Minds & Machines
DOI 10.1007/s11023-017-9421-x

# Illusions in Reasoning

**Sangeet S. Khemlani**[1] · **P. N. Johnson-Laird**[2,3]

Received: 26 February 2016 / Accepted: 17 January 2017

**Abstract** Some philosophers argue that the principles of human reasoning are impeccable, and that mistakes are no more than momentary lapses in "information processing". This article makes a case to the contrary. It shows that human reasoners commit systematic fallacies. The theory of mental models predicts these errors. It postulates that individuals construct mental models of the possibilities to which the premises of an inference refer. But, their models usually represent what is true in a possibility, not what is false. This procedure reduces the load on working memory, and for the most part it yields valid inferences. However, as a computer program implementing the theory revealed, it leads to fallacious conclusions for certain inferences—those for which it is crucial to represent what is false in a possibility. Experiments demonstrate the variety of these fallacies and contrast them with control problems, which reasoners tend to get right. The fallacies can be compelling illusions, and they occur in reasoning based on sentential connectives such as "if" and "or", quantifiers such as "all the artists" and "some of the artists", on deontic relations such as "permitted" and "obligated", and causal relations such as "causes" and "allows". After we have reviewed the principal results, we consider the potential for alternative accounts to explain these illusory inferences. And we show how the illusions illuminate the nature of human rationality.

**Keywords** Illusory inferences · Mental models · Deduction · Rationality

✉ Sangeet S. Khemlani
skhemlani@gmail.com

[1] Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence, 4555 Overlook Drive, Washington, DC 20375, USA

[2] Naval Research Laboratory, Princeton University, Princeton, NJ, USA

[3] Department of Psychology, New York University, New York, NY, USA

Published online: 10 February 2017

> I have never found errors which could unambiguously be attributed to faulty reasoning.
>
> – Henle 1978, p. xviii

> Of course, various kinds of mistakes are frequently made in human reasoning, both by laboratory subjects and in ordinary life. But in all such cases some malfunction of an information-processing mechanism has to be inferred, and its explanation sought…. Our fellow humans have to be attributed a competence for reasoning validly, and this provides the backcloth against which we can study defects in their actual performance.
>
> – Cohen 1981, p. 317

Suppose you are a mechanic who knows that:

If oil in the engine is burning then blue smoke will appear.

You observe that blue smoke does not appear. It seems obvious that the engine is not burning oil, but many reasoners fail to make inferences of the following sort:

If A then B.
Not B.
Therefore, not A.

where *A* can stand for "oil in the engine is burning" and *B* can stand for "there is blue smoke". For several years, one of the present authors gave groups of engineering students a similar inference with an abstract content. A substantial minority always responded, "nothing follows". People, it seemed, do make mistakes in reasoning. Yet, a long-standing view to the contrary is that systematic errors are impossible. This idea goes back to the nineteenth century dogma that logic states the laws of thought. The epigraphs to our paper are two twentieth century versions of the same idea.

Reasoners do indeed make all sorts of errors—they misread, misunderstand, or misremember premises, they import their own premises, they misinterpret the task, they get distracted, and so on and on. As a result their performance fails to reflect their true competence. Aside from these slips, the question remains: do human beings reason validly granted their interpretation of premises, which perforce must be established independently? That is, are they rational in their deductions in that they make valid inferences? Our aim in what follows is to answer this question. Readers will notice that we do not refer to logic in framing the question, because validity is definable without reference to logic: a valid inference is one in which the conclusion is true in all the possibilities to which the premises refer (cf. Jeffrey 1981, p. 1).

Our article has three parts. First, we outline a psychological theory of what is computed when people reason, and of how it is computed. This theory predicts that, according to their own understanding of logical terms, naive individuals—those who have not mastered formal logic or any of its cognate disciplines—should succumb to systematic fallacies. Second, we review studies corroborating this prediction. Third, to helps readers to digest these findings, we examine their implications for others accounts of reasoning and for human rationality.

## 1 The Theory of Mental Models

The Scottish psychologist Craik (1943) suggested that the mind constructs 'small-scale models' of reality that it uses to anticipate events. It appears to construct them as a result of perception (Marr 1982), imagination (Metzler and Shepard 1982), knowledge (Gentner and Stevens 1983), and the comprehension of discourse (Johnson-Laird 1983). Craik himself supposed that reasoning depends on verbal rules, but the modern theory of mental models began with the hypothesis that reasoning could be based on models too. The theory has developed in several expansions, in which the later versions make the same predictions as previous versions but add new predictions to them. The original theory used models to explain syllogistic and spatial reasoning (Johnson-Laird 1983), and it accommodated reasoning on the basis of sentential connectives, such as conjunctions, disjunctions, and conditionals (Johnson-Laird and Byrne 1991, 2002), and reasoning about probabilities based on the proportions of models in which events occurred (Johnson-Laird et al. 1999). From the start, the theory distinguished between intuitive reasoning, which has no access to working memory, and deliberative reasoning, which has access to working memory (e.g., Johnson-Laird 1983, Ch. 6). This distinction was due originally to the late Peter Wason, and it has become familiar in "dual process" theories of many different varieties (for a review, see, e.g., Evans 2008). If people had unlimited working memory capacities and could consider all possibilities consistent with a given set of assertions, then they would never make errors. Hence, the theory provides both an "algorithmic" account of human reasoning performance as well as a "computational" account of reasoning competency (see Marr 1982). Polk and Newell (1995) proposed a version of the model theory that was notable for making no use of counterexamples—a claim that seemed plausible at the time, but that was subsequently refuted (e.g., Johnson-Laird and Hasson 2003).

An expansion of the theory dealt with inductive and abductive reasoning (Johnson-Laird 2006; Johnson-Laird et al. 2004; Khemlani et al. 2013). And Koralus and Mascarenhas (2013) have developed their own version of the model theory, which has its background in formal semantics. In essence, their "erotetic" principle embodies two hypotheses. First, reasoning proceeds by treating successive premises as questions and maximally strong answers to them. A categorical premise such as "It is raining" asks no question, but a disjunctive premise, "It is raining or it is hot," poses the question: which is it? Second, as reasoners interpret each new premise, their asking a certain sort of question allows them to draw classically valid conclusions. The erotetic theory often runs in parallel with the original model theory, but there can be subtle differences between them.

The most recent version of the model theory is one that unifies reasoning about facts, possibilities, and probabilities (Khemlani 2016). This account diverges to a greater extent from the erotetic theory, because it does not allow all classically valid deductions. It is this unified theory that we outline in what follows.

The unified theory explains the computations underlying reasoning, both what they compute and how they compute it. A set of constraints characterizes the deductions that human reasoners draw for themselves: to deduce is to maintain semantic information, to simplify, and to reach a new conclusion (Johnson-Laird and Byrne 1991, p. 22). If they cannot draw such a conclusion, they declare that nothing follows from the premises, even though in logic infinitely many conclusions follow from any set of premises whatsoever, such as the conjunction of all the premises.

The theory explains how individuals make deductions. They envisage the possibilities to which the premises refer: they construct mental models of them, and they draw conclusions based on these models. A conclusion is *possible* if it holds in at least one model of the premises; it is *probable* if it holds in most models of the premises; and it is *necessary*—it is the case—if it holds in all the models of the premises. Likewise, an inference is invalid if it has a counterexample: a model of the premises in which the conclusion does not hold. A crucial feature of models, which distinguishes them from other sorts of proposed representation, such as semantic networks, is that they are iconic insofar as possible, that is, their structure corresponds to the structure of what they represent. For instance, consider a scenario in which a nail is placed to the right of a hammer on a table. A model of the scenario can be represented in the following iconic diagram:

hammer nail

The diagram is iconic because it can be scanned to yield conclusions, e.g., it follows (from scanning) that the hammer is to the left of the nail. No special inference rules are necessary beyond those that control how the representation is built and scanned. In contrast, a symbolic representation of the sort used in computer science and logic, such as:

right-of(nail, hammer)

cannot be scanned to yield inferences. Visual images are iconic, but models underlie images, which studies on mental rotation by Shepard and Metzler (1971) make clear: in their experiments, participants saw two drawings of "nonsense" figures made out of ten blocks glued together to form a rigid object with right-angled joints. Their task was to decide whether the pictures depicted one and the same object. The way they spontaneously carried out the task was to try to rotate the object in the first picture so that they could mentally superimpose it on the object in the second picture. Their decision times increased linearly with the angular difference between the orientations of the main axes of the object in the two pictures. This same result occurred whether the rotation was in the picture plane or in depth. To rotate the object in the picture plane is as though you are merely rotating the picture itself as it rests on top of the table. But, to rotate the object in depth is as though you are turning the actual three-dimensional object away from you or towards you. As Metzler and Shepard (1982, p. 45) wrote: "These results seem to be consistent with the notion that… subjects were performing their mental operations upon internal representations that were more analogous to three-dimensional objects portrayed in the two-dimensional pictures than to the two-dimensional pictures actually presented." In other words, the participants were rotating *iconic* mental models of the objects, which underlie their images of the objects. More recent research shows that images can slow reasoning down (Knauff et al. 2003). Hence, models form the basis of images, but they can be more useful for reasoning than mental images.

Models of a process can simulate the various steps of the process kinematically (Khemlani et al. 2013). Consider Fig. 1, which is a static diagram designed to depict two adjacent gears (see Schwartz and Black 1996). Given the figure, if the leftmost gear is turned clockwise, will the knob on the gear fit into the groove on the rightmost gear? Most reasoners respond that it will. To make the response, reasoners need to build a mental model of the scenario and rotate that model (akin to the Shepard and Metzler 1971, task). The mental rotation is not isomorphic to how gears rotate in real life, since it is unlikely that reasoners build a model of the individual interlocking teeth and their physical interactions. Rather, the simulation is piecemeal and kinematic, i.e., most reasoners have to consider discrete steps of the gear system individually (see also Hegarty 1992).

In addition to simulating physical scenarios, reasoners build models of compound assertions as sets of possibilities. For instance, an exclusive disjunction of the sort:

Either there is a circle or else there is a triangle, but not both has two iconic models to represent the two possibilities to which it refers:

●

▲

where each row in this diagram denotes a model of a separate possibility. However, certain aspects of models are symbolic. For example, an exclusive disjunction containing a negative clause:

Either there is a circle or else there is not a triangle has the mental models:

●

¬ ▲

where "¬" denotes a mental symbol representing negation. Reasoning on the basis of sentential connectives is computationally intractable (Cook 1971), i.e., as the number of distinct 'atomic' propositions in an inference increases, so does the amount of time and memory needed to draw deductions from those premises. Mental models are accordingly based on a *principle of truth* in order to reduce the load on working memory:

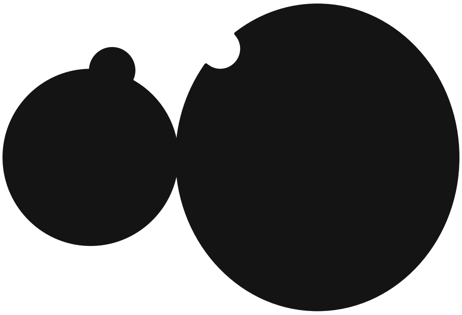

**Fig. 1** A model of two adjacent gears. Reasoners need to build a kinematic mental model that simulates the rotation of the gears to infer whether or not knob on the left gear will mesh with the groove on the right gear

*Mental models represent what is true, but not what is false.*

The models above of the exclusive disjunction with a negation illustrate the principle. They represent only what is possible according to the disjunction. The first model represents the possibility in which the circle occurs, but it does not represent explicitly that it is false that there is not a triangle in this case, i.e., there is a triangle. The second model represents the possibility in which there is not a triangle, but it does not represent explicitly that it is false that there is a circle in this case. Falsity should not be confused with negation: the former is a semantic notion, whereas the latter is a syntactic one. Mental models—indeed, a single mental model—underlie intuitive deductions. But, in certain circumstances—when a task is easy, for example—individuals can deliberate and flesh out mental models into *fully explicit* models, which represent propositions that are false. The fully explicit models of the exclusive disjunction with a negative clause are as follows:

● ▲

¬ ● ¬ ▲

These two possibilities can also be described in a biconditional assertion:

There is a circle if and only if there is a triangle.

Naive individuals tend not to notice the equivalence, which bears out their tendency to rely on mental models. Table 1 summarizes the mental models and the fully explicit models for the main sentential connectives. The difference between the two sorts of model will become clearer in the next section in which we outline their contrasting predictions.

Mental models explain why the opening inference about burning oil and blue smoke is difficult, and why individuals often tend to respond that nothing follows from the premises. A conditional:

**Table 1** The mental models and fully explicit models for assertions based on the principal sentential connectives

| The sentence | The mental models of its possibilities | | The fully explicit models of its possibilities | |
|---|---|---|---|---|
| A and B | A | B | A | B |
| Neither A nor B | ¬ A | ¬ B | ¬ A | ¬ B |
| A or else B, but not both | A | | A | ¬ B |
| | B | | ¬ A | B |
| A or B or both | A | | A | ¬ B |
| | | B | ¬ A | B |
| | A | B | A | B |
| If A then B | A | B | A | B |
| | … | | ¬ A | ¬ B |
| | | | ¬ A | B |
| If and only if A then B | A | B | A | B |
| | … | | ¬ A | ¬ B |

The symbol "¬" denotes negation, and the symbol "…" is a reminder that there are other implicit cases

If A then B

yields one model of the salient possibility (*A* and *B*):

$$\begin{array}{cc} \mathrm{A} & \mathrm{B} \end{array}$$

The further premise *not-B*, as in *there is no blue smoke*, eliminates the model, from which it seems that nothing follows. To draw the correct conclusion (a so-called "modus tollens" inference), reasoners need to consider the possibilities in which *A* is false. The fully explicit models are accordingly:

$$\begin{array}{cc} \mathrm{A} & \mathrm{B} \\ \neg\ \mathrm{A} & \mathrm{B} \\ \neg\ \mathrm{A} & \neg\,\mathrm{B} \end{array}$$

The premise, *not-B*, eliminates the first two models, and only the third model remains. It yields the conclusion, *not-A*: *there is no burning oil*. No other models of the premises exist, and so the conclusion is valid. In contrast, the mental models alone of the conditional, *If A then B*, above, can elicit a valid inference when it is known that *A* holds, i.e., a "modus ponens" inference. Table 2 describes the general principles that dictate how models are combined. The model theory accordingly predicts that reasoners should solve modus ponens inferences intuitively, i.e., with a single mental model, whereas they need to reason deliberately to solve modus tollens inferences, i.e., they need to build fully explicit model, and experiments corroborate the prediction (Johnson-Laird et al. 1992, p. 418).

Several veins run through the evidence for the model theory. Iconicity implies that some inferences call for the construction of more models than others. The resulting load on working memory predicts that inferences depending on multiple models should be more difficult than inferences depending on only one model.

**Table 2** The principles for combining mental models and pairs of fully explicit models

1. The conjunction of a pair of models containing respectively a proposition and its negation yields the null model (of an impossible instance), e.g.:

A B and ¬ A B yield nil

2. The conjunction of a pair of models that are not contradictory yields a model representing all the properties in the models, e.g.:

A B and B C yield A B C

3. The conjunction of a null model with any model yields the null model, e.g.:

A B and nil yield nil

4. If one mental model represents a proposition, *A*, which is not represented in the second mental model, and *A* occurs in at least one of the set of models from which the second model is drawn, then its absence in the second model is treated as its negation (and procedure 2 above applies); otherwise its absence is treated as its affirmation (and procedure 3 above applies). This procedure applies only to mental models

Many studies have shown that multiple-model inferences do take longer and elicit more errors than one-model inferences. Such results occur in reasoning with sentential connectives (Bauer and Johnson-Laird 1993; García-Madruga et al. 2001; Mackiewicz and Johnson-Laird 2012), with negations (Khemlani et al. 2012, 2014), and with quantifiers, such as, "All the artists" and "Some of the artists" (Bucciarelli and Johnson-Laird 1999; Khemlani and Johnson-Laird 2012). No study has reported results to the contrary.

Another vein in the evidence concerns counterexamples. When individuals have to evaluate given inferences, they can detect many sorts of invalid inference. They justify their evaluations typically by pointing out a counterexample, a tendency that is most frequent for putative conclusions that are consistent with the premises but that do not follow from them (Johnson-Laird and Hasson 2003). Counterexamples also tend to suppress inferences that individuals would otherwise make (Byrne et al. 1999; De Neys et al. 2003; Juhos et al. 2015).

Two notable features distinguish the recent unified theory from earlier accounts. First, the possibilities to which assertions refer have the force of conjunctions (Johnson-Laird et al. 2015a). Hence, individuals tend to draw the following sorts of conclusion from disjunctive premises (Hinterecker et al. 2016):

The fault is in the software driving the printer or in the connection to the printer, or both. Therefore:

1. It's possible that it's in the software driving the printer.
2. It's possible that it's in the connection to the printer.
3. It's possible that it's in both the software and the connection.

None of these inferences is valid in any modal propositional logic, i.e., a type of logic that extends reasoning about propositions to reasoning about possibility. They allow that the disjunction could be true even though it is impossible for the fault to be in the software (see Hinterecker et al. 2016). And it is difficult to formulate a modal logic capable of permitting the inferences. To make the inference in (1) above, you need an additional premise ensuring that A is not impossible. (If it were impossible, the premise could still be true but the conclusion false in even the weakest modal logical systems, and so the inference would be invalid.) The simplest additional premise that would render the inference in (1) valid would be:

$$\text{Not}(\text{Impossible}(\text{A}))$$

But, this premise is equivalent to the conclusion to be proved, and so there is no longer any need for the disjunctive premise. We see no obvious way around this problem. Hence, the unified theory postulates that all inferences are in default of information to the contrary. The initial default inference would only be blocked by knowledge that *A* is impossible.

Second, the meanings of disjunctions, unlike those in logic, are not truth functional: they refer, not to truth-values, but to possibilities. The unified theory allows that the meanings of clauses, their referents, and general knowledge, can all *modulate* the interpretation of sentential connectives. Knowledge can block the construction of models, and it can introduce relations between referents in models (see Johnson-Laird and Byrne 2002; Juhos et al. 2012; Quelhas et al. 2010). Modulation can also establish the truth-values of certain assertions a priori, e.g., "If God exists then atheism is false" is true a priori (pace Quine 1953).

## 2 Studies of Illusory Inferences

Our aim now is to examine one feature of the unified theory: its prediction that human reasoners commit systematic fallacies, which we refer to as "illusory" inferences, because they are often compelling. We begin with a simple example so that readers can understand what is at stake. Imagine that you are in a restaurant, and suppose that only one of the following two assertions is true:

> (1) You have the bread.
> (2) You have the soup or the salad, but not both.
> Also, suppose you have the bread. What, if anything, follows? Is it possible that you also have either the soup or the salad? Could you have both?

The rubric "only one of the following assertions is true" establishes an exclusive disjunction between assertions (1) and (2): one is true, and one is false. They therefore yield the following three mental models of the food that you can have:

| | | |
|---|---|---|
| Bread | | |
| | Soup | |
| | | Salad |

Given the further premise that you have the bread, the models predict that you should respond, "no", to the question of whether you could have both the soup and the salad. Reasoners make this response (Khemlani and Johnson-Laird 2009). But they are wrong. The principle of truth predicts their error. Their mental models fail to represent that when assertion (1) is true, assertion (2) is false, and its falsity implies that they either have both the soup and the salad or neither of them. Nevertheless, the inference is compelling, and it is perhaps the simplest illusory inference—a failure to think about the falsity of an exclusive disjunction.

The discovery of illusions came from a computer program implementing the theory of mental models for connectives such as "if" and "or" (Johnson-Laird and Savary 1999). Consider the following exclusive disjunction about a hand of cards:

> If there is a king in the hand then there is an ace in the hand, or else if there isn't a king in the hand then there is an ace in the hand.

The disjunction is exclusive because either there is or there isn't a king in the hand, and the two conditionals state the consequences of these two alternatives. The program produced the following mental models for the disjunction:

$$\begin{array}{rl} \text{King} & \text{Ace} \\ \neg\ \text{King} & \text{Ace} \end{array}$$

They look sensible, and they imply that in either case there is bound to be an ace. A further categorical premise can make the inference even easier:

There is a king in the hand.

It eliminates the second model to leave only the first model, from which it follows at once:

There is an ace in the hand.

However, there was a surprise in the program's fully explicit models for the disjunction. They were:

$$\begin{array}{rl} \text{King} & \neg\ \text{Ace} \\ \neg\ \text{King} & \neg\ \text{Ace} \end{array}$$

These models implied that there was *not* an ace. That seemed impossible, and so the author of the program spent half a day looking in vain for a bug in his program. He then hand simulated the process. The force of the exclusive disjunction is that one conditional is true and one conditional is false. The meaning of conditionals is highly controversial, but if a conditional, such as, "If there is a king then there is an ace," is false, then one possibility is that there is a king and not an ace (Oaksford and Stenning 1992):

$$\text{King} \quad \neg\ \text{Ace}\,.$$

Likewise, if a conditional such as, "If there isn't a king then there is an ace," is false then one possibility is that there is not a king and there is not an ace:

$$\neg\ \text{King} \quad \neg\ \text{Ace}\,.$$

But, one of the two conditionals must be false because they are in an exclusive disjunction, and so there are at least the two possibilities:

$$\begin{array}{rl} \text{King} & \neg\ \text{Ace} \\ \neg\ \text{King} & \neg\ \text{Ace} \end{array}$$

Even granted that there is a king, it fails to follow that there is an ace. The program was right, and its author's intuitions were wrong. The mental models of the premises yield the erroneous conclusion that there is an ace in the hand, but the fully explicit models of the premises established that it was possible—or even necessary, depending on your interpretation of conditionals—that there was not an ace in the hand.

In the light of this analysis, an initial experiment examined the inference:

If there is a king in the hand then there is an ace in the hand, or else if there isn't a king in the hand then there is an ace in the hand.
There is a king in the hand.
What follows?

Nearly all the participants, who drew their own spontaneous conclusions, inferred that there was an ace in the hand (Johnson-Laird and Savary 1999). It is a compelling inference, and they were highly confident that were correct. They were wrong, of course. In contrast, given control premises, such as:

> If there is a king in the hand then there is an ace in the hand, or else there isn't a king in the hand.
> There is a king in the hand.

They correctly inferred that there is an ace. In this case, the mental models of the disjunction are:

$$\begin{array}{ll} \text{King} & \text{Ace} \\ \neg\,\text{King} & \end{array}$$

And the categorical premise eliminates the second model. The fully explicit models of the disjunction are:

$$\text{King} \quad \text{Ace}$$

And the categorical premise leaves them unchanged:

$$\text{King} \quad \text{Ace}$$

So, as the fully explicit models show, the mental models yield the correct response.

A critic rejected this account, arguing that its authors erred, not their participants. The argument was that the participants think that in the illusory premises one conditional rule applies and that the other does not (Lance Rips personal communication, see Johnson-Laird and Savary 1999). Hence, they take the semantics of the disjunction to be akin to an instruction in a programming language:

> If A then do B,
> Else if not-A do B.

In other words, they treat "or else" as referring only to the antecedents of the two conditionals:

> If there is a king or else there is not a king, then there is an ace.

Of course the explanation is post hoc, and it is not obvious why individuals would interpret "or else" in this way, because the lines in the computer program express imperatives and therefore do not have truth values. The experimenters tested what happens with an unequivocal exclusive disjunction expressed as follows:

> One of the following assertions is true and one of them is false:
> If there is a king then there is an ace.
> If there is not a king then there is an ace.

The results confirmed the occurrence of illusory inferences and of correct control inferences. Nevertheless, some critics argued that individuals could still be making the programming interpretation of conditionals, even with the rubric above (Stenning and van Lambalgen 2008). Strictly speaking, the interpretation is wrong given that the rubric states that one conditional is true and one conditional is false. In any case, the critics overlooked quite different sorts of illusion (Johnson-Laird and Savary 1999, Experiment 3), such as:

Only one of the following two assertions is true:
    Albert is here or Betty is here, or both.
    Charlie is here or Betty is here, or both.
This assertion is definitely true:
    Albert isn't here and Charlie isn't here.
What follows?

The mental models of the opening pair of disjunctions (one true, and one false) represent five possible sets of individuals as possibly present:

| | | |
|---|---|---|
| Albert | | |
| | Betty | |
| Albert | Betty | |
| | | Charlie |
| | Betty | Charlie |

The categorical assertion rules out any model representing Albert or Charlie as present, and so it leaves just a single model:

Betty

The participants drew their own conclusions, and indeed most of them (85%) inferred:

Betty is here.

It is an illusory inference. The fully explicit models of the two disjunctions (one true and one false) are:

| | | |
|---|---|---|
| Albert | ¬ Betty | ¬ Charles |
| ¬ Albert | ¬ Betty | Charles |

And, as these models establish, Betty is not present. The experiment examined various other illusions and controls, and the results corroborated the principle of truth.

In these early studies, the illusory nature of inferences seemed likely to be transparent to the participants. It never was. Consider this example (from Goldvarg and Johnson-Laird 2000):

Only one of the following premises is true about a particular hand of cards:
    There is a king in the hand or there is an ace, or both.
    There is a queen in the hand or there is an ace, or both.
    There is a jack in the hand or there is a 10, or both.
Is it possible that there is an ace in the hand?

Every participant responded, "Yes". The mental models of the first premise show that an ace is possible, and so do the mental models of the second premise. Yet, the participants overlooked that if there was an ace in the hand, then both these first two disjunctions would be true—a state of affairs contrary to the rubric that only one of the three disjunctions is true. The problem illustrates an *illusion of possibility*: reasoners infer wrongly that a card is possible. We created a similar *illusion of impossibility* by replacing the two occurrences of "there is an ace" in the problem with, "there is not an ace". The participants succumbed to this illusion too, but they performed very well with comparable control problems. Figure 2 summarizes the results of this study in which the participants carried out 16 inferences, four illusions of possibility, four illusions of impossibility, and four of each of their respective control problems. Half of the illusions were based on disjunctions, and half were based on conditionals. The participants' confidence in their conclusions did not differ reliably between illusory and control problems. As the Figure shows, they were highly susceptible to the illusions but performed well with the control problems, and the illusions of possibility were more telling than those of impossibility. To infer that a situation is impossible calls for a check of every model, whereas to infer that a situation is possible calls only for a single model in which it holds, and so reasoners are less likely to make the inference of impossibility. This difference occurs in slightly harder problems that are not illusory (Bell and Johnson-Laird 1998). When two-premise problems had the heading "One of the premises is true and one is false," the participants still succumbed to the illusions. But, as predicted, the illusions were reduced when reasoners were told to check their conclusions against the constraint that only one of the premises was true (Goldvarg and Johnson-Laird 2000).

Another early study of illusions, and the first one to be published, concerned simple "extensional" probabilities (Johnson-Laird and Savary 1996). An event is probable if it holds in most models of the premises. And a principle of "indifference" posits that models are equiprobable unless there is evidence to the contrary. Hence, as other studies have shown, one event is judged to be more probable than another if it occurs in more models than the other (Johnson-Laird,

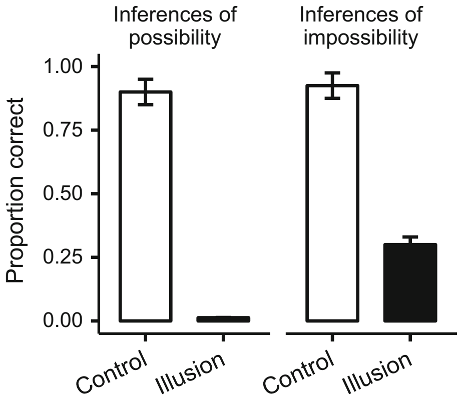

**Fig. 2** The proportions of correct responses to illusions of possibility, illusions of impossibility, and their respective control problems (based on Goldvarg and Johnson-Laird 2000, Experiment 1)

et al. 1999). Illusions yield a twist in these predictions, because individuals should rely on mental models. One study examined problems, such as:

> Suppose that only one of the following assertions is true about a specific hand of cards:
>
> There is a king in the hand or there is an ace in the hand, or both.
> There is a queen in the hand or there is an ace in the hand, or both.
>
> Which is more likely to be in the hand: the king or the ace?

Most participants relied on mental models, and inferred that the ace was more likely to be in the hand than the king. In fact, it is impossible for an ace to be in the hand. A further study extended the phenomena to a variety of different sorts of problem. It also established that illusions could be constructed with just two sentential connectives, e.g.:

> If one of the following assertions is true about a specific hand of cards, then so is the other assertion:
>
> There is a jack in the hand or else there isn't a queen in the hand.
> There is a jack in the hand.

All but one of the participants estimated a higher probability for a jack in the hand than for a queen. The rubric is equivalent, not to an exclusive disjunction, but to a biconditional relation between the two assertions. The mental models of the premises are:

$$\begin{array}{ll} \text{Jack} & \\ \text{Jack} & \neg\,\text{Queen} \\ \ldots & \end{array}$$

Individuals overlook that the biconditional rubric allows that *both* assertions can be false. And the fully explicit models show that the mental models are wrong:

$$\begin{array}{rl} \text{Jack} & \text{Queen} \\ \neg\,\text{Jack} & \text{Queen} \end{array}$$

The participants performed much more accurately with the matched control problem, which is surprisingly similar:

If one of the following assertions is true about a specific hand of cards, then so is the other assertion:

There is a jack in the hand or else there is a queen in the hand.
There isn't a queen in the hand.

They correctly judged that the jack was more likely to be in the hand than the queen.

There is an intimate connection between validity and consistency: a deduction is valid if and only if the negation of its conclusion is inconsistent with its premises (see, e.g., Boolos and Jeffrey 1989). Naive individuals do not understand the concept of "consistency", but they do understand an equivalent task: to assess whether or not the assertions in a set could all be true at the same time. A study showed that there are both illusions of consistency and illusions of inconsistency (Johnson-Laird et al. 2000). Here is a typical example:

> There is a nail on the table and/or there is a bolt on the table, or else there is a bolt on the table and there is a wrench on the table.

On four separate trials, the preceding disjunction was paired with one of four different conjunctions (see below), which according to the model theory should give rise respectively to an illusion of consistency, its matching control inference, an illusion of inconsistency, and its matching control inferences. The participants' task was to answer the following question for all four pairs of assertions:

> Is it possible that both assertions could be true at the same time?

The mental models of what's on the table given the initial assertion above are as follows:

| | | |
|---|---|---|
| nail | | |
| | bolt | |
| nail | bolt | |
| | bolt | wrench |

In contrast, the fully explicit models of the disjunction are:

| | | |
|---|---|---|
| nail | ¬ bolt | wrench |
| nail | ¬ bolt | ¬ wrench |
| ¬ nail | bolt | ¬ wrench |
| nail | bolt | ¬ wrench |

The discrepancy between the two sets of models establishes the following status for the four conjunctions paired with the disjunction on separate trials, which we present with the percentages of correct evaluations of consistency or inconsistency:

> Illusion of consistency: There is a bolt on the table and there is a wrench on the table: 2%
> Control for consistency: There is a nail on the table and there is a bolt on the table: 99%
> Illusion of inconsistency: There isn't a bolt on the table and there is a wrench on the table: 8%
> Control for inconsistency: There isn't a bolt on the table and there isn't a wrench on the table: 95%

These results were typical. All but one of the 128 participants made more errors with illusions than with the controls, and the one exception made no mistakes whatsoever. Because so many experts have themselves succumbed to illusory inferences, we have accumulated many putative explanations for them. They often argue that the premises are complex, ambiguous, artificial, and odd. And so people are confused, and as a result commit fallacies. This argument overlooks the experts' high confidence in their conclusions, and the equally complex, ambiguous, etc. control problems. Indeed, in the present study, the *same* initial premises occurred for illusions and controls. They differed only in the second sentences, which were all conjunctions (see above).

A special experimental task yielded more direct evidence for mental models. The task was to summarize the properties of those objects that the participants thought had consistent descriptions (Legrenzi et al. 2003). Hence, first they judged whether or not a description was consistent, and then for those descriptions that they judged to be consistent, they summarized the properties of the relevant objects. Here is a typical problem:

Only one of the following assertions is true:
The tray is heavy or elegant, or both.
The tray is elegant and portable.
The following assertion is definitely true:
The tray is elegant and portable.
Write a description of the tray: ________________

The mental models of the premises represent the tray as elegant and portable. Most participants described the tray in these terms. The fully explicit models, however, show that it is impossible for the tray to be both elegant and portable, and so there is an illusion of consistency. The matching control problem has the same initial pair of assertions but the following assertion was definitely true: The tray is heavy and elegant. This description fits both the mental models and the fully explicit models. As in previous studies, the participants succumbed to illusions of both consistency and of inconsistency, and performed much more accurately with control problems. Their descriptions of the properties of entities matched the mental models of the premises even when the participants succumbed to illusions of consistency.

Is there any way to ameliorate illusions and to get individuals to reason correctly? According to the model theory, the principle of truth yields the illusions. Therefore, any procedure that leads individuals to think about what is false should improve performance with illusory inferences. The rubric, “Only one of the following two premises is *false*,” did reliably reduce their occurrence (Tabossi et al. 1998), as did the participants’ prior production of false instances of individual premises (Newsome and Johnson-Laird 2006). Experimenters used a different procedure with quantified problems, such as:

Only one of the following statements is true:
At least some of the brown beads are round, or
All the brown beads are round.
Is it possible that all the brown beads are round?

Most participants succumbed to the illusion and responded, “yes” to such problems (Yang and Johnson-Laird 2000a). But, the participants in another study were told to think carefully about the consequences of the truth of the first assertion and the falsity of the second assertion, and then about the consequences of the truth of the second assertion and the falsity of the first assertion. The result was that the difference between controls and illusions vanished—the illusions became easier but the controls became harder (Yang and Johnson-Laird 2000b).

One other method to ameliorate illusory inferences is to use modulation. Many concepts depend on negation and on relations such as conjunction and disjunction, as in the concept of a "ball" in baseball: a pitch that does *not* enter the strike zone *and* is *not* struck at by the batter. The model theory applies to such Boolean concepts, and it postulates that individuals tend to represent only instances of a concept, and for each instance only those properties, affirmative or negative, that a description asserts as holding for the instance, i.e., the principle of truth as applied to concepts. As a consequence, there are predictable conceptual illusions in which individuals envisage as instances of a concept some cases that in fact are non-instances, and vice versa (Goodwin and Johnson-Laird 2010). Consider, for example, this description of a concept:

red and square, or else red.

Its mental models are:

red square
red

Individuals tended to write down both sorts of entity in listing the possible instances of the concept. But, as its fully explicit models establish, there is only one possible instance of the concept:

red ¬ square

A simple change to the content blocked the mental representation of the impossible member. The description:

red and green, or else green

inhibited the participants from thinking of entities of both colors, and so they were more likely to identify the one correct instance of the concept: red and not green.

Illusory inferences have often served as a litmus test for the use of mental models, and so studies have examined illusions in various domains of reasoning using various sorts of task. Table 3 summarizes the main studies, including those that we have described here. It presents a broad description of the task and the domain, and presents an example of a typical illusory problem, always one from several illusions and always tested with similar control problems. The table completes our survey of illusory inferences. None of the studies failed to produce a reliable effect, though not all of the illusions were equally compelling. In general, individuals make systematic and predictable fallacious inferences. We turn to the implications of these results for human rationality and for alternative theories of reasoning.

**Table 3** A summary of studies of illusory inferences and control problems

| The task | Domain | Example of the form of an illusory problem | References |
|---|---|---|---|
| What follows? | Sentential reasoning | Only one is true:<br>If A then B. If not A then B | Johnson-Laird and Savary (1999) |
| What follows? | Causal reasoning | One is true and one is false:<br>A will cause B<br>Not-A will cause B<br>Definitely true: A | Goldvarg and Johnson-Laird (2001) |
| Answer deductive yes/no question | Co-reference | Either j is W and X, or else k is Y and Z. j is W. Is j X? | Walsh and Johnson-Laird (2004) |
| | | Either j & k do W, or l & m do<br>If j does W then will k? | Koralus and Mascarenhas (2013) |
| Answer deductive question about what is present | Reasoning about checkers | Only one is true:<br>If j is there then k is there<br>If j is not there then k is there | Newsome and Johnson-Laird (2006) |
| Is the conclusion possible? | Sentential reasoning | Only one is true:<br>If A then not B. If B then A<br>Is A & not-B possible? | Sloutsky and Johnson-Laird (1999) |
| Is the conclusion possible? | Sentential reasoning | Only one is true:<br>A or B or both. C or B or both. D or E or both<br>Is B possible? | Goldvarg and Johnson-Laird (2000) |
| Is the conclusion possible? Check truth & falsity | Sentential reasoning | One is true and one if false:<br>A & B. B or else C<br>Is only A & B possible? | Khemlani and Johnson-Laird (2009) |
| Is the conclusion possible? | Quantified reasoning | Only one is true:<br>Some X are Y. All X are Y<br>Is all X are Y possible? | Yang and Johnson-Laird (2000a) |
| Is a conclusion possible? Check truth & falsity | Quantified reasoning | Only one is true:<br>Some X are not Y. No X are Y<br>Is no Yare X possible? | Yang and Johnson-Laird (2000b) |
| Is it possible for both statements to be true at the same time? | Quantified reasoning | All E are F iff all B are F<br>None of E is F<br>Possible for both to be true? | Kunze et al. (2010) |
| What are the possible instances? | Boolean concepts | A if and only if B, or else B | Goodwin and Johnson-Laird (2010) |
| Is the conclusion possible? Plus remedial contents | Set membership | Only one is true:<br>1. j is a W or X or both<br>2. j is a Y or X or both<br>3. j is a Z<br>j is not an W or Y. Is j a Z? | Santamaria and Johnson-Laird (2000) |

**Table 3** continued

| The task | Domain | Example of the form of an illusory problem | References |
|---|---|---|---|
| Is a situation possible? | Spatial relations | j is not in the same place as k or else k is not in the same place as l. Possible all three in different places? | Mackiewicz and Johnson-Laird (2012) |
| Could all three assertions be true? | Spatial relations | If j is to left of k then k is to left of l. j is to right of l<br>j is to right of k | Ragni et al. (2016) |
| Is an action permissible? | Deontic reasoning | Permitted only one action:<br>Take j or k, or both<br>Take l or k, or both<br>Permitted to take k? | Bucciarelli and Johnson-Laird (2005) |
| Could both assertions be true? | Sentential reasoning | A &/or B, or else B & C<br>B & C | Johnson-Laird et al. (2000) |
| | | A or else B<br>Not-A or else B | Johnson-Laird et al. (2012). |
| Could assertions all be true? If so, write description | Sentential reasoning | Only one is true:<br>A or B or both<br>B or C or both<br>Definitely true: B & C | Legrenzi et al. (2003) |
| Which is more probable, A or B? | Sentential reasoning | Only one is true:<br>If A then B. If C then B | Johnson-Laird and Savary (1996) |
| Which is more probable, A or B? | Sentential reasoning | Only one is false:<br>A or B or both<br>Not-A or B or both | Tabossi et al. (1998) |

In abbreviations, A–F denote propositions, W–Z denote predicates, and j–m denote individuals or objects

## 3 General Discussion

The unified theory of mental models predicts systematic fallacies. Of course, the notion of a fallacy implies the existence of some normative account of what is rational or free from error. In the past, theorists have often assumed that rationality in deductive reasoning is a matter of drawing conclusions in accordance with the pertinent branch of formal logic. As a reviewer reminded us, too much research has established that human reasoning does not adhere to the rules of classical logic, so that theorists have mostly given up the paradigm of 'rationality equals logic.' Our normative account of deductive rationality therefore depends not on any particular formal logic but rather on the underlying principle of validity to which most logics aspire, which as we implied earlier, can be defined as Jeffrey (1981, p. 1) does: "A valid inference is one whose conclusion is true in every case in which all its premises are true”. One caveat, however, is that there should be at least one case—or possibility in terms of our theory—in which the premises are true. The aim of this caveat is to preclude inferences that are vacuously valid because their premises are inconsistent, and therefore imply any case whatsoever. The criterion of validity yields an immediate account of what it means for a set of assertions to be consistent: no valid inference to the negation of one of the assertions can be made from the rest of them. In order to use validity or consistency as normative, we therefore need to know how people tend to interpret assertions. Let us illustrate how the idea applies to a simple illusion.

Consider the following problem based on two exclusive disjunctions:

> Either the pie is on the table or else the cake is on the table.
> Either the pie isn’t on the table or else the cake is on the table.
> Could both of these assertions be true at the same time?

The mental models of what’s on the table according to the first assertion are:

$$\begin{array}{rl} \text{pie} & \\ & \text{cake} \end{array}$$

and the mental models of what’s on the table according to the second assertion are:

$$\begin{array}{rl} \neg\ \text{pie} & \\ & \text{cake} \end{array}$$

The models have in common the possibility of the cake on the table, and indeed most participants responded: “Yes, the two assertions could both be true at the same time” (Johnson-Laird et al. 2012). But the response is an irrational illusion. The fully explicit models of the first assertion are:

$$\begin{array}{rr} \text{pie} & \neg\ \text{cake} \\ \neg\ \text{pie} & \text{cake} \end{array}$$

And the fully explicit models of the second assertion are:

$$\begin{array}{rr} \neg\ \text{pie} & \neg\ \text{cake} \\ \text{pie} & \text{cake} \end{array}$$

As they show, no possibility is common to both sets, and so the correct answer is that the two assertions could not both be true at the same time. The claim that the illusion is irrational rests on three assumptions. First, an exclusive disjunction, *Either A or else B*, refers to just two possibilities:

$$\begin{array}{rr} \text{A} & \neg\ \text{B} \\ \neg\ \text{A} & \text{B} \end{array}$$

and the other two cases are impossible ( A B, ¬ A¬ B ). Second, a negation such as, *There is not a pie on the table*, is equivalent to the falsity of the corresponding affirmative: *There is a pie on the table*. Third, two assertions could both be true if, and only if, they both refer to a possibility in common. In our view, none of these assumptions is controversial. It follows that reasoners err if they respond that the two assertions in the example could both be true. They are making an irrational evaluation, but the criteria of rationality presuppose no more than assumptions about the common interpretations of negation and exclusive disjunctions, and the concept of validity and its allied notion of consistency.

That people make irrational inferences in cases such as the example above does not imply that they are irremediably irrational. Indeed, with appropriate remedial procedures their performance improves, though seldom to one that is completely without error. Some theorists might argue that in principle human beings have a rational competence but err in performance. It is hard to see how to distinguish empirically between this claim and our point of view. But, we emphasize that the principle of truth, which has its own quasi-rational motivation—it cuts the load on working memory—is not easy to overturn. The fallacies it yields often have the quality of illusions: reasoners are highly confident in their responses (see Johnson-Laird and Savary 1999), and yet they are wrong. Moreover, unlike slips in linguistic performance, the illusions are highly predictable. They are easily elicited and robust in occurrence. Communication can even proceed successfully because speaker and listener alike make the same mistake. For example, a professor of chemistry warned his students:

> A grade of zero will be recorded if your absence [from class] is not excused, or else if your absence is excused other work you do in the course will count …

The mental models of this assertion yield the two possibilities that presumably he and his students had in mind:

| | | |
|---|---|---|
| ¬ excused | zero-grade | |
| excused | | other-work-counts |

But, the fully explicit models of the professor's assertion do not yield these models. What the professor should have asserted is, not a disjunction, but a conjunction of the two biconditionals:

> A grade of zero will be recorded if and only if your absence is not excused, and if and only if your absence is excused then other work you do in the course will count.

If, by magic, the processing limitations of working memory could be eliminated and reasoners always rely on fully explicit models, they would indeed meet our criterion of rationality apart from a few slips in performance here and there. But, in reality, we are all prey to irrational inferences when we are forced to rely on mental models. The errors arise because mental models represent what is true in each possibility but fail to represent what is false. For many inferences in daily life the failure is harmless, but in certain cases it leads to fallacious inferences. The theory predicts these fallacies and, as the previous section showed, their occurrence is robust. Readers may wonder, however, whether there might be another explanation for them.

We mentioned earlier that an alternative account exists for illusions based on the exclusive disjunction of conditionals. Naive individuals—for reasons unknown—interpret these conditionals as akin to those in a programming language:

If A then do B,
Else if not-A then do B.

This account fails with other illusions, and it fails to explain the exclusive disjunction concerning the pie and the cake on the table above.

Psychologists, notably Rips (1994), have proposed theories of reasoning based on formal rules of inference. But, such theories rely only on rules of inference that yield valid deductions, and so it is not easy to see how they might account for illusory inferences, such as the one above. Rules of inference of some sort are often presupposed in analyses of rationality (e.g., Stanovich 1999), and then much of the discussion is about the nature of the rules on which naive reasoners rely, such as normative rules corresponding to those of formal logic or prescriptive rules that take into account the limitations of human reasoning. But, no one has proposed an explanation of illusory inferences based on rules of inference. The model theory, however, makes no use of rules of inference. Other theorists have also attacked the notion that reasoning is about following normative rules. Stenning and van Lambalgen (2008) suggest that reasoning depends on a plurality of logics. Their account distinguishes between credulous reasoning aimed at making a single interpretation of a discourse in which all its utterances are true, and skeptical reasoning aimed at finding only conclusions that are true in all interpretations of the discourse. The account allows that it is possible to make errors in reasoning. Yet, with the exception of the argument about the disjunction of conditionals, which we outlined earlier, it offers no explanation of illusory inferences. We emphasize that an alternative theory of illusions needs to deal, not with just one or two of them, but with a comprehensive sample.

The situation is no better for theories of reasoning relying on probabilistic logic (e.g., Adams 1998). There has been a recent surge of interest in this idea of replacing logic with probability (see the interchange between Johnson-Laird et al. 2015a, Baratgin et al. 2015 and Johnson-Laird et al. 2015b). But, probabilities hardly explain why individuals accept, say, assertions of the sort, *pie or else cake*, and *not pie or else cake*, as consistent with one another. If one of these disjunctions is highly probable, then the other cannot be highly probable, i.e., they are therefore probabilistically inconsistent (Adams 1998, p. 181). Indeed, proponents of the new probabilistic paradigm have not ventured any explanations of either inconsistency or illusory inferences. It would be silly to claim that only the model theory *could* explain illusory inferences, but the illusions were first published 20 years ago, and no other comprehensive theory of them exists outside the model theory and its variant due to Koralus and Mascarenhas (2013).

The occurrence of systematic and predictable fallacies in human reasoning came as a shock to us. A computer program implementing the model theory predicts them (see http://mentalmodels.princeton.edu/models/ for computer programs that yield illusions), and, as we have argued, other current theories of reasoning have yet to explain them. To reason only about the truth is a sensible way to manage computational intractability. It is a reasonable compromise—a tax paid to intractability—and it is a rational enough method of reasoning to have ensured the continued survival of *homo sapiens*. But it does lead to illusions. And, to treat them as momentary lapses (pace Cohen 1981), or not as faults in reasoning itself (pace Henle 1978) stretches credibility. They are the regular consequences of the intuitive way in which humans reason. They change our picture of rationality. Its earlier defenders saw it as monolithic: reasoning was rational, because its underlying logic was impeccable despite occasional glitches in performance. This picture was simplistic. What we now know is that the machinery for reasoning is not unitary, and does not depend on an impeccable logic. Intuitions rely on mental models, which can give rise to predictable fallacies. Deliberations rely on fully explicit models, and so they can override illusions. Preventative methods, such as a conscious assessment of the consequences of falsity, usually come at a cost to reasoning with control problems, which do not require them. Yet, without these prophylactics, reasoners remain open to the illusion that they grasp what is in fact beyond them—that they are rational when in fact they can be systematically irrational.

## References

Adams, E. W. (1998). *A primer of probability logic*. Stanford, CA: Center for the Study of Language and Information.

Baratgin, J., Douven, I., Evans, J. S. B. T., Oaksford, M., Over, D., Politzer, G., et al. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences, 19,* 547–548.

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

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

Boolos, G., & Jeffrey, R. (1989). *Computability and logic* (3rd ed.). Cambridge: Cambridge University Press.

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

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

Byrne, R. M., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40,* 347–373.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317–331.

Cook, S. A. (1971). The complexity of theorem proving procedures. In *Proceedings of the third annual association of computing machinery symposium on the theory of computing*, pp. 151–158.

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

De Neys, W., Schaeken, W., & D'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31,* 581–595.

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

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.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldvarg, Y., & Johnson-Laird, P. N. (2000). Illusions in modal reasoning. *Memory & Cognition, 28,* 282–294.

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

Goodwin, G., & Johnson-Laird, P. N. (2010). Conceptual illusions. *Cognition, 114,* 253–265.

Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1084–1102.

Henle, M. (1978). Foreword to R. Revlin & R. E. Mayer (Eds.), *Human reasoning*. Washington, DC: Winston.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 1606–1620.

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

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press; Cambridge, MA: Harvard University Press.

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

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

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

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 109,* 646–678.

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

Johnson-Laird, P. N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition, 31,* 1105–1113.

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

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015b). Response to Baratgin et al.: Mental models integrate probability and deduction. *Trends in Cognitive Sciences, 19,* 548–549.

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science, 288,* 531–532.

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., Lotstein, M., & Byrne, R. M. J. (2012). The consistency of disjunctive assertions. *Memory & Cognition, 40,* 769–778.

Johnson-Laird, P. N., & Savary, F. (1996). Illusory inferences about probabilities. *Acta Psychologica, 93,* 69–90.

Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition, 71,* 191–229.

Juhos, C., Quelhas, A. C., & Byrne, R. M. J. (2015). Reasoning about intentions: Counterexamples to reasons for actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 55–76.

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

Khemlani, S. (2016). Automating human inference. In U. Furbach & C. Shon (Eds.), *Proceedings of the 2nd IJCAI Workshop on bridging the gap between human and automated reasoning* (pp. 1–4). CEUR Workshop Proceedings.

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. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin, 138,* 427–457.

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

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

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

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, 4,* 559–573.

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives, 27,* 312–365.

Kunze, N., Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2010). Illusions of consistency in quantified assertions. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science, 14,* 131–137.

Mackiewicz, R., & Johnson-Laird, P. N. (2012). Reasoning from connectives and relations between entities. *Memory & Cognition, 40,* 266–279.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Metzler, J., & Shepard, R. N. (1982). Transformational studies of the internal representations of three-dimensional objects. In R. N. Shepard & L. A. Cooper (Eds.), *Mental images and their transformations* (pp. 25–71). Cambridge, MA: MIT Press.

Newsome, M. R., & Johnson-Laird, P. N. (2006). How falsity dispels fallacies. *Thinking & Reasoning, 12,* 214–234.

Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 835–854.

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.

Quine, W. V. O. (1953). Two dogmas of empiricism. In *From a logical point of view* (pp. 20–46). Cambridge, MA. Harvard University Press.

Ragni, M., Sonntag, T., & Johnson-Laird, P. N. (2016). Spatial conditionals and illusory inferences. *Journal of Cognitive Psychology*, *28*(3), 348–365.

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

Santamaria, C., & Johnson-Laird, P. N. (2000). An antidote to illusory inferences. *Thinking & Reasoning, 6,* 313–333.

Schwartz, D., & Black, J. (1996). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology, 30,* 154–219.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171,* 701–703.

Sloutsky, V. M., & Johnson-Laird, P. N. (1999). Problem representations and illusions in reasoning. In *Proceedings of the twenty first annual conference of the cognitive science society*, pp. 701–705.

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.

Tabossi, P., Bell, V. A., & Johnson-Laird, P. N. (1998). Mental models in deductive, modal, and probabilistic reasoning. In C. Habel & G. Rickheit (Eds.), *Mental models in discourse processing and reasoning* (pp. 299–331). Berlin: Elsevier Science.

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

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

Yang, Y., & Johnson-Laird, P. N. (2000b). How to eliminate illusions in quantified reasoning. *Memory & Cognition, 28,* 1050–1059.