CHAPTER ONE

Toward a Unified Theory of Reasoning

P. N. Johnson-Laird[*,†,1], **Sangeet S. Khemlani**[‡]
[*]Department of Psychology, Princeton University, Princeton, NJ, USA
[†]Department of Psychology, New York University, New York, NY, USA
[‡]Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA
[1]Corresponding author: E-mail: phil@princeton.edu

Contents

Abstract

This article describes a theory that uses mental models to integrate deductive, inductive, and probabilistic reasoning. It spells out the main principles of the theory and illustrates them with examples from various domains. It shows how models underlie inductions, explanations, estimates of probabilities, and informal algorithms. In all these cases, a central principle is that the mind represents each sort of possibility in a separate mental model and infers whatever holds in the resulting set of models. Finally, the article reviews what has been accomplished in implementing the theory in a single large-scale computer program, *mReasoner*.

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00001-0

The Psychology of Learning and Motivation, First Edition, 2014, 1-42

1. INTRODUCTION

The capacity to reason underlies mathematics, science, and technology. It is essential for coping with everyday problems—without it, social life would be almost unimaginable. The challenge to psychologists is to explain its underlying mental mechanisms. Since Störring's (1908) pioneering study, they have discovered several robust phenomena. Perhaps the most important is that naive reasoners—those with no training in logic—can make *valid* deductions, that is, inferences in which the conclusion is true in all the cases in which the premises are true (cf. Jeffrey, 1981, p. 1). And they are happy to do so about abstract matters with no ecological validity, as in Sudoku puzzles (Lee, Goodwin, & Johnson-Laird, 2008).

Fifty years ago, psychologists took for granted that human reasoning was rational. Individuals developed deductive competence during childhood, and the psychologists' task was to pin down the nature of the formal logic underlying this ability. As Inhelder and Piaget (1958, p. 305) wrote, "Reasoning is nothing more than the propositional calculus itself". There may be vagaries in performance, but faulty reasoning does not occur (Henle, 1962) or is attributable to local malfunctions in the system—a spanner in the works rather than an intrinsic flaw (Cohen, 1981). Indeed, theories of deduction in cognitive psychology began with accounts based on formal logic (e.g. Braine, 1978; Johnson-Laird, 1975; Osherson, 1974–1976). These views, of course, echo those of Enlightenment philosophers. But, another robust phenomenon that psychologists discovered is that individuals differ in their ability to reason. A few are very good, a few are very bad, and most are somewhere in the middle. Differences in ability are vast, and correlate with the tests of academic achievement, as proxies for measures of intelligence (Stanovich, 1999). For everyone, however, failures are inevitable: complex inferences are computationally intractable.

Nowadays, a consensus exists that that the psychology of reasoning has undergone a deep change—even, some say, a paradigm shift. The accessibility of digital computers was a license for theorizing, and psychological theories of reasoning have multiplied at a startling rate. Quite what the new foundations of reasoning should be is controversial. One view is that humans are rational but in terms of the probability calculus rather than logic (e.g. Oaksford & Chater, 2007; Tenenbaum & Griffiths, 2001). One view is that natural selection has equipped the mind with modules for reasoning about special topics, such as social exchange (Cosmides & Tooby, 2005). One view is that rationality presupposes a normative system, and psychologists should

abandon norms in favor of descriptions (Evans, 2012). One revenant is that logic, or logics, provides the inferential machinery (Rips, 1994; Stenning & van Lambalgen, 2008). We will not try to assess these views, but for those who espouse such a theory, we recommend the answers to two questions as a recipe for resipiscence: Has the theory been implemented in a computer program, and does it predict most of the 60 or more experimental results reported here? The goal of the article, however, is not polemical, but to describe a different theory.

Craik (1943) postulated that thinking was based on making mental simulations of the world to anticipate events. This idea in turn has historical antecedents, although Craik was unlikely to have known them (see Johnson-Laird, 2004, for the history of mental models). Oddly, however, Craik did not consider reasoning, other than to make a casual remark that it was based on "verbal rules" (Craik, 1943, p. 81). In the spirit of Craik, we argue that the mind is neither a logical nor a probabilistic device, but instead a device that makes mental simulations. Insofar as humans reason logically or infer probabilities they rely on their ability to simulate the world in mental models. The application of simulation to reasoning is based on mental models of the possibilities to which the premises refer, and a valid deduction has a conclusion that holds in all these models. This idea was first proposed a generation ago (Johnson-Laird, 1975). Since then, its proponents and critics have revised and extended it in hundreds of publications.

The theory of mental models—the model theory for short—is controversial, as are all current theories of reasoning, and the only way to put it beyond controversy calls for two crucial steps. The first step follows Leibniz (1685, 1952), who dreamt of replacing argument with calculation. It is to implement a unified theory of reasoning in a computer program that, for any inferential task, outputs the responses that human reasoners should make, the respective likelihoods and latencies of these responses, the processes underlying them, and, where relevant, valid or ideal responses. The second step is to show in stringent experiments that the program's predictions are correct. We are a long way from the two steps. Researchers have applied the model theory to many sorts of inferential task, implemented computational models of these applications, and tested the theory experimentally. But, until recently, the work has been piecemeal rather than unified.

What are the essentials of the model theory, and what counts as a mental model? This article is going to answer these questions step by step, and it aims above all to enable readers to understand the model theory without having to read anything else. It illustrates the theory's application to most sorts of reasoning. Its plan mirrors these aims. It begins with an outline of the

main sorts of reasoning. It follows with sections that elucidate each of the theory's main principles. It then considers the role of models in inductive reasoning, explanatory reasoning, reasoning about probabilities, and reasoning that yields informal algorithms. The final section of the article reviews what has been accomplished in unifying the theory, and in implementing it in a single computer program to achieve Leibniz's (and our) dream.

2. WHAT IS REASONING?

Suppose you infer:

If the ink cartridge is empty then the printer won't work.
The ink cartridge is empty.
So, the printer won't work.

You are making a deduction: your inference is valid because your conclusion holds in any case in which the premises hold. Suppose instead you infer:

If the ink cartridge is empty then the printer won't work.
The printer won't work.
So, the ink cartridge is empty.

You are making an induction. Your inference isn't valid because there may be another reason that the printer won't work. Yet, your conclusion may be true, especially if the printer is producing blank pages. For many theorists—Aristotle for one, all inferences fall into one of these two categories: deduction and induction.

Aristotle defined induction as an inference from a particular assertion to a universal one (*Topics*, 105a13). But, inductions are often from the particular to the particular, as is your induction about the printer. Hence, a better way to distinguish between the two sorts of inference is in terms of semantic information (Johnson-Laird, 1983, chap. 2). The more possibilities that an assertion rules out, the more information it conveys (Bar-Hillel, 1964). An inference to a conclusion that refers to all the same possibilities as the premises do, or at least includes them all in what it refers to, is a *deduction*. Consider again your earlier deduction:

If the ink cartridge is empty then the printer won't work.
The ink cartridge is empty.
So, the printer won't work.

The premises refer to just one possibility:

The ink cartridge is empty and the printer won't work.

Hence, your inference is valid because its conclusion holds in the one possibility to which the premises refer. The conclusion therefore does not

increase information. But semantics should not be confused with epistemology: a conclusion may be news to the person who draws it, bringing to mind a novel proposition. An inference to a conclusion that refers to only some of the possibilities to which the premises refer, although it may add some new possibilities too, is an *induction*. Consider again your earlier induction:

If the ink cartridge is empty then the printer won't work.
The printer won't work.
So, the ink cartridge is empty.

The premises refer to two possibilities:

The ink cartridge is empty. The printer won't work.
The ink cartridge isn't empty. The printer won't work.

Your conclusion, however, refers to only one of these two possibilities. It goes beyond the information in the premises, and it is consistent with them, that is, it is possible that the ink cartridge is empty. But, the conclusion does not follow validly. A special case of induction is one that also introduces new ideas to explain something, and this sort of reasoning is known as *abduction*, for example:

If the ink cartridge is empty then the printer won't work.
The printer won't work.
Hence, there's a fault in the connection between the computer and the printer.

Inferences either maintain or throw information away (deductions) or they increase information (inductions). One other relation between the premises and conclusion remains. If they refer to disjoint possibilities, they contradict one another. In logic, any conclusion whatsoever follows validly from a contradiction: the premises don't refer to any possibility in which the conclusion fails to hold because the premises don't refer to any possibility. Naive reasoners, however, reject inferences from self-contradictions. For them, the definition of validity has a rider: a valid inference is one in which the conclusion holds in every possibility to which the premises refer, and there is at least one such possibility.

Reasoners aim to draw conclusions that are true, or at least plausible. But, they also aim to draw novel and parsimonious conclusions, and so they would feel silly just to form a conjunction of all the premises even though such an inference is valid. They know more when they know:

It's raining

than when they know:

It's raining or it's cold, or both.

Yet, the disjunctive conclusion follows validly from the categorical premise. Hence, not all valid deductions are sensible, and it would be silly to make this particular inference because it throws away information by adding a disjunctive

alternative to a premise. So, a theory of deductive competence—of what the inferential system computes—assumes that individuals have the potential to be rational and an awareness of this potential. They abide by the foregoing constraints. In sum: "To deduce is to maintain semantic information, to simplify, and to reach a new conclusion" (Johnson-Laird & Byrne, 1991, p. 22). These constraints are not easy to embody in a theory based on formal logic, and this difficulty explains why such theories, like automated theorem provers in artificial intelligence, focus on the evaluation of *given* conclusions. Inductive competence also aims for parsimony and novelty, but it goes beyond the information given and ultimately aims to explain phenomena.

3. MODELS OF POSSIBILITIES

The fundamental assumption of the model theory is that each mental model represents what is common to a distinct set of possibilities. Hence, an assertion such as:

A triangle is on the right of a circle

has a single mental model, which we depict in this diagram:

The left-to-right axis of the model corresponds to the left-to-right axis of a scene, and the disposition of the triangle and circle in the model corresponds to their disposition in a scene for which the assertion is true. The model represents an indefinite number of possibilities that have in common only that a triangle is on the right of a circle. Of course, the relative sizes of the figures in the model, their distance apart, and so on, play no role in reasoning from the model, but we defer an explanation of how their irrelevance is represented until Section 11.

Everyone prefers to think about just one possibility at a time. Intuitions work in this way. And the theory postulates two separate systems for reasoning, one for intuitions and one for deliberations—a familiar distinction in "dual process" theories of reasoning (see, e.g. Evans, 2008; Kahneman, 2011; Reitman, 1965; Sloman, 1996; Stanovich, 1999; Verschueren, Schaeken, & d'Ydewalle, 2005). The model theory distinguishes between the two systems in computational power, and we have implemented both of them in computer programs (Khemlani & Johnson-Laird, 2012a; Khemlani, Lotstein, & Johnson-Laird, 2012). The intuitive system, which is sometimes known as "system 1", has no access to working memory, and so it can represent only one mental model at a time (Johnson-Laird, 1983, chap. 6), and it

cannot carry out recursive processes, including arithmetical operations such as counting. It lacks even the computational power of a finite-state automaton (Hopcroft & Ullman, 1979) because it can carry out a loop of operations for only a small finite number of times—a restriction that is built into its computer implementation. In contrast, the deliberative system, which is sometimes known as "system 2", has access to working memory, and so it can search for alternative mental models, and carry out recursive processes, such as counting and arithmetical operations, until they overwhelm its processing capacity.

One way in which to overwhelm the deliberative system is to force it to reason about disjunctions. An inclusive disjunction, such as:

There's a triangle or there's a circle, or both

includes the joint possibility of both the triangle and the circle, and so it refers to three sorts of possibility. It therefore calls for three mental models, which we show on separate rows in this diagram:

In this case, spatial relations play no role in the use of the models. An exclusive disjunction, such as:

Either there's a triangle or there's a circle, but not both

exclude the joint possibility, and so it calls for only two mental models:

Models preoccupy system 2, and so more models mean more work. The theory therefore predicts that deductions from exclusive disjunctions should be easier than those from inclusive disjunctions, as when either of the disjunctions above occurs with the categorical assertion:

There isn't a circle.

This assertion eliminates any model in which there is a circle, and so it follows validly in both cases that:

There is a triangle.

Evidence corroborates the prediction (e.g. Johnson-Laird, Byrne, & Schaeken, 1992), and it also shows that inferences from conjunctions, which have just one model, are easier than those based on disjunctions (García-Madruga, Moreno, Carriedo, Gutiérrez, & Johnson-Laird, 2001).

The sorts of inference that can overwhelm the deliberative system are "double disjunctions" (Johnson-Laird et al., 1992), which are from pairs of disjunctive premises, such as:

June is in Wales, or Charles is in Scotland, but not both.

Charles is in Scotland, or Kate is in Ireland, but not both.

What follows?

The two possibilities compatible with the first premise are relatively easy to envisage, but it is difficult to update them with those from the second premise, although the result is just two possibilities:

June in Wales Kate in Ireland

Charles in Scotland

Of course, real mental models represent these spatial relations, and are not phrases, which we use here for convenience. The two models yield the conclusion:

Either June is in Wales and Kate is in Ireland or Charles is in Scotland.

Inferences become even harder when disjunctions are inclusive. In one experiment, 25% of the participants, who were from the general public, drew valid conclusions from exclusive disjunctions, but this figure fell to below 10% for inclusive disjunctions. The result is hardly surprising, but what was striking was the nature of the modal errors: for all the inferences, the participants drew conclusions corresponding to a model of a single possibility. Just under a third of all the participants' responses were conclusions of this sort. The result suggests that when the task was too much for them, they fell back on their intuitions and envisaged just a single model of the premises. So, their conclusions were consistent with the premises but did not follow from them. The performance of undergraduates showed the same pattern. But, when the disjunctions were presented in equivalent electrical circuit diagrams or analogs of them, they performed better and faster (Bauer & Johnson-Laird, 1993). Their conclusions, however, still bore out a failure to consider all the possibilities, and so most errors were at least consistent with the premises.

Mental models represent possibilities, and so the more models that are necessary to make an inference, the harder that inference is to make. Individuals are in danger of overlooking a model. When the deliberative system is vastly overburdened, reasoners may even fall back on the intuitive system and draw a conclusion that is consistent with only a single model. One side effect of the use of models is that reasoners are most unlikely to draw

conclusions that throw semantic information away by adding disjunctive alternatives.

4. ICONS AND SYMBOLS

Mental models are iconic insofar as possible. What "iconic" means is that their structure corresponds to the structure of what they represent (see Peirce, 1931–1958, Vol. 4, paragraph 447). One example is the mental model of the assertion, *the triangle is on the right of the circle*, which we diagrammed in the previous section. Another example is an electrical circuit diagram with the same structure as the circuit it denotes. The great advantage of an icon, as Peirce realized, is that its inspection yields new information. Given the premises:

The triangle is on the right of the circle.

The square is on the right of the triangle.

The intuitive system can build the model:

It yields a new relation, namely, *the square is on the right of the circle*, and so this transitive inference emerges from scanning the model. To establish its validity, reasoners need to call on the deliberative system to check that no alternative model of the premises refutes the conclusion.

When premises are consistent with more than one spatial layout, inferences are more difficult than the preceding example (e.g. Byrne & Johnson-Laird, 1989; Carreiras & Santamaría, 1997; Vandierendonck, Dierckx, & De Vooght, 2004). Likewise, reasoners try to construct initial models that do not call for a rearrangement of entities (e.g. Jahn, Knauff, & Johnson-Laird, 2007; Knauff & Ragni, 2011), and inferences that call for such rearrangements are more difficult than those that do not (e.g. Krumnack, Bucher, Nejasmic, Nebel, & Knauff, 2011). Analogous results bear out the use of iconic representations in temporal reasoning, whether it depends on relations such as "before" and "after" (Schaeken, Johnson-Laird, & d'Ydewalle, 1996a) or on the tense and aspect of verbs, as in:

John has cleaned the house.

John is taking a shower.

John is going to read the paper.

Mary always does the dishes when John cleans the house.

Mary always drinks her coffee when John reads the paper.

What is the relation between Mary drinking coffee and doing the dishes?

Participants inferred that Mary drinks her coffee after doing the dishes, in an experiment that controlled such factors as order of mention (Schaeken, Johnson-Laird, & d'Ydewalle, 1996b).

The earlier example of a transitive inference is child's play (see, e.g. Bryant & Trabasso, 1971). But, many inferences based on iconicity are more complex, such as those that combine both spatial and temporal relations in kinematic simulations (see Section 10). The intuitive system can also mislead adult reasoners. It constructs only a simple model of a typical situation. Given this sort of problem:

Ann is a blood relative of Beth.
Beth is a blood relative of Cal.
Is Ann a blood relative of Cal?

Many adult reasoners respond, "Yes". The relation holds in their model, which represents lineal descendants or siblings. They fail to search assiduously for an alternative model—it takes work to engage the deliberative system, or a clue to a possible alternative model, such as a reminder that people can be related by marriage. Indeed, Ann and Cal could be Beth's parents, not blood relatives of one another (Goodwin & Johnson-Laird, 2005, 2008).

Visual images are iconic, and so some theorists suppose that they play a key role in reasoning (e.g. Kosslyn, 1994, p. 404). They do play a role: they impede reasoning. To see why, it is crucial to distinguish among relations that elicit visual images, such as "dirtier than", relations that elicit spatial relations, such as "on the right of", and relations that are abstract, such as "better than". Individuals are slowest in reasoning from visual relations (Knauff & Johnson-Laird, 2002), but do not differ reliably in reasoning from the other sorts of relation. As an fMRI study showed, only visual relations elicited additional activity in visual cortex (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003). Knauff (2013) tells the whole story: visual imagery is not necessary for reasoning, which is just as well because some relations, such as those between sets, have iconic representations that may not be visualizable.

Not everything can be represented in an icon. A crucial example is a negation, such as:

The triangle is *not* on the right of the circle.

Reasoners could try to list all the alternative affirmative possibilities—the triangle is on the left of the circle, behind it, and so on—but it would be critical, not only to include all the possibilities but also to make explicit that the list is exhaustive. Alas, neither these conditions nor the meaning of negation itself can be represented in an icon. The model theory accordingly introduces a symbol for negation, which is linked to its meaning: a negative assertion

or clause is true if, and only if, its corresponding affirmative is false. This meaning goes back to Aristotle's *De Interpretatione*, and with some exceptions, it holds for English usage (Khemlani, Orenes, & Johnson-Laird, 2012). The mental model of the preceding assertion is denoted in the following diagram:

¬ [● ▲]

where "¬" denotes the symbol for negation, and the brackets symbolize the scope of the negation, that is, what it applies to. Hence, a comparison of this model with an actual scene would yield the value "true" if, and only if, the relevant circle and triangle were not in the spatial relation represented in the embedded model.

Few people grasp the concept of "negation", and so prudent experimenters ask them about the "denial" of assertions. But, even so, most reasoners err in enumerating the possibilities referred to by compound assertions, such as:

> He denied that John was watching TV and smoking, or else Ann was writing a letter.

Once again, however, number of models is the key variable (Khemlani et al., 2012). It is harder to enumerate the possibilities for the denial of a conjunction, A and B, which has three models:

¬ A ¬ B

¬ A B

A ¬ B

than to enumerate the possibilities for the denial of an inclusive disjunction, A or B, which has one model:

¬ A ¬ B

There are plenty of other abstract concepts, such as "possibility", "truth", and "obligation" that transcend iconicity.

5. THE PRINCIPLE OF TRUTH

The model theory postulates a principle of truth: mental models represent what is true, not what is false unless assertions refer to falsity. Here is an example of an exclusive disjunction with a negative clause:

> Either there isn't a triangle or there's a circle.

It has two mental models:

They represent the possibilities in which the disjunction is true, not the possibilities in which it is false. But, the principle of truth applies at a lower level. Both of the preceding models represent clauses in the disjunction only when they are true. In contrast, *fully explicit* models also represent clauses that are false. In the first model above, it is false that there is a circle; and in the second model, it is false that there isn't a triangle, that is, there is a triangle. Hence, the fully explicit models of the exclusive disjunction are:

where we use negation to represent falsity. The fully explicit models show that the disjunction is equivalent to the *biconditional* assertion:

There isn't a triangle if, and only if, there isn't a circle.

Reasoners don't immediately grasp this equivalence—a failure that shows that they rely on mental models, not fully explicit models.

When participants are given a compound assertion, such as a disjunction, and are asked to list what is possible, the principle of truth constrains them, and so they list the possibilities corresponding to mental models (see, e.g. Barres & Johnson-Laird, 2003; Johnson-Laird & Savary, 1995). The advantage of the principle is that it reduces the processing load of reasoning. But, when we implemented the principle in a computer program, we discovered an unexpected downside.

Could both of these disjunctions be true at the same time?

Either the pie is on the table or the cake is on the table, but not both.

Either the pie isn't on the table or the cake is on the table, but not both.

Most people say, "Yes" (Johnson-Laird, Lotstein, & Byrne, 2012). The mental models of what's on the table according to the first disjunction are:

Pie

Cake

And the mental models of what's on the table according to the second disjunction are:

¬ Pie

Cake

The presence of the cake is common to both sets of models, and so it seems that the two assertions can both be true at the same time, that is, when the cake is on the table. In contrast, the fully explicit models of the two disjunctions are as follows:

Pie ¬ Cake

¬ Pie Cake

and:

¬ Pie ¬ Cake

Pie Cake

As readers can see, no model is common to both assertions, and so they cannot both be true at the same time.

The program implementing the model theory predicts these fallacies, and others too. Their occurrence has been corroborated in many sorts of deductions, including those based on:

- Disjunctions and biconditionals of conditionals (Johnson-Laird & Savary, 1999);
- Disjunctions of conjunctions (Walsh & Johnson-Laird, 2004);
- Disjunctions of disjunctions (Khemlani & Johnson-Laird, 2009);
- Disjunctions of quantified assertions (Yang & Johnson-Laird, 2000).

The fallacies tend to be compelling and to elicit judgments of high confidence in their conclusions, and so they have the character of cognitive illusions. Other illusions led to conclusions about what is probable (Johnson-Laird & Savary, 1995), possible (Goldvarg & Johnson-Laird, 2000), and permissible (Bucciarelli & Johnson-Laird, 2005). And still others concerned the evaluation of the consistency of assertions (Legrenzi, Girotto, & Johnson-Laird, 2003). Each study examined several sorts of illusion and matched control problems. Why so many studies of illusions? Because only the model theory predicts them, and so they are a litmus test for the use of mental models.

6. MODELS AS COUNTEREXAMPLES

In reasoning, a counterexample is a possibility that is consistent with a set of premises, but not with a putative conclusion, and so it shows that the conclusion does not follow validly from the premises. The intuitive system can generate at most a single model at a time. To establish the validity of a

conclusion, the deliberative system has to search for alternative models and to show either that no other mental model can be formed from the premises or that the conclusion holds in the alternatives. If the deliberative system creates a model that is a counterexample then it can search for an alternative conclusion that holds in all the models or, if this search fails, declare that no valid conclusion follows from the premises. As we pointed out earlier, a conclusion such as a conjunction of the premises follows validly from any set of premises, and so an alternative conclusion needs to be parsimonious and to establish a new relation not explicitly asserted among the premises. In short, counterexamples are crucial for rationality. Without the ability to create them, individuals can infer conclusions, but they have no ready way to establish their invalidity. So, to what extent do individuals make use of them?

On the one hand, reasoners often fail to use counterexamples when they are drawing conclusions from premises—to the degree that one model-based theory makes no use of them (Polk & Newell, 1995). On the other hand, all the participants in one study spontaneously used them to revise their responses (Bucciarelli & Johnson-Laird, 1999).

There are two sorts of invalid conclusion. One sort contradicts the premises—their respective sets of possibilities are disjoint. The other sort is consistent with the premises, but does not follow from them, that is, there are possibilities to which the premises, but not the conclusion, refer. The model theory predicts that the invalidity of contradictions should be easier to detect than the invalidity of consistent premises: the former don't have a mental model in common with the premises whereas the latter do. The theory also predicts that when individuals are asked to explain why a conclusion does not follow from the premises, they should tend to point out the contradiction in the first case but to exhibit a counterexample in the second case. A study corroborated both of these predictions (Johnson-Laird & Hasson, 2003). The participants were more accurate in identifying invalid inferences in which the conclusion contradicted the premises (92% correct) than those in which the conclusion was consistent with the premises (74% correct). To justify their judgments, they used counterexamples more often for conclusions consistent with the premises (51% of cases) than for conclusions inconsistent with them (21% of cases). Of course, they used other strategies too. One participant, for instance, pointed out that a piece of necessary information was missing from the premises. But, the use of counterexamples correlated with accuracy in the evaluation of the inferences.

An fMRI study contrasted reasoning and mental arithmetic from the same premises (Kroger, Nystrom, Cohen, & Johnson-Laird, 2008).

The participants read a statement of the problem, then three premises, and finally either a conclusion or an arithmetical formula, which they had to evaluate. The experiment included easy inferences that followed immediately from a single premise and difficult inferences that should lead individuals to search for counterexamples, as in this case:

There are five students in a room.
Three or more of these students are joggers.
Three or more of these students are writers.
Three or more of these students are dancers.
Does it follow that at least one of the students in the room is all three: a jogger, a writer, and a dancer?

Most people think first of a possibility in which the conclusion holds. But, those who search for a counterexample may find one, such as this model in which each of the five individuals shown in separate horizontal rows is a student:

Jogger	Writer	
Jogger	Writer	
Jogger		Dancer
	Writer	Dancer
		Dancer

Hence, it doesn't follow that a student is all three. While the participants were reading the premises, the language areas of their brains were active (Broca's and Wernicke's areas), but then other areas carried out the solution to the problems. Right prefrontal cortex and inferior parietal lobe were more active for reasoning than for calculation, whereas regions in left prefrontal cortex and superior parietal lobe were more active for calculation than for reasoning. Right prefrontal cortex—a region known as the right frontal pole—was active only during the difficult inferences calling for a search for counterexamples. Other studies have shown that difficult inferences activate right frontal cortex (Kroger et al., 2002; Waltz et al., 1999). The anterior frontal lobes evolved most recently, they take longest to mature, and their maturation relates to measured intelligence (Shaw et al., 2006). Whether they are activated merely by problems calling for deliberation remains unclear.

7. MODULATION AND THE USE OF KNOWLEDGE

In logic, the interpretation of logical terms is constant, as for its idealized connectives akin to "if", "and", and "or". But, the model theory recognizes that their interpretation in everyday language can be modified by the meanings of the clauses that they connect, the entities referred to in these clauses, and general knowledge. We refer to the process as *modulation*, and we illustrate it with the most notorious case, conditional assertions (see Johnson-Laird & Byrne, 2002).

Conditionals of the grammatical form, *if A then B*, receive a logical interpretation by default. When individuals have to list the fully explicit possibilities to which a conditional refers, they tend to list:

A B

¬ A B

¬ A ¬ B

where *A* and *B* have as values actual propositions (see, e.g. Johnson-Laird & Savary, 1995). Barrouillet and his colleagues have shown that children around the age of 8 years list only one possibility, *A and B*, a conjunctive interpretation; around the age of 11 years, they include another possibility, ¬ *A* and ¬ *B*, a biconditional interpretation; and around the age of 15 years, they list the three possibilities above (Barrouillet & Lecas, 1998). The processing capacity of working memory is a better predictor than chronological age for the number of possibilities that children list (Barrouillet, Grosset, & Lecas, 2000; Barrouillet & Lecas, 1999).

In reasoning, individuals rely on the mental models of conditionals, which consist of an explicit model of the salient case in which both clauses hold, and a content-less placeholder for other possibilities in which the if-clause is false:

A B

. . .

One corollary concerns inferences of the form:

If A then B.

A.

What follows?

Individuals easily infer the conclusion, *B*. It follows at once from the mental models. A contrasting inference is:

If A then B.

Not-B.

What follows?

The second premise eliminates the one explicit mental model, and so it seems that nothing follows—a common response. Only if reasoners flesh out their mental models, or adopt some analogous strategy, can they make the valid inference: *Not-A*. The difference between the two sorts of inference is highly robust. A more striking corroboration of the model theory is that the presentation of the premises in the opposite order improves performance with the difficult inference—it renders unnecessary the need to construct the explicit mental model of the conditional, thus making room for models of the possibilities in which *not-A* holds (see Girotto, Mazzocco, & Tasso, 1997).

Modulation has several effects, and one of them is to block the construction of models of possibilities. A conditional, such as:

If she played a musical instrument then it wasn't the flute

refers to only two possibilities because knowledge that a flute is a musical instrument blocks the construction of the possibility that she didn't play a musical instrument but did play the flute. Hence, the conditional alone yields the conclusion that she didn't play the flute. The principal possibility to which almost all conditionals refer is the one in which both the if-clause and the then-clause hold. Hence, the theory postulates that if a conditional refers to more than one possibility, then this possibility must be one of them. Modulation can accordingly yield the preceding interpretation or the biconditional interpretation. Still other effects of modulation occur with then-clauses that themselves express only a possibility, such as "if Hillary runs then she may win".

Experiments have corroborated that modulation blocks the construction of models (Quelhas, Johnson-Laird, & Juhos, 2010). Consider these two conditionals translated from the Portuguese:

If the dish is lasagne then its basis is pasta.

If the cake is made of eggs then it can be suspiro.

For the first sort of conditional, participants allow that the dish can be pasta but not lasagne. But, for the second sort of conditional, they do not allow that the cake can be suspiro but not made of eggs—all Portuguese know that suspiro is made from eggs. The two sorts of conditionals yield appropriately different patterns of inference.

Consider the inference:

Luisa didn't play soccer.

Therefore, if Luisa played music then she didn't play soccer.

The conditional conclusion refers to three possibilities in which Luisa played, respectively:

Music ¬ Soccer

¬ Music ¬ Soccer

¬ Music Soccer

Hence, the conclusion refers to the possibility to which the premise refers, and so the inference is valid. Yet, most people reject it. Why? One answer is that it is unacceptable because it throws information away, that is, its conclusion also refers to an alternative possibility that conflicts with the premise:

¬ Music Soccer

This conflict, as Orenes and Johnson-Laird (2012) argued, may deter individuals from drawing the inference. If so, then modulation that blocks the conflicting model should yield an acceptable inference, for example:

Luisa didn't play soccer.

Therefore, if Luisa played a game then she didn't play soccer.

The conditional now refers to just two possibilities in which Luisa played:

A game ¬ Soccer

¬ A game ¬ Soccer

The conditional can't refer to the case in which Luisa didn't play a game but played soccer because soccer is a game. So, both the preceding possibilities refer to the same possibility as the premise. In this case, a highly reliable increase occurs in the percentage of participants who accepted the inference. And analogous phenomena occur with inferences to disjunctive conclusions.

Another effect of modulation is to introduce spatial, temporal, or other relations between the if-clause and the then-clause. As a consequence, individuals make different inferences (Quelhas et al., 2010). For example, given these premises:

If Laura got the virus, then she infected Renato.

If she infected Renato, then he went to hospital.

Laura got the virus.

Participants tend to infer that Laura got the virus before Renato went to hospital. But, given these premises:

If Cristina wrote the article, then Marco asked her to write it.

If Marco asked her to write it, then he met her at the meeting.

Cristina wrote the article.

Participants tend to infer that Cristina wrote the article after Marco met her. The temporal inferences depend on the participants' general knowledge about the typical orders of events.

A subtle effect of temporal modulation is illustrated in the following contrasting examples (Juhos, Quelhas, & Johnson-Laird, 2012). The first example is:

If the author writes the book, then the publisher publishes it.

The author writes the book.

What follows?

Individuals tend to infer:

The publisher publishes it.

The second example is:

If the author writes the book, then the publisher publishes it.

The publisher publishes the book.

What follows?

Individuals tend to infer:

The author wrote the book.

The difference is that for the first inference, the participants tended to use the present tense (the experiment was carried out in Portuguese), whereas for the second inference, they tended to use the past tense. As in English, which has no future tense, the present tense in Portuguese can be used to refer to future events. The same phenomenon occurred in inferences from disjunctions. The categorical premise accordingly establishes a reference time, and events prior to it are referred to in the past tense, and events subsequent to it are referred to in the present tense. This sort of modulation is tacit—participants are not usually aware of its effects—but it shows that general knowledge influences the interpretation of conditionals and disjunctions.

A crucial corollary of modulation concerns logical form. A typical formal rule of inference is:

A or B.

Not-B.

Therefore, not-A.

This rule is applicable to any premises that have the corresponding logical forms, which are transparent in logic, because they are defined by its grammar. In language, however, logical forms are far from transparent, and no algorithm exists to determine them because they are not just a matter of grammar. They depend on the possibilities to which assertions refer. So too does validity, and it therefore can be decided only on a case-by-case basis. The model theory makes no use of logical form, but merely the grammatical structure of sentences, and it uses meaning, reference, and knowledge to modulate logical interpretations.

8. INDUCTION AND ABDUCTION

Modulation depends on knowledge, and so it is a bridge from deduction to induction. Inductive inferences yield specific conclusions, generalizations, and, above all, explanations. They depend on knowledge and its availability (Tversky & Kahneman, 1973). An induction may yield a true conclusion; but it may not, even if its premises are true. The engineers in charge of Chernobyl induced that the reactor was intact after the explosion. Their inference was plausible because no nuclear reactor had ever melted down before. But, they were wrong, and their delay in making the correct inference cost lives. Induction is indeed risky.

Logic is "monotonic" in that further premises warrant further conclusions, and no subsequent premise ever calls for a valid conclusion to be withdrawn—not even its direct contradiction. At Chernobyl as in daily life, individuals withdraw conclusions, even valid ones, in the light of subsequent information. Their reasoning is "nonmonotonic". They withdraw some conclusions because they are based on assumptions made by default. They infer, say, that Fido has four legs because Fido is a dog and by default dogs have four legs, but then they discover that poor old Fido has only three legs. This process of withdrawing conclusions based on default assumptions is integral to the model theory (Johnson-Laird & Byrne, 1991). But, retractions also occur in other cases. You believe, say, that if someone pulls the pistol's trigger then it will fire. Someone pulls the trigger. Yet, the pistol does not fire. Hence, there is a conflict between a valid inference from your beliefs—that the pistol fires—and the incontrovertible fact that it doesn't fire. So, you have to withdraw your conclusion and modify at least one of your beliefs. Artificial intelligencers have devised various systems of nonmonotonic reasoning to deal with such cases, but these approaches have grown increasingly remote from psychological plausibility (see Brewka, Dix, & Konolige, 1997). In fact,

at the heart of human performance is the abduction of explanations that resolve inconsistencies (Johnson-Laird, Girotto, & Legrenzi, 2004). It is these explanations that, as a by-product, yield revisions to beliefs.

An inconsistent set of assertions is a potentially serious matter in daily life. For example, disasters at sea are often a consequence of a conflict between a mariner's mental model and reality (Perrow, 1984). The ability to detect inconsistencies is accordingly one hallmark of rationality. Reasoners can do it by trying to construct a single model of all the relevant information. If they succeed, they evaluate the information as consistent; but if they fail, they evaluate it as inconsistent (see Johnson-Laird et al., 2004, for corroboratory evidence). Once they have detected an inconsistency, they can use their knowledge to try to explain it. The rest of this section focuses on such explanations, that is, abductions.

The basic units of explanations are causes and their effects. In the case of an inconsistency, the effect makes possible the facts of the matter. According to the model theory, causation refers to what is possible and to what is impossible in the co-occurrence and temporal sequence of two events (Frosch & Johnson-Laird, 2011; Goldvarg & Johnson-Laird, 2001). A computer program implementing this account constructs mental models of the premises, as in the pistol example, detects the inconsistency, and uses its models of causal relations to build a chain resolving the inconsistency, for example, *a person emptied the pistol and so there were no bullets in the pistol* (Johnson-Laird et al., 2004). Such an explanation is bound to repudiate at least one previous belief, which reasoners can modify to refer to a situation that was once possible, but that did not occur, as in the counterfactual conditional, *if a person hadn't emptied the pistol and there were bullets in the pistol then the pistol would have fired* (see Byrne, 2005). Experimental evidence showed that individuals are usually able to create such explanations, which tend to refute the conditional premise (Johnson-Laird et al., 2004). Individuals rate the cause and effect as more probable than either the cause alone or the effect alone—a fallacy in which a conjunction is wrongly judged to be more probable than its constituents (Tversky & Kahneman, 1983).

A study of abduction (Johnson-Laird et al., 2004) examined such inferences as:

If a pilot falls from a plane without a parachute then the pilot dies.

This pilot didn't die. Why not?

Some participants made a valid deduction:

The pilot didn't fall from a plane without a parachute.

But, other participants made explanatory abductions, such as:

The pilot fell into a deep snowdrift and so wasn't hurt.
The plane was on the ground and he [sic] didn't fall far.
The pilot was already dead.

An inadvertent demonstration of the imaginative power of human abductions used pairs of sentences chosen at random from pairs of stories, also chosen at random (see Johnson-Laird, 2006, chap. 14). The result was pairs of sentences such as:

Celia made her way to a shop that sold TVs.
Maria had just had her ears pierced.

The participants' task was to describe "what is going on" in such scenarios. To the experimenters' surprise, the participants were usually able to comply. The task was easier in another condition in which the sentences were edited minimally to ensure that they both referred to the same individual:

Celia made her way to a shop that sold TVs.
She had just had her ears pierced.

Typical examples of the participants' responses in this case were:

She's getting reception in her earrings and wanted the shop to investigate.
She wanted to see herself wearing earrings on closed-circuit TV.
She won a bet by having her ears pierced, using money to buy a new TV.

What was striking was how rarely individuals were stumped for an explanation. Human reasoners are adept at abductions—they outperform any existing computer program. Their explanatory ability underlies superstitions (Johnson-Laird, 2006, chap. 14). It also underlies science, but scientists test putative explanations: they search for counterexamples.

A long-standing view of a rational reaction to inconsistency is encapsulated in William James's remark: "[The new fact] preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty" (James, 1907, p. 59). Cognitive scientists have often defended the same view (e.g. deKleer, 1986; Gärdenfors, 1992; Harman, 1986; cf. Elio & Pelletier, 1997, for results to the contrary). Naive individuals, however, are much more concerned to *explain* inconsistencies because explanations can help them to decide what to do. They readily sacrifice minimalism for this goal. For instance, when reasoners are asked what follows from inconsistent premises, they spontaneously offer an explanation that resolves the inconsistency, and they judge that such explanations are more probable than revisions to the premises that restore consistency (Khemlani & Johnson-Laird, 2011). Once they have formulated such an explanation, a striking phenomenon occurs. It becomes harder for them to detect the inconsistency in comparison with cases in

which instead of explaining the inconsistency, they rate which assertion is more surprising. They seem to have explained the inconsistency away, perhaps by reinterpreting the generalization in the premises as holding by default so that it is less vulnerable to contrary facts (Khemlani & Johnson-Laird, 2012b). In sum, reasoners are able to resolve inconsistencies. They tend to do so by using knowledge to abduce causal models that explain the origins of the conflicts. This reasoning usually makes sense of the inconsistency, although on some occasions it fails to yield any explanation whatsoever.

9. PROBABILITIES: EXTENSIONAL AND INTENSIONAL

Induction is often uncertain, and uncertainty implies probability. Individuals who know nothing of the probability calculus happily infer probabilities. How they make such inferences should be part of a unified theory of reasoning. Following Tversky & Kahneman (1983), we distinguish between *extensional* reasoning in which the probability of an event is inferred from the different mutually exclusive ways in which it can occur and *nonextensional* (intensional) reasoning in which the probability of an event is inferred from some relevant evidence or index. In principle, extensional reasoning is deductive, whereas nonextensional reasoning is inductive, and much of it, as Tversky & Kahneman (1973, 1983) showed, depends on heuristics. In daily life, probabilistic reasoning may mix extensional and nonextensional processes. Hence, we explain how the model theory applies to both of them.

Mental models represent possibilities, and so simple extensional inferences can be made on the assumption that each possibility is equiprobable barring evidence to the contrary (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). The probability of an event is accordingly the proportion of models in which it holds. The theory allows that models can also be tagged with numerals denoting their probabilities. Similar principles underlie other theories of probabilistic reasoning (e.g. Falk, 1992; Shimojo & Ichikawa, 1989). The model theory, however, assigns equiprobability, not to events but to models of events. Classical probability theorists, such as de Laplace (1995); (originally published in 1819), advocated an analogous principle of "indifference", but ran into difficulty because events can be partitioned in different conflicting ways (Hacking, 1975). Mental models merely reflect the probabilities that the individual constructing them assigns to events: different individuals can therefore partition events in different ways without self-contradiction.

A simple extensional problem (from Johnson-Laird et al., 1999) is:

In the box, there is a green ball or a blue ball or both.

What is the probability that both the green and the blue balls are there? The mental models of the premise are:

Green

Blue

Green Blue

Hence, the equiprobability principle predicts correctly that individuals will tend to estimate the probability of *green and blue* as 1/3. But, a telltale sign of mental models is that individuals succumb to illusory inferences about probabilities. Here is an example:

> There is a box in which there is at least a red marble or else there is a green marble and there is a blue marble, but not all three marbles. What is the probability that there is a red marble and a blue marble in the box?

The mental models of the premise represent two possibilities:

Red

Green Blue

They imply that the red and blue marbles cannot occur together, and so their probability is zero. And most people make this estimate. However, the disjunction means that when its first clause is true, its second clause is false, and it can be false in three ways:

Red Green ¬ Blue

¬ Green Blue

¬ Green ¬ Blue

When the second clause of the premise is true, the first clause is false:

¬ Red Green Blue

Hence, there are four distinct possibilities for what's in the box, and, on the assumption of equiprobability, the probability of green and blue is, not zero, but 25%. The participants in an experiment performed much better with the control problems than with the illusory problems of this sort (Johnson-Laird et al., 1999).

Turning to nonextensional reasoning, consider this question about a unique event:

What is the probability that Hillary Clinton is elected US President in 2016?

Some psychologists argue that such probabilities are meaningless because probabilities concern only the natural frequencies with which events occur (e.g. Cosmides & Tooby, 1996). However, naive reasoners are happy to make estimates for unique events, and, as we have observed, their estimates correlate reliably over different contents, ranging from politics to climate (Khemlani et al., 2012). The psychological mystery about such estimates is what mental processes underlie them, and, in particular, where do the numbers come from?

We proposed a dual process theory (see Section 3) in which the intuitive system given, say, the question about Hillary, adduces evidence, such as: *Hillary was a very effective Senator; and many effective Senators have become President*. It uses a mental model representing this evidence to construct a primitive non-numerical representation of a degree of belief in the proposition. This iconic representation is akin to the following sort of diagram:

in which the left vertical represents impossibility, the right vertical represents certainty, and the pointer at the end of the line corresponds to the strength of the particular belief, such as, *Hillary will be elected President*. The intuitive system can translate this representation into the sorts of description that a non-numerate individual would use, such as: "it's as likely as not".

The deliberative system can map the degrees of belief represented in an icon into a numerical estimate. Because this system has access to working memory, it can carry out proper arithmetical operations. It can also try to keep track of the complete joint probability distribution (the JPD). Given two unique events, such as the election of Clinton in 2016 and the Democrats gaining control of Congress, the JPD consists in the set of probabilities for each possible combination of the affirmations and negations of the relevant propositions:

Hillary is President & Democrats control Congress	35%
Hillary is President & not (Democrats control Congress)	30%
Not(Hillary is President) & Democrats control Congress	15%
Not(Hillary is President) & not(Democrats control Congress)	20%

The JPD provides all the information needed to estimate any probability concerning the domain. There are many different sets of probabilities from

which the values of the JPD can be inferred. For instance, if you know the values of the three probabilities in each of the following triples then, in principle, you can infer that values of the probabilities in the JPD, where, say, *A* denotes "Hillary is President" and *B* denotes "Democrats control Congress":

P(A), P(B), P(A and B)
P(A), P(B), P(A or B, or both)
P(A), P(B), P(A | B)

The last of these triples includes P(A | B), which is the conditional probability of *B* on the assumption that *A* occurs.

Granted the limited ability of the intuitive system to carry out loops of operations (see Section 3), it is capable of only a small number of primitive analogs of arithmetical operations of the sort found in infants (Barth et al., 2006; Dehaene, 1997; Xu & Spelke, 2000) and adults in non-numerate cultures (Gordon, 2004). It can add two pointers, subtract one from another, take their mean, and multiply a proportion signified by one pointer by another—all within the bounds between certainty and impossibility and all in crude error-prone ways. The theory accordingly postulates that to estimate the probability of a conjunction of events, reasoners should tend to split the difference between them, but some may take the proportion of a proportion. The latter is a more complex operation (in terms of Kolmogorov complexity, see Li & Vitányi, 1997), and so it should tend to be used less often. Reasoners should likewise make analogous inferences in estimating conditional and disjunctive probabilities.

We implemented the intuitive and deliberative systems in a computer model and tested its predictions in experiments (Khemlani et al., 2012). The results showed that the participants concurred in the rank orders of their estimates of the probabilities of unique events. For example, they agreed that the US is more likely to make English the official language of the country (the average estimate was a probability of 46%) than to adopt an open border policy (an average estimate was a probability of 15%). Hence, they are to some extent relying on knowledge and mental processes in common. They tended to estimate the probability of a conjunction by taking the mean of their estimates of the probabilities of its conjuncts. This tendency was even evident in the overall means, for example, their mean estimate of the conjunction of the US adopting an open border policy and making English the official language was 26%, a

value falling between their mean estimates of the two conjuncts. It yields a violation of the JPD, that is, the negative probability in the third conjunction shown here:

English	Open borders:	26%
English	¬ Open borders:	20%
¬ English	Open borders:	-11%
¬ English	¬ Open borders:	65%

Violations of the JPD, however, were smaller when the conjunction came last as opposed to first in the sequence of judgments. When it was last, the participants had already made numerical estimates of the probabilities of its conjuncts, and so they could use a deliberative procedure, such as taking a proportion of a proportion. This method is appropriate only for independent events, but a prior study established that they were not independent.

The model theory of probabilities dispels some common misconceptions. Probabilistic reasoning isn't always inductive. Extensional estimates can be deductively valid, but they can also yield illusory values. Likewise, nonextensional estimates of unique events depend on intuitions, and the resulting violations of the JPD suggest that the probability calculus is not native to human cognition. Individuals simulate events, but their restricted repertoire of intuitive methods leads them into error.

10. MENTAL SIMULATIONS AND INFORMAL PROGRAMS

Is there one sort of thinking that depends on mental simulation and that cannot be explained in any other terms? In our view, there is. It is the thinking that underlies the creation of algorithms and computer programs. Expert programming is an intellectual discipline that depends

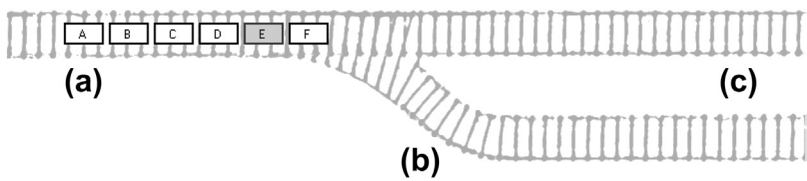

Figure 1.1 The railway domain with an example of an initial configuration in which a set of cars is on the left side (a) of the track, the siding (b) can hold one or more cars while other cars are moved to the right side of the track (c).

on knowledge of programming languages. Hence, our studies focused on how nonprogrammers formulate algorithms in everyday language (Khemlani & Johnson-Laird, 2013). To make the task easy, the programs concerned the railway domain shown in Figure 1.1. The participants had to rearrange the cars in a train on the left track using the siding so that they arrived on the right track in the required new order. In the terminology of automata theory, the siding acts as a "stack" on which to store cars temporarily. Items on the siding move back to the left track, which therefore also functions as a stack. Automata with two stacks are equivalent to Universal Turing machines. Hence, if cars can also be added and removed, the railway serves as a general-purpose computer (Hopcroft & Ullman, 1979).

Most people can solve rearrangement problems in the railway domain. They use a simple variant of "means-ends" analysis in which they work backward from the required goal, invoking operations relevant to reducing the difference between the current state and the goal (e.g. Newell, 1990, Newell & Simon, 1972). For rearrangement problems, they need only envisage each successive car in the goal. Suppose, for instance, they have to rearrange the order ABCD into ACBD. The starting state is:

ABCD[ ]

where the square brackets denote the contents of the siding, which is empty at the start. Their immediate goal is to get D to the far end of the right track:

[ ]…D

So, they move D from left to right track:

ABC[ ]D

The next partial goal is to get B to the right track, and so they need to move C out of the way onto the siding:

AB[C]D

Now, they can move B to the right:

A[C]BD

They move C off the stack:

AC[ ]BD

The next move is intriguing. They should move both A and C together from left to right track. But, if reasoners perseverate, they may move only C to the right track. Their solution won't be minimal because they then have to make a separate move of A to right track. Our initial study investigated

all 24 possible rearrangements of four cars, and the participants easily solved each of them, but they did tend to perseverate: every participant made one or more unnecessary moves.

In the principal experiment, the participants, who were not programmers, had to formulate algorithms for three sorts of rearrangement: *reversals* in which, say, ABCDEFGH on left track becomes HGFEDCBA on right track; *palindromes* in which, say, ABCDDCBA becomes AABBCCDD, and *parity sorts* in which, say, ABCDEFGH becomes ACEGBDFH, that is, cars in odd-numbered positions precede those in even-numbered positions. Each solution calls for recursion, that is, a loop of operations. *Primitive* recursion in the theory of recursive functions corresponds to a loop carried out *for* a given number of times, a so-called "for-loop", whereas *minimization* corresponds to a loop carried out *while* a given condition holds, a "while-loop" (see Rogers, 1967). While-loops are more powerful than for-loops because only they can compute certain functions. Indeed, when a while-loop is entered, there may be no way to determine how many times it will repeat before it yields an output or whether it will ever halt to yield an output. But, how do nonprogrammers formulate informal algorithms? The task isn't deductive: they can deduce the consequences of a program, but they can't create it using deduction alone (Kitzelmann, Schmidt, Mühlpfordt, & Wysotzki, 2002). Likewise, they don't rely on probabilities any more than they do for Sudoku puzzles (Lee et al., 2008). The one viable method is to simulate a solution to a problem, observe what happens in the simulation, and translate these observations into a description. The simulation depends on a kinematic sequence of mental models representing successive states of the world, real or imaginary (Johnson-Laird, 1983, chap. 15).

Let's examine the process in more detail. The first step is to solve two different examples of the relevant rearrangement problem. Without two examples differing in numbers of cars, rearrangements are ambiguous. The solution to reversing a train of four cars is as follows:

ABCD[ ], A[BCD], [BCD]A, B[CD]A, [CD]BA, C[D]BA, [D]CBA, D[ ]CBA, [ ]DCBA

As this protocol illustrates, only three sorts of move are possible, and they occur in these summaries of simulations that solve reversals of trains of four and five cars:

S3 R1 L1 R1 L1 R1 L1 R1

S4 R1 L1 R1 L1 R1 L1 R1 L1 R1

where "S3" means move three cars to the siding from left track, "R1" means move one car to right track from left track, and "L1" means move one car to left track from the siding. The second step uses the two summaries to work out the loop of moves they contain and any moves before or after it (pace Miller, 1974,1981; Pane, Ratanamahatana, & Myers, 2001). The loop in the simulations above is (R1 L1). But, how many times should it be iterated? There are two ways to answer this question, depending on whether reasoners are formulating a while-loop or a for-loop. The simpler way is to observe the conditions in the simulation when the loop halts, which are respectively:

D[ ]CBA
E[ ]DCBA

The condition that halts the loop is that no cars are left on the siding, and so the while-loop should continue as long as the siding isn't empty. The alternative is to compute the number of times that a for-loop should be executed. It calls for the solution of a pair of simultaneous linear equations to obtain the values of *a* and *b* in:

Number of iterations = $a \times$ length of the train + b.

The final step maps the structure of the solution into an informal description. We implemented this entire process in a computer program, which constructs programs for any rearrangement problem based on a single loop. It produces a for-loop and a while-loop in Lisp and translates the while-loop into informal English. Each of these functions solves any instance of the relevant class of rearrangements. Table 1.1 presents its solutions for the three sorts of rearrangement: reversals, palindromes, and parity sorts.

If individuals use simulation to devise algorithms, then they should tend to use while-loops rather than for-loops because it is easier to observe the halting condition of a while-loop than to solve simultaneous equations. The overall difficulty of formulating an algorithm should depend on its Kolmogorov complexity, which is the length of its shortest description in a given language, such as Lisp (Li & Vitányi, 1997). A good proxy is the number of instructions. In Table 1.1, the functions for reversals and palindromes call for four instructions, whereas parity sorts call for five instructions. Within a given level of complexity, another factor should also affect difficulty: the mean number of operands (i.e., cars) per move. This measure distinguishes reversals, which have 1.38 operands per move for eight cars, from palindromes, which have 1.75 operands per move for eight cars. Hence, the three sorts of problem should increase in difficulty from reversals through palindromes to parity sorts.

Table 1.1 Loops for Computing Minimal Solutions to Three Sorts of General Problem: Reversals, Palindromes, and Parity Sorts Using "for"-loops and "while"-loops and Their Informal Description (from the Output of the Computer Program *mReasoner* for Abducing them)

	While-loops	
For-loops	**Lisp**	**Informal English**
a) Reversals (e.g., ABCDEFGH ⇒ HGFEDCBA)		
(setf track (S (+ (* 1 len) -1) track)) (loop for i from 1 to (+ (* 1 len) -1) do (setf track (R 1 track)) (setf track (L 1 track))) (setf track (R 1 track))	(setf track (S (+ (* 1 len) -1) track)) (loop while (> (length (second track)) 0) do (setf track (R 1 track)) (setf track (L 1 track))) (setf track (R 1 track))	Move one less than the cars to siding. While there are more than zero cars on siding. Move one car to right track. Move one car to left track. Move one car to right track.
b) Palindromes (e.g., ABCDDCBA ⇒ AABBCCDD)		
(setf track (S (+ (* ½ len) -1) track)) (loop for i from 1 to (+ (* 1/2 len) -1) do (setf track (R 2 track)) (setf track (L 1 track))) (setf track (R 2 track))	(setf track (S (+ (* ½ len) -1) track)) (loop while (> (length(first track)) 2) do (setf track (R 2 track)) (setf track (L 1 track))) (setf track (R 2 track))	Move one less than half the cars to siding. While there are more than two cars on left track. Move two cars to right track. Move one car to left track. Move two cars to right track.

(Continued)

Table 1.1 Loops for Computing Minimal Solutions to Three Sorts of General Problem: Reversals, Palindromes, and Parity Sorts Using "for"-loops and "while"-loops and Their Informal Description (from the Output of the Computer Program *mReasoner* for Abducing them)—(Cont'd)

	While-loops	
For-loops	**Lisp**	**Informal English**
c) Parity sorts (e.g., ABCDEFGH ⇒ ACEGBDFH)		
(loop for i from 1 to (+ (* ½ len) -1) do (setf track (R 1 track)) (setf track (S 1 track)) (setf track (R 1 track)) (setf track (L (+ (* ½ len) -1) track)) (setf track (R (+ (* ½ len) 0) track))	(loop while (> (length (first track)) 2) do (setf track (R 1 track)) (setf track (S 1 track))) (setf track (R 1 track)) (setf track (L (+ (* ½ len) -1) track)) (setf track (R (+ (* ½ len) 0) track))	While there are more than two cars on left track. Move one car to right track. Move one car to siding. Move one car to right track. Move one less than half the cars to left track. Move half the cars to right track.

Our experiment corroborated the predictions. Individuals were able to create informal algorithms, even though they had no access to the railway domain while they carried out the task. They formulated algorithms for the three sorts of problems, once for trains of eight carriages, and once for trains of any length, in a counterbalanced order. Performance with trains of eight cars was at ceiling, but with trains of any length, it corroborated the predicted trend in accuracy and in time. Likewise, the participants used many more while-loops than for-loops. The use of while-loops correlated with accuracy, whereas the use of for-loops had a negative correlation with accuracy. Differences in ability were striking: the best participant was correct on every problem, whereas the worst participant was correct for only a third of the eight-car problems and for none of problems with trains of any length.

The ability to create algorithms is useful in daily life: loops of operations are ubiquitous in everything from laying a table to shutting down a nuclear reactor. Intelligent individuals are able to carry out this task, and our results corroborated an account that bases their thinking on the ability to make simulations. It is difficult to see how else they could create programs.

11. TOWARD A UNIFIED THEORY

We have now described the main principles of the model theory, illustrated them in various sorts of reasoning, and outlined programs implementing special cases of the theory. But, these programs are a dozen separate pieces, often employing ad hoc notations, and so an urgent task is to integrate them. We have therefore begun to integrate the disparate parts within a single unified theory, implementing it in a large-scale computer program, *mReasoner* (available at http://mentalmodels.princeton.edu/models/mreasoner/). The problem is to bring together various sorts of reasoning (e.g. relational, sentential, modal, causal, quantificational) with various sorts of task (e.g. formulating conclusions, evaluating given conclusions, evaluating consistency) in a way that predicts various sorts of phenomena (e.g. accuracy, latency, effects of modulation). Our aim here was to describe the more important insights that have emerged so far.

A mental model can represent the spatial relations among a triangle, circle, and square (as in Section 4), but the size of the figures, their distance

apart, and so on, may not be intended to represent their real sizes or distances apart. Analogous issues occur with mental models of other sorts of assertion. For instance, a quantified assertion, such as:

Some of the actors are bakers

has the following sort of iconic model shown in this diagram of four individuals:

Actor Baker

Actor Baker

Actor

Baker

The numbers of mental tokens in this case are not intended to represent the actual numbers of actors or bakers. Only the overlap between the two sets is iconic. When reasoners search for an alternative model of a set of premises, they can modify all but the essentials of a model. So, how does the system keep track of the essentials? A single uniform answer is that it relies on the meanings of assertions. Hence, *mReasoner* uses a grammar, a lexicon, and a parser to construct representations of meanings, that is, *intensional* representations (for an account of their construction, see Khemlani, Lotstein, & Johnson-Laird, submitted for publication). They are then used to build *extensional* representations, that is, mental models. Both sorts of representations are crucial in reasoning, and to illustrate this point, we consider the intuitive and deliberative systems in reasoning from quantified assertions.

The model theory treats the intensions of quantified assertions as relations between sets (see Boole, 1854; Cohen & Nagel, 1934). The advantages of this treatment are twofold. First, it dovetails with a long-standing treatment of quantifiers in the model theory (Johnson-Laird, 1983, chap. 15): a set is represented iconically as a set of mental tokens, and a quantified assertion is represented as a relation between such sets. Second, it works for all quantifiers in natural language, including those such as "more than half of the artists", which cannot be defined using the quantifiers that range over entities in logic (Barwise & Cooper, 1981). Here are some illustrative examples of this treatment of quantifiers, which the intensions of assertions capture:

All A are B	$A \subseteq B$	(A is included in B.)
Some A are B	$A \cap B \neq \varnothing$	(Intersection of A and B is not empty.)
No A is a B	$A \cap B = \varnothing$	(Intersection of A and B is empty.)
Some A are not B	$A - B \neq \varnothing$	(Set of A that are not B is not empty.)
Most A are B	$\lvert A \cap B \rvert > \lvert A - B \rvert$	(Cardinality of intersection > cardinality of A that are not B.)
More than half the A are B	$\lvert A \cap B \rvert > \lvert A \rvert / 2$	(Cardinality of intersection > ½ of cardinality of A.)

A corollary is that determiners, such as "most", have parameters specifying such matters as the minimal cardinality of A, the cardinality of the quantifier, "most A", constraints on the relation between the two cardinalities, and so on.

The intuitive system can construct a model of an assertion and update the model according to subsequent assertions. Hence, given the premises of a syllogism:

Some of the actors are bakers.

All the bakers are colleagues.

It updates the model above to:

Actor	Baker	Colleague
Actor	Baker	Colleague
Actor		
	Baker	Colleague

In general, affirmative premises are added so as to minimize the number of distinct sorts of individual, whereas negative premises are added so as to maximize the number of distinct sorts of individual.

Once the intuitive system has an initial model, it can draw a conclusion establishing a new set-theoretic relation, that is, a relation that is not asserted in the premises. In the past, the model theory has eschewed heuristics, but it now embodies them to frame both the quantifier in the conclusion (its mood) and the order of the terms that occur in it: "actors" and "colleagues" (its figure). When two premises differ in mood, one of them dominates the other in determining the mood and figure of the initial conclusion. The order of dominance reflects two principles governing valid inferences:

- A negative premise can yield only valid conclusions that are negative.
- Within both negative and affirmatives, a particular premise—one based, for instance, on "some"—can yield only valid conclusions that are particular.

The resulting order of dominance for syllogisms is accordingly:

Some _ are not _ > No _ are _ > Some _ are _ > All _ are_

In our example, the first premise is dominant: "artists" is its subject, and so "artists" is the subject of the conclusion, and the term in the other premise, "colleagues", is in the predicate of the conclusion, that is, "some of the artists are colleagues". Analogous principles apply to other sorts of premise. They account for the well-known figural effect that occurs in syllogistic reasoning, for example, the tendency to infer the conclusion above rather than the converse, "some colleagues are artists" (see, e.g. Bucciarelli & Johnson-Laird, 1999). The order of dominance is the same as the order invoked in Chater and Oaksford (1999) from probabilistic considerations. But, since our principles derive from valid inferences, and yield only conclusions that hold in initial mental models, they do not depend on probabilities. The heuristics operate without storing any information in working memory, and so they are rapid, but fallible.

The deliberative system makes a recursive search for alternative models falsifying an initial conclusion. When the system finds a counterexample, it formulates a new conclusion if one is possible or else declares that no definite conclusion follows about the relation between the end terms. It searches for counterexamples using the same operations as did participants working with external models in the form of cutout shapes (Bucciarelli & Johnson-Laird, 1999): adding a new individual to a model, breaking an individual into two, and moving a property from one individual to another. As a meta-analysis showed, the resulting theory embodied in *mReasoner* outperforms all current theories of syllogistic reasoning (see Khemlani & Johnson-Laird, 2012c; Khemlani et al., submitted for publication).

The implementation of the unified theory has so far focused on unique probabilities and quantified assertions. It draws its own conclusions from quantified premises, evaluates given conclusions about what is necessary or about what is possible, formulates counterexamples to putative conclusions, and evaluates whether or not a set of quantified assertions is consistent. It carries out these tasks using both the intuitive system, which builds initial models, and the deliberative system, which searches for alternative models. Hence, it is able to predict human inferences. It provides the beginnings of a unified computational account, which we are extending to accommodate sentential and relational reasoning.

12. CONCLUSIONS

The psychology of reasoning would have been simpler if human beings were logicians or probabilists. Logic and the probability calculus are not native mental faculties but cultural discoveries. Some individuals master these technologies; some do not. And, in our culture, most individuals have smatterings of them at best. As the model theory predicts, deductive and probabilistic inferences are difficult and fallible. An awareness of the occurrence of errors led Aristotle and his intellectual descendants to devise logical and probabilistic calculi. These technologies, however, are unlikely foundations for human reasoning. So, what is? We have argued that it is mental simulation. Reasoners build models of premises and base their inferences on them. This view seems undeniable for reasoning that creates informal algorithms. The evidence we have presented shows that it applies also to all the main domains and tasks of reasoning, from deductions based on sentential connectives to inductions about the probabilities of unique events. But, its manifold applications are a source of its main weakness—its potential disintegration into a bunch of separate subtheories. Their unification is viable because each subtheory is constrained by the main principles of the theory. What is much harder is to implement a computer program that predicts the responses that reasoners make to any inferential task, but *mReasoner* is a step toward that goal.

ACKNOWLEDGMENTS

This research was supported by a National Science Foundation Grant No. SES 0844851 to the first author to study deductive and probabilistic reasoning and by a National Research Council Research Associateship to the second author. We are grateful to more colleagues than we can name here, but many of them can be found among the references.

REFERENCES

Bar-Hillel, Y. (1964). *Language and information processing*. Reading, MA: Addison-Wesley.

Barnes, J. (Ed.), (1984). *The complete works of Aristotle*. Princeton, NJ: Princeton University Press.

Barres, P., & Johnson-Laird, P. N. (2003). On imagining what is true (and what is false). *Thinking & Reasoning*, *9*, 1–42.

Barrouillet, P., & Lecas, J.-F. (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, *67*, 209–253.

Barrouillet, P., & Lecas, J.-F. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning*, *5*, 289–302.

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

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. S. (2006). Nonsymbolic arithmetic in adults and young children. *Cognition*, *98*, 199–222.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

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

Boole, G. (1854). *An investigation of the laws of thought*. London: Macmillan.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*, 1–21.

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning: An overview*. Stanford, CA: CLSI Publications, Stanford University.

Bryant, P. E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, *232*, 456–458.

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

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

Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory & Language*, *28*, 564–575.

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

Carreiras, M., & Santamaría, C. (1997). Reasoning about relations: spatial and nonspatial problems. *Thinking & Reasoning*, *3*, 191–208.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*, 191–258.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method*. London: Routledge & Kegan Paul.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317–370.

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

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *Evolutionary psychology handbook* (pp. 584–627). New York: Wiley.

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

Dehaene, S. (1997). *The number sense*. Oxford, UK: Oxford University Press.

deKleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, *28*, 127–162.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, *21*, 419–460.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Evans, J. St. B.T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *18*, 5–31.
Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition*, *43*, 197–223.
Frosch, C.A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*, 280–291.
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.
Gärdenfors, P. (1992). Belief revision: an introduction. In P. Gärdenfors (Ed.), *Belief revision* (pp. 1–20). Cambridge, UK: Cambridge University Press.
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. (2000). Illusions in modal reasoning. *Memory & Cognition*, *28*, 282–294.
Goldvarg, Y., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.
Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, *112*, 468–493.
Goodwin, G. P., & Johnson-Laird, P. N. (2008). Transitive and pseudo-transitive inferences. *Cognition*, *108*, 320–352.
Gordon, P. (2004). Numerical cognition without words: evidence from Amazonia. *Science*, *306*, 496–499.
Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.
Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press, Bradford Book.
Henle, M. (1962). On the relation between logic and thinking. *Psychological Review*, *69*, 366–378.
Hopcroft, J. E., & Ullman, J. D. (1979). *Formal languages and their relation to automata*. Reading, MA: Addison-Wesley.
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge, Chapman & Hall.
Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, *35*, 2075–2087.
James, W. (1907). *Pragmatism—a new name for some old ways of thinking*. New York: Longmans, Green.
Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York, NY: McGraw-Hill.
Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
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., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition*, *31*, 1105–1113.
Johnson-Laird, P. N., & Savary, F. (1995). How to make the impossible seem probable. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the seventeenth annual conference of the cognitive science society*. Mahwah, NJ: Erlbaum.
Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: a novel class of erroneous deductions. *Cognition*, *71*, 191–229.
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., 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., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640–661.

Johnson-Laird, P. N., Lotstein, M., & Byrne, R. M. J. (2012). The consistency of disjunctive assertions. *Memory & Cognition*, *40*, 769–778.
Johnson-Laird, P. N. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 7–54). Springdale, NJ: Erlbaum.
Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press. (Cambridge, MA: Harvard University Press).
Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 179–212). New York: Psychology Press.
Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
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*. New York, NY: Farrar, Strauss, Giroux.
Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition*, *37*, 615–623.
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. (2012a). The processes of inference. *Argument and Computation*, 1–17. (iFirst).
Khemlani, S., & Johnson-Laird, P. N. (2012b). Hidden conflicts: explanations make inconsistencies harder to detect. *Acta Psychologica*, *139*, 486–491.
Khemlani, S., & Johnson-Laird, P. N. (2012c). Theories of the syllogism: a meta-analysis. *Psychological Bulletin*, *138*, 427–457.
Khemlani, S., & Johnson-Laird, P.N. (2013). *Mental simulation and the construction of informal algorithms*. In M. Knauff M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2012). The probabilities of unique events. *PLoS ONE*, 7, 1–9. (Online version).
Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. *A unified theory of syllogistic reasoning*, submitted for publication.
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.
Kitzelmann, E., Schmidt, U., Mühlpfordt, M., & Wysotzki, F. (2002). Inductive synthesis of functional programs. In J. Calmet B. Benhamou, et al. (Eds.), *Artificial intelligence, automated reasoning, and symbolic computation* (pp. 26–37). New York: Springer.
Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, *30*, 363–371.
Knauff, M., & Ragni, M. (2011). Cross-cultural preferences in spatial reasoning. *Journal of Cognition and Culture*, *11*, 1–21.
Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, *15*, 559–573.
Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.
Kosslyn, S. M. (1994). *Image and brain*. Cambridge, MA: MIT Press.
Kroger, J. K., et al. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cerebral Cortex*, *12*, 477–485.
Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, *1243*, 86–103.
Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., & Knauff, M. (2011). A model for relational reasoning as verbal reasoning. *Cognitive Systems Research*, *11*, 377–392.
de Laplace, P.-S. (1995). *Philosophical essay on probabilities*. New York: Springer- Verlag. (Originally published in 1819).

Lee, N.Y. L., Goodwin, G. P., & Johnson-Laird, P. N. (2008). The psychological problem of sudoku. *Thinking & Reasoning*, *14*, 342–364.
Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science*, *14*, 131–137.
Leibniz, G. W. (1685, 1952). The art of discovery. In P. P. Weiner (Ed.), *Selections—Gottfried Wilhelm Leibniz*. New York: Charles Scribners. (Originally published 1685).
Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer-Verlag.
Miller, L. (1974). Programming by non-programmers. *International Journal of Man–Machine Studies*, *6*, 237–260.
Miller, L. (1981). Natural language programming: styles, strategies, and contrasts. *IBM Systems Journal*, *20*, 184–215.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford: Oxford University Press.
Orenes, I., & Johnson-Laird, P. N. (2012). Logic, models, and paradoxical inferences. *Mind & Language*, *27*, 357–377.
Osherson, D. N. (1974). 1976). *Logical abilities in children*. (Vols. 1–4). Hillsdale, NJ: Erlbaum.
Pane, J. F., Ratanamahatana, C. A., & Myers, B. A. (2001). Studying the language and structure in non-programmers' solutions to programming problems. *International Journal of Human-Computer Studies*, *54*, 237–264.
Peirce, C. S. (1931–1958). Collected papers of Charles Sanders Peirce. (Vols. 8). C. Hartshorne, P. Weiss, & A. Burks, (Eds.), Cambridge, MA: Harvard University Press.
Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York: Basic Books.
Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*, 533–566.
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.
Reitman, W. R. (1965). *Cognition and thought*. New York: Wiley.
Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
Rogers, H. (1967). *Theory of recursive functions and effective computability*. New York: McGraw-Hill.
Schaeken, W. S., Johnson-Laird, P. N., & d'Ydewalle, G. (1996a). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.
Schaeken, W. S., Johnson-Laird, P. N., & d'Ydewalle, G. (1996b). Tense, aspect, and temporal reasoning. *Thinking & Reasoning*, *2*, 309–327.
Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, *440*, 676–679.
Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: theoretical and experimental analyses of the "problem of three prisoners". *Cognition*, *32*, 1–24.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.
Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.
Störring, G. (1908). Experimentelle Untersuchungen über einfachen Schlussprozesse. *Archiv für die gesamte Psychologie*, *11*, 1–27.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*, 292–315.

Vandierendonck, A., Dierckx, V., & De Vooght, G. (2004). Mental model construction in linear reasoning: evidence for the construction of initial annotated models. *Quarterly Journal of Experimental Psychology, A*, *57*, 1369–1391.

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, *11*, 278–293.

Walsh, C., & Johnson-Laird, P. N. (2004). Co-reference and reasoning. *Memory & Cognition*, *32*, 96–106.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119–125.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74B*, 1–11.

Yang, Y., & Johnson-Laird, P. N. (2000). Illusions in quantified reasoning: how to make the impossible seem possible, and vice versa. *Memory & Cognition*, *28*, 452–465.