Mental models and the algorithms of deduction

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

Mental models and the algorithms of deduction

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

Abstract

This chapter begins with an outline of logic, and of the attempts to use it as a theory of human deduction. The fatal impediments to this approach led to the model theory in which models based on the meanings of premises yield deductive conclusions. And the chapter describes in detail the implementation of this theory's account of deductions based on sentential connectives such as "if", and how this simulation led to the discovery of systematic but compelling fallacies. The chapter outlines how algorithms based on models simulate deductions of the spatial relations among objects. And it points out the problems that need to be solved to extend accounts of elementary inferences from quantified assertions, such as "All artists are imaginative", to deal with multiply-quantified relations, such as "Everyone loves anyone who loves someone". One alternative to the model theory is the idea that human deduction relies on probabilities. This approach concerns only which inferences people make, not the underlying mental processes by which they're made. The model theory fills the gap, because it applies to the deductions of probabilities, both those based on frequencies or proportions, and those based on evidence pertinent to unique events, such as the probability that Trump is re-elected President of the USA. The chapter ends with an account of why theories of human deduction need to be simulated in computer programs.

Key words: Deduction – Mental Logic – Mental models – Probabilities - Quantifiers – Selection task - Spatial inferences

1. Introduction

Pose the following problem to a smart eight-year-old:

> All machines can break down.
>
> Alexa is a machine.
>
> What follows?

and the child is likely to reply:

> Alexa can break down.

So, as experiments confirm, human beings unschooled in logic are able to make deductions. Yet, this easy deduction defeats Alexa, Siri, and other virtual assistants. To build machines that reason, students of reasoning need to know the answers to three questions: 1. which deductions do human reasoners make? 2. how do they make them? And 3. how can computers simulate them? The goal of this chapter is to describe the main efforts to simulate human deduction. It aims to provide its own intellectual life-support system so readers can understand it without having to consult anything else. It proceeds from the main approach to human reasoning that has led to computational simulations – the theory of mental models, a remote descendant from logic that is no longer compatible with its classical branch, the predicate calculus. Here and throughout this chapter, the term "orthodox logic" refers to this calculus, whose basic principles are presented below. The "model theory" refers to the most recent version of the theory of mental models (e.g., Khemlani, Byrne, & Johnson-Laird, 2018). And the term "assertion" does double duty: it refers both to a declarative sentence and to the proposition – which can be true or false – that the sentence expresses depending on its context.

Theories of thinking have a crucial though often neglected goal: they need to explain their own creation. So, theories of reasoning must explain themselves. They cannot depend solely on the sort of machine learning embodied in current programs in artificial intelligence (AI). Because language leads to reasoning, and because people can verbalize their thoughts, theories of their reasoning must explain how people understand discourse. Their simulations call for explicit grammar, lexicon, and parser; a module that simulates the mental representations humans compute when they comprehend language and thought; and a reasoning engine to make deductions and other inferences. Three main sorts of theory of the deductive component of the engine exist: those that depend on mental models of the world (e.g., Khemlani et al., 2018), those that depend on a "mental logic" of rules from a logical calculus (e.g., Rips, 1994), and those that depend on the probability calculus (e.g., Oaksford & Chater, 2020). The latter theories aim to account only for which inferences individuals make, not how they make them.

The chapter accordingly deals with these topics:

- The basic concepts of logic and deduction.
- Mental logic and its critical differences from human deductions.
- The first algorithmic account of human reasoning.
- The algorithms that underlie model-based reasoning.
- Simulations of spatial reasoning.
- Simulations of reasoning about properties.
- Simulations of probabilistic reasoning.

Why should cognitive scientists simulate human reasoning? The chapter concludes with an answer to this question.

2. Basic concepts in logic and deduction

Deduction has two goals: to yield valid inferences and to assess consistency. An inference from premises to a conclusion is *valid* provided that the conclusion is true in every case in which the premises are true (Jeffrey, 1981, p. 1). A set of assertions is *consistent* provided they can all be true at the same time. Validity and consistency are independent of any logic, and interdependent on one another. An inference is valid if the negation of its conclusion is not consistent with its premises; and a set of assertions is consistent if there is no valid deduction of the negation of one of the assertions from the others. Logics depends on the concept of validity: the rules and axioms of a logic determine which inferences are valid. Orthodox logic, for instance, allows for valid inferences from inconsistent premises; indeed, any conclusion whatsoever follows from them. In daily life, reasoners do not draw deductions from inconsistencies. Hence, a rider is necessary for everyday validity: people draw deductions from consistent information. Naive individuals – the term refers to those with no training in logic or cognate disciplines – can make deductions that are valid in orthodox logic. No procedure can decide whether or not an inference is valid in this logic, that is, if the inference is valid, then it can be proved, but if it is invalid, no algorithm can be guaranteed to prove its invalidity. Orthodox logic contains the sentential calculus, i.e., a more rudimentary system that deals only with connections between sentences or clauses. The sentential calculus handles deductions that depend on negation, and simplified versions of such sentential connectives as *if*, *or*, and *and*. It is computationally intractable (and so the more complex predicate calculus is too) in that as the number of different assertions in inferences increases, the

amount of time and memory needed to establish validity increase even faster – to the point that deductions soon exceed the capacity of any finite device, such as the human brain (Cook, 1971).

A logic has three parts. Its first part is a grammar that specifies all and only those assertions to which the logic applies. Its second part is its proof theory, which consists of formal rules of inference, perhaps supplemented with axioms, that allow proofs that derive conclusions from premises. A typical formal rule of inference is:

If A then B.

A.

Therefore, B.

where the capital letters *A* and *B* denote assertions, which can be compounds containing further connectives, or else atoms that do not. A typical axiom (or postulate) is:

For any *x*, and any *y*, if *x* is on the left of *y* then *y* is on the right of *x*.

where the variables refer in a consistent way to entities in a spatial domain. An example of a formal proof is as follows, where the first two assertions are premises:

1. If Pat is on the right of Viv then they are opposite Ross.
2. Viv is on the left of Pat.
3. Therefore, Pat is on the right of Viv. (from line 2 and the axiom above)
4. Therefore, they are opposite Ross. (from lines 1 and 3, and the formal rule above).

The third part of a logic is its semantics, which defines the meanings of logical terms and allows assessments of the validity of inferences. Orthodox logic defines the meanings of connectives, such as its analogs of *if* and *or*, as true or false depending on the truth values of the clauses that they connect. The *material* conditional of logic, *if A then*

B, concerns four cases, depending on whether each of *A* and *B* is true or false. And orthodox logic defines the material conditional as false only in case *A* is true and *B* is false. In any other case, it is true. (The four cases can be spelt out explicitly in a "truth table".) So, unlike everyday conditionals, *If A then B* in logic is true whenever *A* is false. And it is true in case B is true.

To apply orthodox logic to a set of sentences, the first task is to recover their *logical forms* in order to match them to formal rules of inference, such as the rule above. This task is trivial when sentences are unambiguous, as in the case of a grammar that yields only their logical forms. But, for natural language, the task is extraordinarily difficult – to the point that no algorithm exists to carry it out. Natural language can yield ambiguous sentences, and content and context have a massive effect on the assertion that a sentence makes. Logical forms in natural language depend on meanings, e.g., the phrase, "Take the cookie and you'll get smacked," conveys a conditional assertion, *If A then B*, not a conjunction, *A and B*. But, when a reasoner has represented the meanings of assertions, those representations can be the basis of reasoning, and logical forms become superfluous.

A natural language has a mental lexicon of the meanings of words, and a grammar with rules that also account for how the meaning of an assertion is composed from the meanings of its grammatical parts, which in turn are composed from the meanings of their parts, and so on . . . down to the meanings of words or morphemes. A parser uses semantic principles attached to the syntactical rules to carry out this process of composition. Its results can be ambiguous. The simulations of deduction that we describe

below contain elementary versions of each of these components: a lexicon, a grammar, and a parser.

3. Mental logic and deduction

Early psychologists of reasoning took for granted that reasoners rely on orthodox logic (e.g., Beth & Piaget, 1966), and they sought to understand how the mind formulates that logic. Naive reasoners have no awareness of axioms. So, theorists converged on the hypothesis of unconscious rules of inference akin to those in the proof theory for the sentential calculus (e.g., Braine, 1978; Johnson-Laird, 1975; Osherson, 1974-6). Rips (1994) described a mental logic close to orthodox logic, and he implemented the theory in a computer program called PSYCOP (for the psychology of proof). Its inputs were logical forms – so it evaded the problem of recovering them from natural language – and it relied on two sorts of rules of inference. One sort, such as the rule above: *If A then B*; *A*; therefore, *B*, allows a person to reason forwards from premises to reach a conclusion. In contrast, a formal rule, such as:

A.

Therefore, A or B, or both.

where *B* can be any assertion whatsoever, can be applied to its own conclusion. In which case, it yields, for instance:

Therefore, (A or B, or both) or C, or both.

It can apply to this conclusion too, and so on in an infinite chain of deductions. PSYCOP curbs the rule. It is relegated to the second set of rules that can be used only to reason backwards from a given conclusion towards the premises. Even though the theory did not

allow individuals to infer their own conclusions (cf. our opening example of an inference), it was the high point of accounts of human deduction based on mental logic.

One premonition of problems to come concerned the following rule, which holds in logic for the material conditional:

It is not the case that if A then B.

Therefore, A and not B.

PSYCOP excluded this rule, because it included only those that "the individual recognizes as intuitively sound" (Ibid. p.104). In fact, most people do not accept this rule, and take the denial of the conditional to be: *If A then not B*.

What has become clear since PSYCOP is that the idea that everyday deductions depend on orthodox logic has several fatal impediments. The first is that the logic allows infinitely many valid conclusions to follow from any set of premises (e.g., the chain of inferences introducing *or* above).

The second impediment is that given any premises, even self-contradictory ones, orthodox logic never implies that a valid conclusion should be retracted. Consider, for instance, the following premises:

The Prime Minister lied to the Queen.

If the Prime Minister lied to the Queen then he resigned.

Both logic and common sense suggest the conclusion:

The Prime Minister resigned.

But suppose that did not happen. Orthodox logic and common sense now part company. Logic says nothing. The fact contradicts the conclusion, but in logic a self-contradiction implies any conclusions whatsoever. Hence, orthodox logic is *monotonic*, because with

more premises, more conclusions follow. It never requires a conclusion to be retracted, not even one that facts contradict. Common sense says, on the contrary: give up the conclusion, think again about the premises, and try to find an explanation that reconciles the inconsistency. Everyday reasoning is therefore nonmonotonic (or "defeasible"): more premises can lead to the retraction of earlier conclusions and to the revision of premises. Some theorists propose that nonmonotonic logics – systems designed to handle the withdrawal of conclusions – underlie human reasoning (Stenning & Van Lambalgen, 2012), and defeasibility is built into the model theory (Johnson-Laird, Girotto, & Legrenzi, 2004).

The third problem concerns the consistency of a set of assertions, that is, whether they can all be true at the same time. People tend to reject inconsistent assertions if they notice the inconsistency: at least one of them must be false. Logic has rules for proving conclusions, but it is not obvious at once how to use them to assess the consistency of a set of assertions. In fact, a general method is: if the negation of one assertion in the set follows from the other assertions, then the set is inconsistent. Otherwise, after an exhaustive but fruitless search for a proof, the set is consistent. The procedure seems implausible in everyday life. And experiments show that contrary to its prediction, consistency is not harder to deduce than inconsistency – it can even be easier (e.g., Johnson-Laird et al., 2000). How people decide whether or not assertions are consistent has a simple procedure: just determine whether or not the assertions have a model. Meanwhile, the implausibility of orthodox logic for reasoning in daily life may explain why it has not led to a simulation of deductions from everyday assertions as opposed to their logical forms.

4. The first algorithmic theory of human reasoning

The first algorithm designed to simulate an element of human reasoning was a step towards a plausible general theory. The algorithm was formulated to explain a striking phenomenon of how people test hypotheses. Wason (1968) devised a task that examines the potential evidence that naive individuals select to test the truth or falsity of a general hypothesis, such as:

If people have cholera then they are infected with a bacterium

or its equivalent:

All people who have cholera are infected with a bacterium.

There are two sensible ways to test the hypothesis. One way is to examine a sample of people who have cholera and check whether they are all infected with a bacterium. Another way, albeit less practical, is to test a sample of people who are not infected with a bacterium and check whether any of them have cholera. Each method rests on the principle that a person with cholera who is not infected with a bacterium is a counterexample that establishes the falsity of the hypothesis. Popper (1959) argued that potential falsifiability distinguishes a science, such as astronomy, from a non-science, such as astrology. Wason therefore designed his "selection task" to test whether naive individuals grasp the importance of counterexamples.

In the original version of the task (Wason, 1968), the experimenter lays four cards out in front of a participant:

E K 2 3

The participant knows that each card has a letter on one side and a number on the other

side. The task is to select all and only those cards to turn over to determine the truth or falsity of the general hypothesis:

> If there is a vowel on one side of a card then there is an even number on the other side.

Most people select the E card alone, many select both the E and 2 cards, and a few select the three cards E, 2 and 3. What's striking is how few people select the two cards: E and 3. Yet, they are the only two cards needed to evaluate the hypothesis. The K card is irrelevant, as people realize, because whatever is on its other side cannot refute the hypothesis. But, so too is the 2 card, for the same reason. Yet, the 3 card is crucial: if there is an A on its other side, it is a counterexample to the hypothesis, and thereby falsifies it.

The failure to select a potential counterexample shocked psychologists and philosophers (see Ragni, Cola, and Johnson-Laird, 2018, for the history). Defenders of human rationality argued that the task was a trick, that it was overcomplicated, and that it was impossible for human reasoners to be irrational. Yet, this claim is like arguing that it is impossible to break the rules of bridge, because, if you do, you are no longer playing bridge (Ramsey, 1990, p. 7).

Johnson-Laird and Wason (1970a) published a theory and an algorithm for how people carry out the selection task. The algorithm was in a flowchart, not a program, because computers were not accessible to psychologists in those days. It assumed that individuals used the meaning of the hypothesis to guide their selection of evidence. It implemented Wason's idea of two processes in reasoning: a reliance on intuition, now known as "system 1", and, somewhat rarer, a switch to deliberation, now known as

"system 2". So, the theory was an instance of what nowadays is called a "dual process" account (see also Sun, 2016, for an architectural account of dual process theories). The alternative theories of the selection task – and there are at least 16 of them – focus on what is computed rather than how.

The algorithm works as follows (see Ragni et al., 2018): it first makes a list of those items of potential evidence to which the hypothesis refers. If the general conditional, *if p then q*, is taken to imply its converse, *if q then p*, then both *p* and *q* are listed as potential evidence. Otherwise, only *p* is on the list. With no insight into the role of counterexamples, the algorithm selects the items on the list. But, with partial insight, it adds any further item that could verify the hypothesis. So, if *q* is not on the list, it is selected now, because it could verify the hypothesis. But, if there are no such further items, the algorithm adds any item that could falsify the hypothesis. So, if *q* is already on the list, the simulation adds *not-q* because it can falsify the hypothesis, yielding the selection of three items: *p*, *q*, and *not-q*. With insight into falsification from the outset, the algorithm selects only items that are potential counterexamples to the hypothesis, i.e., *p* and *not-q*.

A recent computer simulation used probabilistic parameters governing the interpretation of the conditional and whether insight occurs. A meta-analysis of 228 experiments corroborated the algorithm's principal predictions: the selection of an item is dependent on other selections rather than independent of them, the selections tend to be the four predicted sets of items listed above, and manipulations such as the use of hypotheses about everyday matters enhance the selection of potential counterexamples. Only one other theory was consistent with these predictions, and it was ruled out by its

inability to predict the selection of the three cards, *p*, *q*, and not-q, other than by guesswork. Yet, this selection was the most frequent in one study (Wason, 1969). The simulation fit the data from the experiments well. Its code and that of all the model-based programs referred to this article are available at https://www.modeltheory.org/models/.

Science and the selection task rely on general hypotheses. Their interpretation in logic as material conditionals has several implausible consequences. One of them is that a conditional, such as:

If anything is a quark then it forms composite particles

is equivalent to its contrapositive:

If anything does *not* form composite particles then it is *not* a quark.

The equivalence yields a well-known "paradox" of confirmation (Hempel, 1945). For example, a duck-billed platypus corroborates the hypothesis about quarks, because a platypus does not form composite particles and is not a quark. But, matters are still worse, because if quarks do not exist, then the general hypothesis about them is bound to be true. Its truth is vacuous, because it can be false only in case a quark exists – and does not form composite particles. The mental model theory of reasoning was formulated to solve such puzzles as the paradox of confirmation.

5. The algorithms that underlie model-based reasoning

The model theory asserts that people do not use logical rules to reason, but instead envisage the possibilities compatible with the meanings of premises. They build mental models that represent these possibilities. The crucial distinguishing characteristic of a mental model is that it is *iconic*, that is, it has the same structure as what it represents. The

human reasoning engine operates on the principle that a conclusion follows from the premises provided that they have no model that is a counterexample. What complicates reasoning are the meanings of assertions. Consider the following weather report:

It's rainy or cold, or both.

From this disjunction, people make the following deductions (Hinterecker, Knauff, & Johnson-Laird, 2016):

It is possible that it's rainy.

It is possible that it's cold.

It is possible that it's rainy and cold.

The disjunction refers to a conjunction of these three exhaustive possibilities, and rules out as impossible the case in which it is not rainy and not cold. Each possibility holds in default of knowledge to the contrary. So, if a discovery reveals that it isn't rainy, then this fact eliminates two of the possibilities above, and it follows that it's cold, because that's the only possibility. But, if in fact it isn't cold either, then the disjunction is false: the facts have ruled out all the possibilities to which it refers. In short, the model theory's semantics for sentential connectives is that they refer to exhaustive conjunctions of possibilities that each hold by default. However, because a conjunction, *and*, refers to just one possibility, it asserts a fact.

The semantics ensures that the model theory is nonmonotonic. And it has a striking consequence: none of the inferences above is valid in orthodox logic. The relevant logic has to deal with possibilities – it is a *modal* logic, of which there are infinitely many distinct sorts (e.g., Hughes & Cresswell, 1996). A persistent misconception of the model theory is that it has the same semantics as logic (e.g., Oaksford & Chater, 2020, p. 12.3).

To understand how they differ, consider the first conclusion above: *it is possible that the weather is rainy*. For most people, the inference is obviously valid. But, here is a counterexample: suppose that it is impossible that it is rainy, but it is cold. The disjunctive premise that *it's rainy or cold or both* is true, but the conclusion that *it is possible that it's rainy* is false – in fact, it is impossible. And so the inference is invalid in all normal modal logics. In the model theory, the inferences are valid by default, i.e., new information can overturn them. Sentential connectives therefore have a default semantics: reasoning in daily life is nonmonotonic. Table 1 below illustrates algorithms for model-based reasoning: it shows how computational implementations make use of this semantics to build and reason with models.

The model theory postulates a default semantics for conditionals too. An assertion such as:

If it's rainy then it's cold

asserts that *it is possible that it's rainy*, which in turn presupposes that *it is possible that it isn't rainy* (Johnson-Laird & Ragni, 2019). So, the conditional can be paraphrased as:

It is possible that it's rainy and that it's cold, and it is possible that it's not rainy.

This paraphrase unpacks into an exhaustive conjunction of three default possibilities:

It is possible that it's rainy and that it's cold.

It is possible that it's not rainy and that it's not cold.

It is possible that it's not rainy and that it's cold.

Individuals make these inferences, which are listed in the order in which children make them as the capacity of their working memories increases (see, e.g., Barrouillet & Lecas, 1999). Conditionals presuppose the possibility that their *if*-clauses do not hold, and the

key point about presuppositions is that they are true for both the affirmation of an assertion and its negation, e.g., *it has stopped raining* presupposes that it was raining, and so too does *it has not stopped raining*. The negation of the conditional above is therefore:

If it's rainy then it is not cold.

In a program simulating sentential reasoning, the intuitive system 1 represents possibilities using *mental* models in which each model of a possibility represents only those clauses in the conditional that hold in that possibility. The mental models of a conditional, *If A then B*, are:

```
A        B
  .  .  .
```

The first model represents the default possibility of *A and B*, and the second model allows for other possibilities such those in which *not-A* holds. (If *A* or *B* is itself a compound assertion then its semantics is taken into account in building the models.) In contrast, the deliberative system 2 represents the conditional by fleshing out mental models into *fully explicit* models representing all the assertion's clauses in each model, using negation (symbolized as "¬") to represent their falsity in the possibility. So, the fully explicit models of the conditional are as follows, where the possibilities of *not-A* are presuppositions, and each default possibility in the conjunction is shown on a separate line:

```
   A        B
¬  A     ¬  B
¬  A        B
```

The program takes the meanings of negation (*not*) and of conjunction (*and*) to be

fundamental, and it uses these meanings to define all the other connectives. For instance, an exclusive disjunction, *Either A or else B but not both*, is defined for system 2 as the following conjunction of two default possibilities:

```
  A     ¬ B
¬ A       B
```

Since sentential connections can be embedded, as in, *A and (C or D or both),* the system operates recursively. For instance, *B* above might denote the models for the assertion, *C or D or both*.

The meaning of negation refers to the complement of the set of models for the assertion that is negated. For example, the complement of the following set of models (for the biconditional assertion *if and only if A then B*):

```
  A      B
¬ A    ¬ B
```

is:

```
  A    ¬ B
¬ A      B
```

So, a set and its complement exhaust all the possible combinations of the items and their negations. But, negation ignores presuppositions, because they hold for the negated assertions too. Hence, the negation of a conditional, *If A then B*, yield the models:

```
  A    ¬ B
¬ A    ¬ B
¬ A      B
```

And they are the models of the conditional: *If A then not B*.

Conjunction is needed for compound premises, because it is part of the meaning of each connective. It is also needed to conjoin the models for one premise with those for another premise (see Table 1 for an example of how spatial models can be combined).

Table 1. Seven basic functions that underlie model-based reasoning illustrated for spatial reasoning: the name of the function, its input, its output, and pseudo-code for its algorithm. The appropriate function is called as a result of a procedure that checks which referents in a premise already occur in at least one model. Spatial models have three deictic axes: left-right, above-below, and front-behind. Algorithms refer to additional functions not included in the table, e.g., `RETRIEVE`, `ADD`, and `COMBINE`, whose operations are self-explanatory.

Function	Input	Output	Algorithm
1. START a mental model	Premise: *d is to the right of e*	Spatial model: e d	1. **RETRIEVE** subject (*d*) and object (*e*) of premise. 2. **RETRIEVE** semantics of spatial relation. 3. **ADD** tokens to a model to satisfy semantics. 4. **RETURN** model.
2. UPDATE a mental model by adding a referent	Model & premise: e d *d is to the left of f*	Spatial model: e d f	1. **IF** subject (*d*) not in model: 2. **ADD** subject to model according to semantics. 3. **ELSE IF** object (*f*) not in model 4. **ADD** object to model according to semantics. 5. **RETURN** model.
3. UPDATE a mental model by adding a relation	Model & premise: e d *e is larger than d*	Spatial model e d	1. **MODIFY** subject and object to satisfy semantics of relation. 2. **VALIDATE**(model, premise)
4. VALIDATE that an assertion holds in a model	Model & assertion: e d *d is to the right of e*	Truth value: True	1. **IF** subject (*d*) and object (*e*) satisfy relation in model: 2. **IF** system 1 enabled: 3. **RETURN** True. 4. **ELSE IF** system 2 enabled: 5. **SEARCH**(model, assertion) for counterexample.

			6. **ELSE** 7. **IF** system 1 enabled: 8. **RETURN** False. 9. **ELSE IF** system 2 enabled: 10. **SEARCH**(model, assertion) for example.
5. CONJOIN two models according to a relation between referents in each of them	2 models & premise 1: e d 2: f g *f is above d*	Spatial model: f g e d	1. **IF** subject (*f*) occurs in model 1 and object (*d*) occurs in model 2 **OR** subject occurs in model 2 and object occurs in model 1: 2. **COMBINE** models 1 and 2 according to relation (or its converse) to make a new model; ADD new axis to model if necessary. 3. **RETURN** new model.
6. SEARCH for a counterexample to a conclusion	Model & conclusion: d e f *f is to the right of e*	Spatial model & evaluation d f e *Conclusion is possible*	1. **FOR** each *R* in a set of revisions to model, where *R* satisfies premises: 2. **IF** *R* satisfies conclusion: 3. **RETURN** *R* and *conclusion is possible* 4. **ELSE** 5. **RETURN** model and *conclusion is necessary*
7. SEARCH for an example of a conclusion	Model & conclusion d e f *f is to the left of e*	Spatial model & evaluation d e f *Conclusion is impossible*	1. **FOR** each *R* in a set of revisions to model, where R satisfies premises: 2. **IF** *R* satisfies assertion: 3. **RETURN** *R* model and *conclusion is possible* 4. **ELSE** 5. **RETURN** model and *conclusion is impossible*

We illustrate how conjunction operates for models of compound assertions. It begins with two sets of models, such as:

```
  A      B
¬ A    ¬ B
```

and:

```
  B    ¬ C
¬ B      C
```

It then forms their pairwise conjunctions – but if a model from one set contains an element, such as *B*, and a model from the other set contains its negation, ¬*B*, it would be a self-contradiction, and so it does not return a model and moves on to the next pairwise conjunction. The conjunction of the two sets of models above proceeds as follows:

```
  A      B    and     B    ¬ C    yields            A      B    ¬ C.
  A      B    and   ¬ B      C    do not conjoin because B contradicts ¬B.
¬ A    ¬ B    and     B    ¬ C    do not conjoin because ¬B contradicts B.
¬ A    ¬ B    and   ¬ B      C    yields          ¬ A    ¬ B      C.
```

The result is therefore the conjunction of these two models of default possibilities:

```
  A      B    ¬ C
¬ A    ¬ B      C
```

The semantics of negation and conjunction suffice to capture the meaning of the basic sentential connectives. Table 2 describes the semantics for the mental models of system 1 and for the fully explicit models of system 2.

Recent computational models contain several refinements that are needed to simulate human reasoning (Khemlani et al., 2018; Khemlani & Johnson-Laird, 2020). They include:

- a component that uses a knowledge-base to modulate the interpretation of compound assertions by blocking possibilities,

- a defeasible (i.e., nonmonotonic) component that retracts a conclusion in the face of a contradictory fact, withdraws a premise to restore consistency, and seeks a causal explanation in the knowledge-base to resolve the original inconsistency,
- a component that simulates the verification of assertions and that can construct counterfactual assertions, which describe events that were once possible but that did not occur (see, e.g., Byrne, 2005).

Table 2: The semantics of compound assertions depending on sentential connectives (in systems 1 and 2), where *A* and *B* stand for atomic or compound assertions. Each assertion yields a conjunction ("and") of models of default possibilities, which are each shown in a separate row. Each row shows a model, which is, in turn, a conjunction of models of clauses or their negations ("¬"), or a mental model with no explicit content ("...").

Assertion	**Semantics for mental models in system 1**	**Semantics for fully explicit models in system 2**
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
A or B or both.	A B A B	A ¬B ¬A B A B
A or else B but not both.	A B	A ¬B ¬A B

All of the computational models implement the model theory's general principles about deductive conclusions, which follow in default of knowledge to the contrary:

- If a conclusion holds in all the models of the premises then it is *necessary* given the premises.
- If it holds in most of the models of the premises then it is *probable*.
- If it holds in some model of the premises then it is *possible*.
- If it holds in none of the models of the premises then it is *impossible*.

Likewise, a set of assertions is consistent if they have a model, and inconsistent if a model cannot be built from the premises (i.e., a situation in which the program constructs an empty model). The principal components for simulating deduction are illustrated for spatial reasoning in Table 1 above.

A major and unexpected consequence of the original simulations of the model theory is that intuitive reasoning based on models led to the discovery of many compelling illusions, which only deliberation with fully explicit models can correct (Khemlani & Johnson-Laird, 2017). Here is an example based on two exclusive disjunctions:

Either there's fog or else there's snow.

Either there isn't fog or else there's snow.

Can both of these assertions be true at the same time?

The mental models of the two disjunctions are respectively:

```
fog
        snow
```

and:

```
¬ fog
```

```
snow
```

A model of snow is common to both disjunctions, and so individuals should respond, "yes, the two assertions can both be true". However, the fully explicit models of the two disjunctions are:

```
  fog    ¬ snow
¬ fog      snow
```

and:

```
¬ fog    ¬ snow
  fog      snow
```

No possibility is common to these two sets of models: for one disjunction it snows without fog, and for the other disjunction it snows with fog. Their conjunction yields an empty model. Most people judge that the two disjunctions can both be true, but these fully explicit models show that doing so is wrong.

The model theory elucidates the earlier description of the "paradox" of confirmation. A conditional hypothesis, *If A then B*, calls for two conditions to hold for it to be true. First, there must be an instance in which *A and B* hold, because the other possibilities to which conditional refers also hold for its negation, *if A then not B*. Second, there must be no instances in which *A* and *not B* hold, because they refute the conditional. The hypothesis about quarks therefore demands the existence of quarks that form composites, and the non-existence of quarks that do not form composites. So, a duck-billed platypus is irrelevant to the truth or falsity of the hypothesis.

6. Deductions of spatial relations

The inferences in the previous section concern relations between clauses, but many sorts of deduction depend on relations within them. These relations can occur in scenes,

diagrams, and descriptions, and people can make deductions from any of these sources. Deductions from descriptions of temporal relations are complicated, because they depend on several distinct features of language – tense and aspect, connectives such as "before" and "during" (e.g., Kelly, Khemlani, & Johnson-Laird, 2020), and the temporal consequences of different sorts of verb (Schaeken, Johnson-Laird, & d'Ydewalle, 1996). Likewise, when individuals make deductions from descriptions of algorithms that carry out permutations of a sequence of entities, they rely on kinematic models in which spatial relations change over time (see Khemlani et al., 2013).

Simple but representative cases of relational deductions concern spatial layouts. Consider this inference (from Johnson-Laird, 1975):

> The black ball is directly beyond the cue ball.
>
> The green ball is on the right of the cue ball, and there is a red ball between them.
>
> So, if I move so that the red ball is between me and the black ball, then the cue ball is to the left.

The deduction is deictic in that it depends on the speaker's point of view. It also depends on deictic interpretations of phrases such as "on the right". It is possible to frame axioms that capture their logical properties, and to use logic to make such deductions. But, the evidence is overwhelming that naive individuals base their inferences instead on mental models of spatial layouts (Byrne & Johnson-Laird, 1989; Knauff, 2013; Ragni & Knauff, 2013; Tversky, 1993). Both sets of authors have developed simulations for deictic spatial deductions.

The first model-based algorithm of spatial deductions illustrates the principal functions that simulations need in order to use models to make inferences. Its parser constructs a representation of the meaning of each premise. For the premise:

The triangle is on the right of the circle

it constructs a semantics that specifies which axis is incremented in order to locate the triangle in relation to the circle, i.e., keep adding 1 to the value on the left-right axis of the location of the circle, and hold its values on the front-back and up-down axes constant. The code representing this semantics is used in all the main functions for constructing and manipulating models (see Table 1).

What happens in the simulation depends on the current context, i.e., on which entities, if any, are already represented in a model. This context can elicit any one of seven basic procedures, which are typical for deductions in general. Three of them occur in the processes of system 1:

1. **Start a new model**. The procedure inserts an item representing a referent into a new model.
2. **Update a model with a new referent**. The procedure puts an item representing the new referent into the model according to its relation to a referent already there.
3. **Update a model with a new relation.** The procedure puts it into the model provided that is consistent. Otherwise, it returns the empty model, but system 2 calls procedure (7) below.
4. **Validate** whether an assertion about a relation between referents is true or false in existing models. System 1 returns the truth value. If it is true, system 2 calls procedure (6) below, which searches for a model that is a counterexample to the

assertion; if it is false, system 2 calls procedure (7) below, which searches for an example of the assertion.

The remaining three procedures depend on access to more than one model, and therefore occur only in system 2:

5. Combine two existing models into one according to a relation holding between a referent in one model and a referent in another model.

6. Search for a counterexample, i.e., a model in which an assertion is false. If the search fails then the assertion follows as necessary from the previous premises. If the search succeeds then the assertion follows only as a possibility.

7. Search for an example, i.e., a model in which the assertion is true. If the search fails then the assertion is inconsistent with the previous assertions, and it is retracted. In some simulations, this result elicits a defeasible component that amends the premises and searches for a causal explanation that resolves the inconsistency (see, e.g., Johnson-Laird et al., 2004). If it succeeds then the assertion follows as a possibility.

Table 1 provides examples of how these procedures operate for spatial reasoning.

One point bears emphasis. The simulation of system 2's searches for counterexamples and examples works because the system has access to the representations of the semantics of a premise. Without this access, it would be impossible for the system to keep track of whether or not an alternative model still represents the premises. When a description is consistent with more than one layout, system 1 builds whichever model requires the least work.

This idea lies at the heart of PRISM, a more recent model-based simulation of two-dimensional spatial deductions (Ragni & Knauff, 2013). It implements such

reasoning using principles similar to those of the earlier algorithm, e.g., its initial preferred mental models are constructed without disturbing the arrangement of entities already in the model. But, PRISM introduces several innovations. The most important is that its prediction of the difficulty of an inference reflects, not the search for an alternative model, but the number of operations required to construct it, which depends on local transformations of the initial model. Those models that call for a longer sequence of these transformations are therefore likely to be overlooked. The source code of both simulations can be found on the model theory's website (https://modeltheory.org/models/).

The spatial algorithms have no need for postulates to capture logical postulates of relations, such as the transitivity of the deictic sense of "on the right of", because they are emergent properties from the use of meanings to construct models. Hence, a model of these two assertions:

The triangle is on the right of the circle.

The circle is on the right of the square.

yields the transitive conclusion:

The triangle is on the right of the square.

No model of the premises is a counterexample to it, and so it follows necessarily.

This emergence of logical properties has a further advantage in that it accounts for a different sort of spatial reasoning – deductions that depend on the intrinsic parts of entities (see Miller & Johnson-Laird, 1976, Sec. 6.1.3). Consider these assertions:

Matthew is on Mark's right.

Mark is on Luke's right.

Luke is on John's right.

They can refer to the deictic positions of the four individuals from the speaker's point of view, but they can also refer to their positions in terms of the intrinsic right-hand sides of human beings. A model of these spatial relations depends, first, on locating Mark, then using his bodily orientation to establish the intrinsic axes that specify his right-hand side. The same sort of simulation to the deictic ones above can then insert a representation of Matthew on the lateral plane passing through the right-hand side of Mark. So, if the four individuals are seated down one side of a rectangular table (as in Leonardo's *Last Supper*) then the transitive conclusion, *Matthew is on John's right*, follows. But, if they are seated around a circular table, transitivity depends on the size of the table, and on how close they are sitting to one another, e.g., Matthew could be sitting opposite John, or even on his left-hand side. These vagaries reflect those of the different situations (Johnson-Laird, 1983, p. 261), and no known simulations of this sort of spatial inference exist.

7. Deductions with quantifiers

Quantifiers are phrases such as, *all musicians*, *some painters*, and *no sculptors*. The most complex inferences depend on quantifiers, and the mReasoner program simulates several sorts of quantified deductions (see Khemlani & Johnson-Laird, 2020, and the model theory's website for the program). The simulation treats quantified assertions as relations between sets – an idea that goes back to Boole (1854) and that was adopted early in the development of the model theory, because it is the only way that models can have the same structure as the situations that they represent (Johnson-Laird, 1983, p. 137 et seq.). So, the meaning of the assertion:

Some musicians are painters

is that individuals exist common to both sets. This semantics generalizes to quantifiers that cannot be defined in orthodox predicate logic, such as: “more than half the musicians”. Table 3 presents a representative set of quantifiers and their set-theoretic meanings, which a computational model implements. Its intuitive system works with a single model at a time. It can construct various models of a given assertion in order to accommodate differences in reasoning between individuals and within individuals from one occasion to another. A typical model of the quantified assertion above is:

```
musician    painter
musician    painter
musician
            painter
```

Each row represents a different possible individual who exists in default of knowledge to the contrary. If neither individual of the sort represented in the first two rows exists then the assertion is false.

Table 3. Representative quantified assertions, and their set-theoretic meanings in formal notations and informal paraphrases, where *A* and *B* denote sets of entities.

Quantified assertions	**Set-theoretic meanings**	**Informal paraphrases**
All A are B.	$A \subseteq B$	Set A is included in set B.
Some A are B.	$A \cap B \neq \emptyset$	Intersection of A and B is not empty.
No A is a B.	$A \cap B = \emptyset$	Intersection of A and B is empty.
Some A are not B.	$A - B \neq \emptyset$	Set of A’s that are not B’s is not empty.
Most A are B.	$\lvert A \cap B \rvert > \lvert A - B \rvert$	Cardinality of intersection of A and B is greater than that of A’s that are not B’s.
More than half of A’s are B’s.	$\lvert A \cap B \rvert > \lvert A \rvert / 2$	Cardinality of intersection of A and B is greater than that of half of A’s.

The simulation elucidates how individuals draw immediate inferences from one quantified assertion to another, such as the inference from *All A are B* to the intuitive conclusion *All B are A*, which is possible but not necessary, and to the deliberative conclusion, *Some B are A*, which is necessary granted that *A*'s exist. As in the spatial algorithm, the simulation can add information from a subsequent assertion to update a model (see Table 1). Hence, the following premises are those for a *syllogism*, that Aristotle was the first to study, and that has had a long influence on logic and on psychological studies of deduction:

Some musicians are painters.

All painters are imaginative.

The second premise updates the model above of the first premise to yield the following typical model:

```
musician    painter     imaginative
musician    painter     imaginative
musician
            painter     imaginative
```

The intuitive system 1 relies on heuristics in order to scan the model in order to draw a conclusion. One heuristic reflects the order in which the model is constructed, and another reflects the traditional idea that a negative premise calls for a negative conclusion, and a premise with "some" calls for a conclusion with "some". As a result, system 1 delivers this conclusion from the model above: *some musicians are imaginative*.

The deliberations of system 2 can search for an alternative model of the premises, and if they find one, they can attempt to formulate a new conclusion that satisfies all the current models of the premises. This search relies on the sorts of operation that individuals used when they reasoned with different cut-out shapes to represent different

individuals, e.g., their most frequent operation was to add a new sort of individual to a model, albeit one consistent with the premises (see Bucciarelli & Johnson-Laird, 1999, Experiment 3). The resulting simulation gives a more accurate account of syllogistic reasoning than other rival theories (Khemlani & Johnson-Laird, 2020, and for descriptions of these theories, see Khemlani & Johnson-Laird, 2012). It also allows for deductions about possible sorts of individual, e.g.:

It is possible that only musicians who are painters are imaginative.

No complete simulation of reasoning with quantifiers exists. And the completion of the present account needs a solution to the recursive structure of quantifiers, as in these examples:

Every one of more than three of the seven girls…

Most of the teachers of all the children of some of the employees…

It needs an account of multiple quantifiers in an assertion (Johnson-Laird, 2006, Ch. 11), as in the following sequence of two deductions:

Chuck loves Di.

Everyone loves anyone who loves someone.

So, everyone loves Chuck.

So, everyone loves everyone.

It needs an account of quantified properties, whose analysis in logic calls for the "second order" predicate logic (see Jeffrey, 1981, Ch. 7):

Some member of the Royal family has all the desirable properties of a princess.

One desirable property of a princess is to be beautiful.

So, some member of the Royal family is beautiful.

Finally, it needs an account of inferences hinging on connectives and quantifiers, e.g.:

Either Chuck loves Di or he doesn't.

Everyone loves anyone who loves someone.

So, either everyone loves everyone or no-one loves anyone.

The conclusion follows of necessity from the premises, but the inference is difficult because it depends on the repeated updating of models of the premises. For example, if Chuck loves Di, then everyone loves Chuck. The second premise above can be used again to update the model of this situation in order to represent that everyone loves everyone (see Cherubini & Johnson-Laird, 2004).

8. Deductions of probabilities

Some psychologists argue that deductions depend, not on logic, but on probabilities – an approach called the "new paradigm" (see, e.g., Oaksford & Chater, 2020). One crux is the new paradigm's treatment of the probability of conditionals. It takes the probability of *If A then B* to equal the conditional probability of *B* given *A*, an equality that philosophers sometimes refer to as "the Equation". For the model theory, the probability of a conditional should also fit the Equation provided individuals bear in mind that cases of *not-A* are presuppositions. As described in Section 5, a conditional, *if A then B*, presupposes the possibility of *not-A*, which therefore holds for the negation of the conditional. It follows that the probability of the conditional is the proportion of cases of *A* in which *B* occurs, because cases of *not-A* are irrelevant. Unlike the new paradigm, however, the model theory postulates that probabilities underlie inferences only when tasks implicate them, and evidence corroborates this assumption. Individuals

deduce different conclusions from: *If the wine is Italian then it is red* than from *If the wine is Italian then it is probably red* (Goodwin, 2014).

A long-standing puzzle, which the new paradigm does not solve, is how people deduce numerical probabilities from assertions that make no reference to them. One way is "extensional" (Tversky & Kahneman, 1983). They assume in default of knowledge to the contrary that each model represents an equiprobable possibility, and deduce the probability from the proportion of models of these exhaustive possibilities in which the event occurs, or from the sum of the frequencies of each of these possibilities (Johnson-Laird et al., 1999). For example, the assertion:

> 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

has the following two mental models of what is in the box:

```
red
        green   blue
```

On the assumption that the two models are equiprobable, they yield a probability of ½ that the box contains a green and a blue marble, and a probability of zero that it contains a red and a green marble. An experiment corroborated these predictions. However, the fully explicit models of the assertion are:

```
  red    green ¬ blue
  red  ¬ green   blue
  red  ¬ green ¬ blue
¬ red    green   blue
```

They show that the two previous probabilities should both be ¼. So, as other findings corroborated, mental models predict deductions of extensional probabilities, and granted that models are equiprobable, system 2 yields valid deductions of them. These predictions follow from a computational simulation (https://modeltheory.org/models).

No extensional method is feasible to deduce the probability of a unique event, such as:

Trump is re-elected President of the US.

A big mystery about such inferences, which people are happy to make, is where the numbers come from and what determines their magnitudes. A theory and a computer implementation of it solve the mystery (Khemlani, Lotstein, & Johnson-Laird, 2015). The program deduces the probability of a unique event in the same way as an extensional deduction except that the models it uses are not of the event, but of evidence pertinent to it. The first step of inferring, say, the probability of Trump's re-election is to call to mind relevant evidence, such as:

Most incumbent US Presidents who run again are reelected.

Individuals build a single mental model of such incumbents to represent this belief:

```
incumbent    reelected
incumbent    reelected
incumbent    reelected
incumbent
```

The first three rows represent incumbents who are reelected, but the last row represents an incumbent who is not reelected. The numbers of individuals in the model are not fixed, and can be modified during an inference, or even tagged with deduced numerical values from other evidence, provided that they do not contravene the meaning of the assertion. Because Trump is an incumbent, the model can be sampled to yield a representation of the probability of his reelection. The intuitive system 1 constructs a representation of this probability. It is "pre-numerical" because it represents a magnitude in the same way as infants and non-numerate adults do (see, e.g., Carey, 2009). The

following diagram depicts the representation, in which for convenience the main axis is from left to right:

|–––––– |

The left vertical represents impossibility, the right vertical represents certainty, and the proportional length of the line between them represents a probability. It can be translated into a description such as: "The re-election of Trump is *very likely*: it is *highly possible*."

Individuals are likely to consider other evidence, such as:

Politicians who are known liars tend not to be re-elected.

The probability inferred from this evidence has to be combined with the previous probability. Most people do not know the correct way to form the conjunction of two probabilities. According to the model theory, they seek an intuitive compromise, and so the simulation sets up a pointer, ^, to represent the probability based on the second piece of evidence within the representation of the first probability:

|––^––– |

The simulation then shifts the pointer and the right-hand end of the line towards one another. The two meet at a point corresponding to their rough average. It represents the compromise probability of the event. The theory postulates that intuition uses the same procedure to deduce the probability of a disjunction from the probabilities of its two clauses.

In contrast, the deliberative system 2 can map analog magnitudes representing probabilities into numerical values. The major impediment to the rationality of system 2 is ignorance. Individuals who have not mastered the probability calculus do not know how to compute the probability of compounds, such as conjunctions, disjunctions, or

conditional probabilities. They can grasp that the probability of the conjunction of two independent events is their product, that the probability of a disjunction of inconsistent events is the sum of their probabilities, and that the conditional probability of *A* given *B* is the subset of the possibilities of *B* in which *A* occurs. The algorithm embodies these principles, and experimental results have corroborated the errors in estimates that often violate the principles of the probability calculus (Byrne & Johnson-Laird, 2019; Khemlani et al., 2015).

9. Conclusions

Psychological theories of deductive reasoning can take too much for granted, so that what they predict about a particular inference is often difficult to figure out (Johnson-Laird, 1983, p. 6). They may not predict anything. It is too easy to construct psychological theories if they concern only what conclusions people make and not how they make them. For instance, the existence of over a dozen theories of syllogistic reasoning is embarrassing for cognitive science (see Khemlani, 2020). Few of them have computational simulations. Simulations of the model theory yielded surprising predictions about human rationality, such as inferences that are cognitive illusions (see Section 5).

An account solely of what the mind computes can be embarrassing in another way. Its computer implementation may reveal its intractability. For instance, several theories extend Ramsey's (1990, p. 155) idea of how to determine the credibility of a conditional: granted that its *if*-clause is consistent with a stock of knowledge, assess the likelihood of its *then*-clause in that same stock. Yet, a check of whether the *if*-clause is

consistent with a set, say, of ten beliefs takes far too long to be realistic. In the worst case, it can take 2^{10} assessments. A viable theory of deduction must explain how humans overcome such intractability. Hence, a prophylactic for all these problems is to ensure that a theory accounts for human mental processes too, and to develop a simulation of them. The preceding account shows how to base such simulations on mental models to capture people's intuitive mistakes, biases, and default assumptions, as well as their ability to overcome their intuitions.

References

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

Beth, E. W., & Piaget, J. (1966). *Mathematical epistemology and psychology*. Dordrecht, The Netherlands: Reidel.

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.

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

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

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

Byrne, R. M. J., & Johnson-Laird, P. N. (2019). *If* and *or*: real and counterfactual

possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 760–780.

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

Cherubini, P., & Johnson-Laird, P. N. (2004). Does everyone love everyone? The psychology of iterative reasoning. *Thinking & Reasoning*, *10*, 31-53.

Cook, S. A. (1971). The complexity of theorem proving procedures. *Proceedings of the Third Annual Association of Computing Machinery Symposium on the Theory of Computing*, *3*, 151–158.

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

Hempel, C. G. (1945). Studies in the logic of confirmation, Parts I and II. *Mind*, *54*, 1–26, 97–121. http://dx.doi.org/10.1093/mind/LIV.213.1

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.

Hughes, G. E., & Cresswell, M. J. (1996). *A new introduction to modal logic*. London: Routledge.

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

Johnson-Laird, P. N. (1975). Models of deduction. In Falmagne, R. (Ed.) *Reasoning: Representation and Process*. Springdale, NJ: Erlbaum. Pp. 7-54.

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

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

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

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. (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., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*, 130950.

Johnson-Laird, P. N., & Wason, P. C. (1970a). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*, 134–148. http://dx.doi.org/10.1016/0010-0285(70)90009-5

Johnson-Laird, P. N., & Wason, P. C. (1970b). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, 22, 49–61. http://dx.doi.org/10.1080/14640747008401901

Kelly, L., Khemlani, S., & Johnson-Laird, P.N. (2020). Reasoning about durations. Manuscript in press at the *Journal of Cognitive Neuroscience.*

Khemlani, S. (in press). Psychological theories of syllogistic reasoning. In M. Knauff and W. Spohn (Eds.), Handbook of Rationality. Cambridge, MA: MIT Press.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science*, 1-38. DOI: 10.1111/cogs.12634

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*, 427-457.

Khemlani, S., & Johnson-Laird, P. N. (2013). Cognitive changes from explanations. *Journal of Cognitive Psychology*, *25*, 139-146.

Khemlani, S., & Johnson-Laird, P.N. (2017). Illusions in reasoning. *Minds and Machines*, *27*, 11-35.

Khemlani, S., & Johnson-Laird, P. N. (2020). Reasoning about sets: a unified computational theory. Manuscript under review.

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

Khemlani, S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, *110 (42)*, 16766–16771. http://www.pnas.org/cgi/doi/10.1073/pnas.1316275110.

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

Newell, A. (1973). You can't play 20 questions with nature and win. In Chase, W.G., (Ed), *Visual information processing*. New York, NY: Academic Press.

Oaksford, M., &, N. Chater (1996). Rational explanation of the selection task. *Psychological Review*, *103*, 381–391.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 12.1–12.26. https://doi.org/10.1146/annurev-psych-010419- 051132

Osherson, D. N. (1974-6). *Logical abilities in children*. (Vols. 1-4). Hillsdale, NJ: Erlbaum.

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Ragni, M., Dames, H., & Johnson-Laird, P. N. (2019). A meta-analysis of conditional reasoning. In preparation.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological review*, *120*, 561-588.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses. *Psychological Bulletin*, *144*, 779-796. http://dx.doi.org/10.1037/bul0000146

Ramsey, F. R. (1990). *F. R. Ramsey, philosophical papers.* Mellor, D. H. (Ed.) Cambridge: Cambridge University Press.

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

Schaeken, W, Johnson-Laird, P. N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.

Sun, R. (2016). *Anatomy of the mind: Exploring psychological mechanisms and processes with the Clarion cognitive Architecture*. Oxford University Press, New York.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315.

Tversky, B. (1993) Cognitive Maps, Cognitive Collages, and Spatial Mental Models. In Frank, A.U. and Campari, I. (Eds.) Spatial Information Theory: A Theoretical Basis for GIS, Proceedings COSIT '93. Lecture Notes in Computer Science, 716,

pp.14-24, Springer: Berlin. Article · September 1993 DOI: 10.1007/3-540-57207-4_2

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Wason, P. C. (1969). Regression in reasoning? *British Journal of Psychology*, *60*, 471-480.