of its *if*-clause does not establish that a conditional is true, because this fact alone does not establish that the other relevant cases are possible. The paradoxical inference about Donald Trump is invalid, because the premise does not establish that all the cases to which the conditional refers are possible.

The hypothesis that compound assertions refer to conjunctions of possibilities is borne out in a study in which participants deduced conclusions about what is possible (Hinterecker, Knauff, & Johnson-Laird, 2015). We invite readers to consider whether the following sort of inference is valid:

A or B or both.
Therefore, it is possible that A.

It may help to consider a particular example from the experiment, such as:

Scientists will discover a cure for Parkinson's disease in 10 years or the number of patients who suffer from Parkinson's disease will triple by 2050, or both.
Therefore, it is possible that scientists will discover a cure for Parkinson's disease in 10 years.

The participants thought that the inference is valid, and they also inferred conjunctive conclusions of the sort:

It is possible that A and B.

Yet, none of these conclusions is valid in any logic. The first inference is invalid, because *A* could be a self-contradiction. The premise could still be true in this case, but self-contradictions are impossible, and so the conclusion would be false. The second inference is invalid because *B* could imply *not-A*. The premise could still be true in this case, but the conclusion would again assert that a contradiction is possible: *A and not-A*. The proofs of the two inferences in logic

therefore call for an additional premise to rule out the potential contradictions. A plausible additional premise for the first inference is:

Not necessarily (not A).

And a plausible additional premise for the second inference is:

Not necessarily (B materially implies not-A).

Alas, each of these premises turns out to be equivalent to the respective conclusion to be proved, and so the actual premise for the inference (*A or B or both*) becomes superfluous. It is not obvious what the additional premises should be. But the inferences are straightforward in the model theory. The premise refers to a conjunctive set of mental models of possibilities:

A
B
A B

The conclusions follow validly and at once from these models. And if *A* were self-contradictory, or *B* were to imply *not-A*, modulation would block the construction of the relevant models. The general moral extends to conditionals: they too refer to conjunctive sets of possibilities.

Many conditionals in daily life have main clauses that are imperatives, and imperatives do not have truth values, e.g.:

If you owe money, then pay at least some of it back.

Actions that imply compliance with the *then*-clause count as satisfying the request, e.g., the debtor pays all the money back. Because the model theory treats conditionals as referring to possibilities, it captures such inferences. When the *if*-clauses of conditional requests or bets are false, the listener is under no obligation to do anything. When the *if*-clauses of conditional instructions in computer programs are false, control likewise passes to the next instruction. But, when the *if*-clauses of conditional assertions in daily life are false, the conditionals can nevertheless be true.

Abduction, nonmonotonicity, and programming

Abduction and nonmonotonicity

Researchers have devised various nonmonotonic logics, which allow conclusions to be weakened or withdrawn in the face of conflicting evidence (see Stenning & Varga, this volume, and Oaksford & Chater, this volume). These logics, however, overlook a major psychological problem. In our earlier example, why have your friends not returned from shopping? You need to abduce a resolution of the inconsistency between beliefs and the facts: you need a causal explanation of what has gone wrong. In the model theory, causal relations are deterministic, not probabilistic (e.g., Goldvarg & Johnson-Laird, 2001), because probabilities cannot tell you what is wrong with the claim: *mud causes rain* (Pearl, 2009). Indeed, probabilities cannot distinguish between cause and correlation, or between causing and enabling (Frosch & Johnson-Laird, 2011).

The model theory of how reasoning resolves inconsistencies is implemented in a computer program (Johnson-Laird, Girotto, & Legrenzi, 2004). To illustrate its operations, consider the following problem:

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

Causal links are possibilities ordered in time (Khemlani, Barbey, & Johnson-Laird, 2014). The first two premises in the problem elicit a mental model of such a causal link:

pulled trigger gun fired

The fact that the gun did not fire contradicts the preceding mental model:

not (gun fired).

The conjunction of two models with contradictory elements usually yields the null model, which represents contradictions, and from which nothing follows. But the model theory embodies a simple nonmonotonic principle. When a contradiction arises from two separate premises, one premise can take precedence over the other, and a fact takes precedence over a contradictory element in other models. In our example, the fact contradicts the *then*-clause in the model but takes precedence over it, and so the model of what needs to be explained is:

pulled trigger not (gun fired).

Abduction aims to simulate a causal chain that explains why pulling the trigger did not fire the gun. It elicits models from knowledge of what prevents guns from firing. Human reasoners are likely to think about which explanation is most plausible: the gun was broken, the gun jammed, the safety catch was on, the gun had no bullets, and so on. The program chooses at random from them, for example:

gun broken.

This model, in turn, triggers a search for what was its cause, for example:

gun dropped.

The result is a causal simulation in which dropping the gun caused it to break, and so, when someone pulled the trigger, the gun did not fire. The conditional premise is no longer true, and the program adds a rider to express a counterfactual qualification:

If the gun hadn't broken, then it would have fired.

As the theory predicts, individuals spontaneously abduce explanations of inconsistencies. They judge such explanations as more probable than simpler revisions to the premises that restore consistency without explaining the contradiction (Johnson-Laird et al., 2004;

Khemlani & Johnson-Laird, 2011). Because explanations resolve inconsistencies, they have a surprising side effect. They make it harder to detect inconsistencies. A series of studies presented participants with pairs of assertions, such as:

If a person is bitten by a viper, then the person dies.
Someone was bitten by a viper, but did not die.

The participants then carried out two tasks: they created an explanation and they judged whether or not both assertions could be true at the same time (Khemlani & Johnson-Laird, 2012). The order of the two tasks had a robust effect on performance. Prior explanations made inconsistencies harder to detect by 20% or more. The effect probably occurs because explanations resolve inconsistencies. Hence, a different prior task, assessing how surprising the events were, had no such effect. In sum, inconsistencies elicit explanatory abductions, and the potential precedence of one model over another obviates the need for a special nonmonotonic logic.

Abduction and programming

A program is an explanation of how to carry out a computation. So, abductions underlie the development of programs. Psychologists have studied novice programmers writing code in a programming language, but the model theory has inspired studies of how individuals who know nothing of programming create informal programs (Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). The test-bed for these studies is a toy railway depicted on a computer screen, and Figure 19.1 shows its initial state for a simple problem. Only three sorts of move can occur: one or more cars move from the left track to the right track, from the left track to the siding, or from the siding back to the left track. (Moves from the right track are not allowed.) The siding functions like a working memory: items (cars) can be stored on the siding, while others move from input (left) to output (right). The left track also functions as a working memory, because cars can shuttle between it and the siding. In principle the computational power of the railway is therefore equivalent to a universal Turing machine – a device that can carry out any sort of program (Hopcroft & Ullman, 1979).

We studied rearrangements of the order of cars – of which there are many, such as rearranging the six cars *abcdef* on the left track into the reverse order *fedcba* on the right track. The abduction of programs for rearrangements according to the model theory calls for three steps, which are each implemented in a computer program, *mAbducer*, that develops its own programs to make rearrangements; that is, it is an automatic programmer.

The first step is to solve some instances of the relevant rearrangement for trains with different numbers of cars. Although just three sorts of move are permissible, trial and error works for only simple rearrangements of a few cars. Some sorts of problems, such as the Tower of Hanoi,

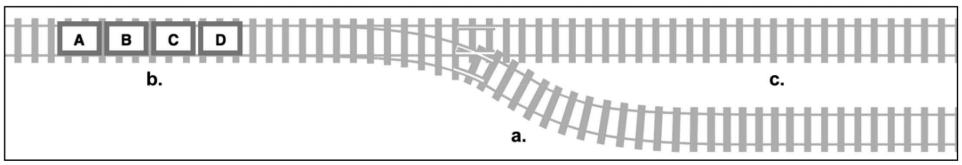


Figure 19.1 The railway environment used in studies of informal programming. Cars can enter the siding (a) only from the left side of the track (b), and exit from it only to the left track. They can also move from the left side of the track to the right side (c)

can be solved using a means–ends analysis in which you choose moves to reduce the difference between the current state and the goal. But for rearrangements, you need to decompose the goal into separate parts. *mAbducer* uses a schematic and kinematic model:

abcdef [–] –

This diagram represents a model containing six cars on the left track, no cars on the siding, which is represented by the square brackets, and no cars on the right track. *mAbducer* finds parsimonious way to make any rearrangement. It matches cars on different parts of the track with the current goal, and finds the best move. Suppose, for example, that the goal is to rearrange the cars above into the reverse order *fdceba* on the right track. The obvious first move is to move all the cars apart from *a* to the siding:

a[bcdef] –

A loop of moves can now be repeated. Move one car to the right, and one car from the siding to the left:

b[cdef]a

The goal is updated to *fdceb*, because *a* is now in its rightful position. The loop continues until the siding is empty. The final move is of car *f* over to the right. Human reasoners lack a parsimonious procedure for solving rearrangements. Every participant in a study failed to find the solution with the fewest moves for at least one sort of rearrangement (Khemlani et al., 2013). A common blunder was perseveration: when participants had moved a single car, they moved it again, overlooking a move of two cars – the car they'd just moved and the one behind it.

The second step in abducing a program is to determine the structure of the solutions to the same rearrangement of different numbers of cars. A program to rearrange trains of any length is bound to contain a loop of operations. Hence, *mAbducer* simulates solutions in order to recover a required loop of moves (such as the one above) and any moves before or after the loop. *mAbducer* uses the results of its simulations to construct two alternative programs: one uses a *for*-loop, which is repeated *for* a fixed number of times, and the other uses a *while*-loop, which is repeated *while* a given condition holds. *While*-loops are more powerful than *for*-loops, because they can carry out computations that *for*-loops cannot. *mAbducer* has to solve a pair of simultaneous equations to determine the number of repetitions for a *for*-loop (see Khemlani et al., 2013). It has only to inspect its simulations to determine the conditions in which a *while*-loop halts: for example, for a reversal the *while*-loop halts as soon as the siding is empty. Hence, naïve individuals should be more likely to discover *while*-loops, even though they are more powerful computational devices.

The program translates the structures of solutions into the programming language Common Lisp. It also translates the *while*-loop version of a program from Lisp into informal English. Here is its description of the program for a reversal:

Move one less than the [number of] cars to the siding.
While there are more than zero cars on the siding,
    move one car to the right track,
    move one car to the left track.
Move one car to the right track.

Table 19.2 Four rearrangements of cars in a train, their minimal number of moves, the Kolmogorov complexity of programs for solving the rearrangements, and the results of experiments on programming the rearrangements and on making deductions from them (Khemlani et al., 2013)

The name of a rearrangement and an example of it	*Minimal number of moves*	*Kolmogorov complexity of programs*	*Participants' percentages of correct programs*	*Participants' percentages of correct deductions from programs*
Reversal: abcdef ⇒ fedcba	12	1,288	90	41
Palindrome: abccba ⇒ aabbcc	6	1,295	70	35
Parity sort: abcdef ⇒ acebdf	7	1,519	63	32
Reverse parity sort: acebdf ⇒ abcdef	9	1,771	-	23

In an experiment, naïve participants formulated programs for rearrangements in their own words. As the theory predicts, they preferred to use *while*-loops rather than *for*-loops. Here is a typical example of what one participant said in abducing a *while*-loop for a reversal:

> Move all cars to the right of *A* to the side [i.e., the siding]. Then move *A* to the right. Shift *B* to left, then right. Shift *C* to left, then right . . . repeat until pattern is reached.

The participants' difficulty in making a rearrangement depended on the minimal number of moves that it calls for. However, their difficulty in abducing a program to carry out the rearrangement didn't depend on its minimal number of moves, but on the complexity of the program. Complexity reflects various factors, such as the number of instructions in the loop, but a simple metric is the number of symbols in a program written in a standard language. Table 19.2 shows this Kolmogorov complexity, where we multiplied the number of characters and spaces in *mAbducer*'s Common Lisp programs by the number of bits in a character; that is, 7 for ASCII characters. Table 19.2 also shows the percentages of the participants' correct programs. The decline in accurate programs with the increase in their complexity was robust. Reversals are the hardest rearrangement to solve – they take the longest time, because they call for the most moves. But, their programs are of the least complexity, and they are the easiest to program – naïve individuals are more likely to formulate an accurate program and to do so in a shorter time than for the other rearrangements.

The third and final step in programming is to test a program by deducing its consequences for new inputs. Individuals should carry out this task by simulating the effect of each instruction in the program on a given train. Suppose the initial train on the left track is *abcdef*. What is the order of the cars on the right track as a result of carrying out the following program, which works for trains of any length?

> While there are more than two cars on the left track,
> move one car to the right track,
> and move one car to the siding.

Move one car to the right track.
Move one less than half the number of all the cars to the left track.
Move half the number of all the cars to the right track.

Readers are invited to try to imagine the effect without recourse to pencil and paper. Here is the required sequence of moves:

Starting state:	abcdef[] -,
Carry out the two moves in the loop:	abcd[e]f
And again:	ab[ce]df
First move after the loop:	a[ce]bdf
Second move after the loop:	ace[]bdf
Third move after the loop:	-[]acebdf

This rearrangement is the parity sort shown in Table 19.2. The difficulty of the required mental simulation should again depend on the Kolmogorov complexity of the program. Table 19.2 presents the percentages of correct deductions in an experiment, and they corroborated the predicted trend that complexity predicts (Khemlani et al., 2013).

The abduction of programs for making rearrangements in the railway domain calls for the simulation of moves in kinematic models. Ten-year-old children are able to abduce informal programs for rearrangements of six cars. When they were not allowed to move the cars, their gestures were a clear outward sign of inward simulations. These gestures appeared to help them to keep track of where the cars were on the tracks. When they couldn't gesture, their programs were less accurate (Bucciarelli, Mackiewicz, Khemlani, & Johnson-Laird, 2015).

In sum, abduction simulates causal chains to resolve inconsistencies. Its kinematic simulations underlie programming. And they are pre-eminent in the creation of scientific and technological hypotheses (see Chapters 25–27, Johnson-Laird, 2006).

Induction and probabilities

Proportions and probabilities

Casual inductions about probabilities are ancient – Aristotle refers to the probable as a thing that happens for the most part – but until the formulation of the probability calculus in the 17th century, no complete normative theory of numerical probabilities existed. In daily life, reasoners use two main sorts of probabilistic reasoning (Tversky & Kahneman, 1983). They make deductions of probabilities from knowledge of the possible ways in which events can occur – for instance, they deduce that the probability that two tossed coins both land "heads" is a quarter, given that the probability of one coin landing "heads" is a half. And they make inductions about the probabilities of unique events from non-numerical evidence, for example the probability that Donald Trump is elected as president of the United States.

Deductions about probabilities rest on simple principles (Johnson-Laird et al., 1999). Reasoners assume that models of possibilities are equiprobable unless they have evidence to the contrary. They infer that the probability of an event is equal to the proportion of such models in which the event occurs. For example, consider the following problem:

There is a box in which there is a black marble, or a red marble, or both.
Given the preceding assertion, what is the probability of the following situation?
In the box there is a black marble and a red marble.

Experts may respond that the problem is ill posed because it says nothing about relative probabilities. Naïve individuals, however, construct mental models of the three possible contents of the box:

black	
	red
black	red

They assume that they are equiprobable, and so they infer a probability of about 33% (Johnson-Laird et al., 1999). Mental models predict systematic errors, because they don't represent what is false. Here's an example:

> There is a box in which there is at least a red marble, or else there is a green marble and a blue marble, but not all three marbles.
> Given the preceding assertion, what is the probability of the following situation?
> In the box, there is a red marble and a blue marble.

The mental models of the premise are as follows:

red		
	green	blue

So, most reasoners infer that the probability of red and blue in the box is zero. But, the response is an illusion. The fully explicit models of the premise show that when one disjunct is true the other is false, and there are three ways in which *green and blue* can be false when *red* is true:

red	green	not-blue
red	not-green	blue
red	not-green	not-blue
not-red	green	blue

As the second of these models shows, red and blue marbles can occur together, and so a proportional estimate of their probability is, not zero, but 25%.

The great difficulty for naïve individuals – and the great motivator for the invention of the probability calculus – is conditional probability. Many puzzles hinge on this fact. A simple one is:

> The Smiths have two children. One of them is a girl. What's the probability that the other is a girl?

The intuitive response is a half. It treats the question as concerning absolute probabilities. But, because the problem establishes a fact, in reality it asks for a conditional probability: given that one child is a girl, what's the probability that the other child is too? The inference calls for reasoners to envisage a model of all four possible pairs of children:

First-born	*Second-born*
girl	girl
girl	boy
boy	girl
boy	boy

It contains three cases in which one child is a girl. There is a subset within them in which the other child is a girl, and it has only one member. So, the correct estimate is a third (see Nickerson, 1996, for the subtleties in such problems). Conditional probabilities appear to lie at the

edge of human competence, because they call for inferences about subsets of models. They can be inferred only using fully explicit models in system 2.

The probabilities of unique events

To study the induction of the probabilities of unique events, we asked participants to estimate the probabilities of various possibilities, such as that U.S. companies will focus their advertising on the Web next year, and that the *New York Times* will become more profitable. For some probabilists, such questions verge on the nonsensical, because there are no definitive frequencies from which their answers can be inferred (e.g., Cosmides & Tooby, 1996). For "Bayesians" such as ourselves (see also Oaksford & Chater, this volume), probabilities correspond to degrees of belief, and so the questions make sense. Naïve individuals are happy to respond too, and their estimates concur to some degree about the relative probabilities of different events; for instance, they estimated the probability of a focus on Web advertising as 69% but the probability that the *Times* would become more profitable as only 41% (Khemlani, Lotstein, & Johnson-Laird, 2012). The profound mystery in such estimates is: where do the numbers come from?

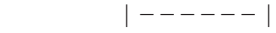
The model theory postulates that individuals adduce evidence from their general knowledge, such as:

Most U.S. newspapers will continue to be unprofitable.

It yields a corresponding model, with a small number of tokens representing U.S. newspapers – a number that can vary from occasion to occasion in a probabilistic way (see Khemlani, Lotstein, Trafton, & Johnson-Laird, 2015):

newspaper	unprofitable
newspaper	unprofitable
newspaper	unprofitable
newspaper	

People know that the *New York Times* is a U.S. newspaper, and so they translate the proportion in this model into a primitive analog representation of the probability that the paper will continue to be unprofitable:



This diagram denotes an analog model of a magnitude. Its left-hand end represents *impossibility*, its right-hand end represents *certainty*, and its length represents a non-numerical probability. Further evidence can push its length one way or another. When individuals estimate different probabilities for two events, *A* and *B*, they are uncertain about the probability of their conjunction and tend to compromise by computing a primitive average of their probabilities (Khemlani et al., 2012, 2015). Such estimates are easy to make using the analog model, but they violate the probability calculus. It calls for the multiplication of one probability, *A*, by the conditional probability of the other, *B*, given *A*. But the estimates are typical. And they occur for the probabilities of other sorts of compounds, including disjunctions and conditional probabilities. These estimates and those of their components, *A* and *B*, are therefore often subadditive – that is, they yield an exhaustive set of alternative possibilities with probabilities summing to more than 100% (Khemlani et al., 2015). AuQ1 AuQ2

In a study in which individuals assessed whether disjunctions implied the possibility of various sorts of conjunctions (Hinterecker et al., 2015), the participants were sensitive to probabilities, e.g., they judged conclusions that they accepted to be more probable than conclusions that they rejected. But they went wrong in estimating the probabilities of the four exhaustive possibilities:

A & B
A & not-B
not-A and B
not-A and not-B.

According to the model theory, naïve estimates of the probability of each conjunction should tend to be a primitive average of the probabilities of the two conjuncts, and so the theory predicts subadditivity. Multiply this effect for four conjunctions, and its extent should be considerable. In fact, it was shocking: the four probabilities summed on average to 191%. Such estimates are irrational: they render individuals vulnerable to a "Dutch book", that is, a set of bets in which they are bound to lose money. The degree of subadditivity is therefore not good news for the new paradigm, which, as we mentioned earlier, seeks to replace logic with probability. And it casts doubt on the idea that the rationality of naïve reasoners is more appropriately measured against the probability calculus than against logic.

Another result contrary to the new paradigm is that reasoners distinguish between the inferential consequences of conditionals, such as:

If the FDA approves a drug, then it is safe for human consumption.

and those that refer to probabilities, such as:

If the FDA approves a drug, then probably it is safe for human consumption.

Notwithstanding claims that probabilities are intrinsic to conditional reasoning (Oaksford et al., 2000), a series of nine experiments, which examined various sorts of deduction, demonstrated robust differences in the way participants interpreted the two sorts of conditional (Goodwin, 2014). Conditionals that do not make explicit reference to probabilities tend not to be interpreted as probabilistic.

Readers should not conclude that reasoning is a deterministic process. Its machinery is not clockwork: the same premises yield different conclusions on different occasions. Part of the *mReasoner* program models a probabilistic process for inferences such as:

All Greeks are athletes.
Some athletes are Greeks.
Can both of these assertions be true at the same time? (Yes.)

A typical mental model of the first assertion also represents all athletes as Greek. This model satisfies the second assertion, and so the inference is easy. But the task is harder with this pair of assertions:

All Greeks are athletes.
Some athletes are not Greek.

The typical mental model of the first assertion doesn't satisfy the second assertion, and the correct affirmative response depends on finding an alternative and atypical model:

Greek	athlete
Greek	athlete
Greek	athlete
	athlete
	athlete

Yet, as in almost all studies of reasoning, these inferences are seldom uniform. A probabilistic mechanism accounts for this variation. *mReasoner* implements it using parameters governing three probabilities (Khemlani, Lotstein et al., 2015). The first parameter constrains the number of individuals represented in a model – the number is always small, but it varies. The second parameter governs the sorts of individual represented in a model – the likelihood that they are typical for an assertion as opposed to drawn from the set of all possible individuals satisfying the assertion. The third parameter governs the chances that system 2 is engaged to make a deliberate search for an alternative to system 1's intuitive mental model. The program responds in a different way to the same inference on different occasions. It behaves like a biased roulette wheel. We ran it many thousands of times to find optimal values for the parameters. Their values accounted for the variations in both the intuitive and deliberative systems of human reasoning. The reasoning engine is probably probabilistic. AuQ3

Conclusions

What is common to the model theory of deduction, abduction, and induction? The answer is that all three rely on models of possibilities. Reasoners can construct models from descriptions and from perception, and they can retrieve them from knowledge. They deduce conclusions from models. They connect one set of models to another to abduce causal explanations. They determine that one model should take precedence over another that contradicts it, and thereby retract conclusions that facts refute. They use kinematic models in order to determine the conditions in which a loop of operations should halt, and to deduce the consequences of informal programs. They induce probabilities for unique events by transforming a proportion in a model into an analog representation of a magnitude. On the whole, experiments corroborate the theory. Compound assertions refer to conjunctive sets of possibilities, and so conditionals can be true when their *if*-clauses are false. Deductions are easier when fewer models of possibilities are at stake, and so it is easier to reason from biconditionals than conditionals. When facts contravene conclusions, individuals are spontaneous in explaining the inconsistency rather than in amending the inconsistent descriptions. They create programs that tend to use *while*-loops rather than *for*-loops. The difficulty of the task and of deducing the consequences of programs depends on the complexity of the programs. They compromise when probabilities conflict, and so their estimates of compound events can be subadditive to a massive extent. The model theory is far from perfect. Its integrative computer model, *mReasoner*, is as yet incomplete, though it does illustrate how to integrate probabilities and deduction, and how to distinguish intuitive reasoning (system 1) from deliberative reasoning (system 2). Experimental results may yet overturn the theory. At present, however, it elucidates a wider variety of inferences than alternative accounts, and it has led to robust findings that challenge these alternatives.

References

Adams, E. W. (1998). *A primer of probability logic*. Stanford, CA: CSLI Publications.

Barrouillet, P., Grosset, N., & Lecas, J. F. (2000). Conditional reasoning by mental models: Chronometric and developmental evidence. *Cognition*, *75*, 237–266.

Bucciarelli, M., Mackiewicz, R., Khemlani, S. S., & Johnson-Laird, P. N. (2015). *Simulations and gestures in children's creation of algorithms*. Manuscript under submission.

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

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

de Finetti, B (1964). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & J. Smokler (Eds.), *Studies in subjective probability* (pp. 93–158). New York: Wiley. (Originally published in 1937.)

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

Girotto, V., Mazzocco, A., & Tasso, A. (1997). The effect of premise order in conditional reasoning: A test of the mental model theory. *Cognition*, *63*, 1–28.

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.

Goodwin, G. P., & Johnson-Laird, P. N. (2015). *The truth of conditional assertions*. Manuscript under submission.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2015). *Modality, probability, and mental models*. Manuscript under submission.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley.

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

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*, 71–115.

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

Johnson-Laird, P. N. (1993). *Human and machine thinking*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N. (2006). *How we reason*. New York: 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., Byrne, R. M. J., & Schaeken, W. S. (1992). Propositional reasoning by model. *Psychological Review*, *99*, 418–439.

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

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*, 201–214.

Johnson-Laird, P. N. and Tagart, J. (1969). How implication is understood. *American Journal of Psychology*, *82*, 367–373.

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

Kahneman, D. (2011) *Thinking, fast and slow*. London: Allen Lane.

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*, 276–288.

Khemlani, S., & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica*, *139*, 486–491.

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. (2012). The probability of unique events. *PLOS-ONE*, 7, 1–9. Online version.

Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2015). Naive probability: Model-based estimates of unique events. *Cognitive Science*, *39*, 1216–1258. (Published on line, 2014).

Khemlani, S., Lotstein, M., Trafton, J. G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *Quarterly Journal of Experimental Psychology*. On line.

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*(42), 16766–16771.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*, 541–559.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2014). The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica*, *151*, 1–7.

Nickerson, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin*, *120*, 410–433.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 883–899.

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, *53*, 238–283.

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

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology*, *63*, 1716–1739.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Schroyens, W. (2010). Mistaking the instance for the rule: A critical analysis of the truth-table evaluation paradigm. *Quarterly Journal of Experimental Psychology*, *63*, 246–259.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.