

# **Inside Psychology: a science over 50 years**

Edited by

Pat Rabbitt (Emeritus Professor,  
University of Manchester, UK)

Senior Research Associate, Department of Experimental  
Psychology, University of Oxford, UK

and

Senior Research Fellow, Department of Psychology,  
University of Western Australia, Perth, WA

**OXFORD**  
UNIVERSITY PRESS

---

## Reasoning

Phil Johnson-Laird

Why do researchers become interested in one topic rather than another? The answer, at least for me, seems to have more to do with events in life than with personality or intellect. In 1961, I arrived at University College London (UCL) to study psychology. It came about in this way. One day when I was still at school, my father told me that he could no longer afford to go on paying the fees and that I would have to leave. With great disappointment, I left at the end of the term. I also left without being able to take the examinations necessary to enter university ('O-levels' and 'A-levels'). I worked for five years under contract as a quantity surveyor—a job whose tedium was alleviated only by my moonlighting as a pianist in a modern jazz quintet. (Our bass player, Ian Keen, is now a distinguished anthropologist.) When my contract expired, I quit surveying. My deferment from national service in the military expired, and I was immediately called up. One of my intellectual heroes, however, was Bertrand Russell, and his arguments had convinced me that Britain's possession of nuclear weapons was a moral and political mistake. So, I was a conscientious objector, and the tribunal that heard my case sent me to work for 'two years and sixty days in hospitals and other vital services'.

Towards the end of this period when I was working in a bakery, I got married, and my wife and I realized that I needed a career. I wanted a job that would be interesting, and the way to get one, I thought, was to go to university. I should have liked to study science, but I needed O-levels and A-levels, and had to work for them by myself with no access to a laboratory. So, I had to study 'arts' subjects at A-level. But, an arts subject at university didn't seem likely to lead to an interesting job—a friend of mine had sat at the feet of Leavis in the Cambridge English Department, and now worked in a public library. The idea of a career within academia struck me as impossible—a realistic assessment, I think, for someone nearly 25 years old with no qualifications. I considered studying philosophy—the influence of Russell, again—but it too seemed unlikely to prepare me for a stimulating job. Then I discovered psychology. It led on to all sorts of interesting possibilities, from ergonomics

to clinical practice. And so, as a result of these deliberations, I applied to UCL and, to my considerable amazement, was accepted.

For my subsidiary subject, I chose logic—in part, under the influence of Russell, and in part because I thought (wrongly) that it was easy—I had passed it at A-level after six weeks of study. The late Bernard Williams was the lecturer in introductory logic, and his wit and enthusiasm boosted my interest. But I discovered, as a subject in one of the late Peter Wason's experiments, that I was as susceptible as everyone else to logical error. Peter tested his own subjects, sitting in the corner of the room, smoking a pipe, and radiating more than a passing resemblance to Sherlock Holmes. Afterwards we chatted, and he explained where I had gone wrong. It was fascinating, and my interest in the psychology of thinking probably dates to that conversation.

The course at UCL had no lectures on reasoning, but textbooks had something to say on the topic: the 'atmosphere' of premises—the choice of words such as 'all' or 'none'—biased reasoners to draw certain conclusions rather than others. The texts, however, had nothing to say about the mental processes of reasoning. For that, one had to go to continental psychology and the works of Piaget. His theory was not easy to understand—an ominous sign was the existence of exegetical works purporting to explain it. But, in essence, he proposed that children's intellectual development culminated in a set of principles akin to those of formal logic (Inhelder and Piaget 1958).

I graduated well enough to be accepted for UCL's postgraduate programme, and Wason agreed to be my adviser. He had just returned from an exciting year at the Harvard Center for Cognitive Studies, co-directed by Jerome Bruner and George Miller, and he had carried out a pioneering study into the effects of negation in sentences, which helped to initiate the study of psycholinguistics. As an adviser, he had a miraculous way of getting his students to work on topics that interested him, without any overt direction on his part. And so I found myself working on the pragmatics of the passive voice. I didn't realize it at the time, but Wason was an extraordinary psychologist. He had a *genius*—no other word will do—for devising provocative experimental tasks. That was the one big purpose of his work, and so he had little interest in theories. His dictum was: psychologists should never quite know why they are carrying out an experiment. It took me a while to grasp that his way of doing psychology was unique.

The 1960s were a good time for budding academics, because there were more jobs than people to fill them. No sooner had I completed my doctorate than I was offered a lectureship at UCL. And, at last, I turned to reasoning as a research topic. I discovered a precursor to Piaget, the nineteenth century logician, George Boole, who had described what he took to be the 'laws of thought'

in his algebra for the logic of *not*, *and*, and *or* (Boole 1854). The task for psychologists, it now seemed to me, was to carry out experiments to pinpoint the particular formal rules of inference in the mind. It would not be easy, because an indefinite number of different ways existed in which to formalize logic.

The idea of a tacit mental logic continues to have its adherents (e.g. Rips 1994), and it is implemented in many computer programs for proving logical theorems. But, Wason, though he did not realize it at first, had already made a major dent in this approach to reasoning. He had devised an ingenious task that required individuals to select evidence that could refute a hypothesis. Inhelder and Piaget (1958) had written that, given a hypothesis of the form:

*If p then q*

individuals should try to refute it by searching for a counter-example, namely the conjunction of p and not-q. Their view was that reasoning was '*nothing more than the propositional calculus*' (p. 305)—an interpretation of Boolean algebra in which variables such as p and q have values that are propositions. In Wason's (1966) experiment, the participants had to test the hypothesis:

*If a card has a vowel on one side then it has an even number on the other side.*

They had to select those cards that needed to be turned over to find out whether this hypothesis was true or false about four cards laid out on the table in front of them: A, B, 2, 3. They knew that each card had a letter on one side and an even number on the other. According to Piaget, they should select the A card (the p in *if p then q*), and the 3 card (not-q). They did indeed select A, but they almost all failed to select the not-q card.

Wason and I worked together for three years to try to find out what was going on. His most striking discovery was that when the task concerned a journey, and the hypothesis was 'If I travel to Manchester then I go by car', the participants were more likely to make the correct selections of p and not-q—the cards bearing 'Manchester' and 'train' (see Wason and Johnson-Laird 1972). Likewise, Paolo Legrenzi, Maria Sonino Legrenzi, and I observed a striking improvement when the hypothesis concerned potential violations to a postal regulation akin to one in force in the UK: 'If a letter is sealed then it has a 5 penny stamp on it.' We had unwittingly invented the so-called 'deontic' selection task. What struck us, however, was the utter failure of correct performance on this task to transfer to the standard selection task (Johnson-Laird *et al.* 1972).

The selection task has launched a thousand studies, but no consensus yet exists about what ability it taps. At the time, however, I had no doubt that the

effect of content was an embarrassment to the view that the mind relies on a formal logic. Formal logic is blind to content, and so Wason's discovery marked a turning point in the study of reasoning. The question for psychologists now became: is there a theory that accounts for the effects of content? Such a theory now exists, but it emerged only over the course of some years. In pursuit of an answer, I spent a year working on the meanings of words with George Miller at the Institute for Advanced Studies in Princeton. Our paper on the topic turned over several years into a book (Miller and Johnson-Laird 1976)—a book so long that few individuals even claim to have read it.

With help from my work with Miller, I developed a theory of reasoning based on the idea that we understand the *meaning* of what we reason about. We use the meanings of words, the grammatical relations amongst the words, and general knowledge to compose the meanings of the premises. These meanings enable us to construct mental models of the possibilities compatible with the premises (Johnson-Laird 1983). Craik (1943) had argued that we construct mental models of the world in order to anticipate events, but he had taken for granted that reasoning depends on verbal rules. Several other theorists proposed that we represent discourse in models (e.g. van Dijk and Kintsch 1983), and the idea is no longer controversial (Garnham 2001). However, once we have constructed models to represent the situations that a discourse describes, we could use the same models as a basis for reasoning. A conclusion is valid if it holds in all the models of possibilities consistent with the premises. And the theory predicts that the more possibilities we have to represent, the harder reasoning should be. It therefore offers an explanation of errors in reasoning, and contrasts with the idea—revived in the 1990s—that naive individuals don't reason at all, but are either prey to 'atmosphere' effects or rely on probabilistic considerations. The current popularity of Sudoku puzzles seems to refute these accounts: the solution of the puzzles depends on pure deduction.

Consider this inference:

*None of the artists is a beekeeper.*

*All the beekeepers are chemists.*

*What follows?*

Few of us draw the logically correct conclusion:

*Some of the chemists are not artists.*

Why is the inference so difficult? The answer, according to the model theory, is as follows. We start by envisaging a possibility in which the first premise is true. We construct a model symbolized in the following diagram, where each

line represents a separate individual, and the number of individuals is small but arbitrary:

<i>artist</i>		
<i>artist</i>		
	<i>beekeeper</i>	
	<i>beekeeper</i>	

Two of the individuals in the model are artists and two of them are beekeepers, but, of course, a real mental model represents individuals, not words, which I use here for simplicity. We use the second premise to update the model in as simple a way as possible:

<i>artist</i>		
<i>artist</i>		
	<i>beekeeper</i>	<i>chemist</i>
	<i>beekeeper</i>	<i>chemist</i>

This model suggests the conclusion that none of the artists is a chemist, or its converse, and many of us do draw these invalid conclusions. The 'atmosphere' of the premises supports these conclusions too, but according to the model theory they arise from the process of reasoning itself. In order to reach the correct conclusion, we need to realize that there can be chemists who are not beekeepers, and to envisage that these chemists could be artists:

<i>artist</i>		<i>chemist</i>
<i>artist</i>		<i>chemist</i>
	<i>beekeeper</i>	<i>chemist</i>
	<i>beekeeper</i>	<i>chemist</i>

This model refutes the conclusion that none of the artists is a chemist, and its converse. Yet, it does yield a conclusion also supported by the initial model: some of the chemists are not artists, namely, those who are beekeepers. In contrast, those problems that yield only a single model of the premises are easy for us, and even for children.

When I moved to the Medical Research Council's Psychological Research Unit in Cambridge, UK, Ruth Byrne came from Trinity College, Dublin, to work with me. We began with a study of simple spatial reasoning, which corroborated the model theory's main prediction. It was easier for the participants to reason from descriptions compatible with one layout than from descriptions compatible with multiple layouts (Byrne and Johnson-Laird 1989).

Subsequent studies showed the same effect for reasoning about temporal relations amongst events (e.g. Schaeken *et al.* 1996). But the main problem that Byrne and I confronted was to extend the model theory to the analogues of Boole's connectives in natural language: *if*, *or*, and *and*. It took three separate steps.

The first step was to postulate that individuals construct models of the possibilities compatible with assertions containing connectives. An 'exclusive' disjunction, such as:

*There is a king in the hand or else there is an ace, but not both.*

is compatible with two possibilities. In one there is a king in the hand (and not an ace), and in the other there is an ace in the hand (and not a king). In contrast, an 'inclusive' disjunction, such as:

*There is a king in the hand or else there is an ace, or both.*

is compatible with three possibilities: the two preceding ones, and the possibility in which both the king and the ace are in the hand. The model theory accordingly predicts that reasoning from an exclusive disjunction should be easier than reasoning from an inclusive disjunction. The prediction is crucial, because theories based on formal rules of inference make the opposite prediction. They treat an exclusive disjunction as a calling for an additional inference over and above an inclusive disjunction (e.g. Rips 1994). The results corroborated the model theory (Johnson-Laird and Byrne 1991; García-Madruga *et al.* 2001).

The second step depended on a major assumption: the principle of truth, which stipulates that mental models represent only what is true. Hence, the exclusive disjunction above has the mental models shown in this diagram, where each line denotes a separate possibility:

<i>king</i>	
	<i>ace</i>

Here, 'king' denotes that there is a king in the hand, and 'ace' denotes that there is an ace in the hand. Indeed, when individuals are asked to list the possibilities compatible with the assertion, they tend to list just these possibilities (Johnson-Laird and Savary 1999). However, the first model contains no information about the ace, and the second model contains no information about the king. 'Fully explicit' models of the two possibilities represent this information:

<i>king</i>	<i>not-ace</i>
<i>not-king</i>	<i>ace</i>

where 'not' is used to show that the corresponding affirmative propositions are false. In other words, the force of 'or else' is that one proposition in the disjunction is true and the other proposition is false. Only fully explicit models, however, represent the status of both propositions in both possibilities. The principle of truth eases the load on working memory, but it exacts an unexpected cost.

At each stage in its development, we implemented the model theory in computer programs, and just occasionally the output of these programs surprised us. The biggest surprise came from a program based on the principle of truth. Its output contained what seemed to be an egregious error. The premises were:

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

*There is a king.*

When the program followed the principle of truth, it represented the first premise in these mental models:

<i>king</i>	<i>ace</i>
<i>not-king</i>	<i>ace</i>

Given the second premise—the categorical assertion that there is a king—the program eliminated the second model and drew the conclusion: there is an ace. However, when the program used fully explicit models, representing both what is true and what is false, it drew the bizarre conclusion that there is *not* an ace.

Nearly everyone draws the conclusion that there is an ace (Johnson-Laird and Savary 1999). Yet, it is an illusion, and the program's conclusion from fully explicit models is correct: there is not an ace. To understand why, you need to recall two assumptions that I have already made. The first assumption is that one proposition in an exclusive disjunction is true and the other proposition is false—they can't both be true. The second assumption is that a conditional of the form, if p then q, is false in the possibility in which p and not-q occur. Granted that the first premise in the inference above is an exclusive disjunction, it can be abbreviated as:

*If king then ace, or else if not-king then ace.*

Suppose that the first conditional, if king then ace, is true. The second conditional is therefore false (i.e. not-king and not-ace both hold). This case is compatible with the truth of the first conditional, and so one possibility is:

<i>not-king</i>	<i>not-ace</i>
-----------------	----------------

Now, suppose that the second conditional, if not-king then ace, is the one that is true. The first conditional is therefore false (i.e. king and not-ace both hold).

This case is compatible with the truth of the second conditional, and so another possibility is:

*king not-ace*

The premises allow only these two possibilities, and so, even granted the presence of the king, it follows that there is not an ace.

You may think of an alternative rationale leading reasoners to infer to the contrary, that there is an ace. Perhaps they interpret the disjunction, not as exclusive but as inclusive. Perhaps they interpret the conditional as implying its converse. But, even granted either of these interpretations, or both of them, it still doesn't follow that there's an ace. Another possibility is that reasoners take the first premise to mean:

*If there is a king or if there isn't a king then there is an ace.*

Yet, the fallacy occurs even when the two conditionals are stated separately and the participants are told, 'One of these assertions is true and one of them is false'. A more powerful result, however, is that illusions of many other sorts are predicted by the principle of truth, and reasoners are highly susceptible to them (Johnson-Laird 2006). A theory based on formal rules of inference might be able to explain the illusions. So far, no such theory has been forthcoming.

The third and most difficult step in formulating the model theory was to give a proper account of conditionals, that is, sentences of the form: if p then q, which have perplexed philosophers for millennia. Byrne and I, however, assumed that their complexities arise from interactions among a number of simple components (Byrne 2005; Johnson-Laird and Byrne 2002). We proposed that the core meaning of a conditional, such as 'If there is a king then there is an ace', is compatible with three possibilities:

<i>king</i>	<i>ace</i>
<i>not-king</i>	<i>ace</i>
<i>not-king</i>	<i>not-ace</i>

In fact, children start by interpreting conditionals as compatible with just the first of these possibilities, later they add the third possibility, and by early adolescence they list all three possibilities (Barrouillet *et al.* 2000). But, the meanings of the clauses in conditionals, and general knowledge, can modulate the core interpretation. One effect of modulation is to prevent the construction of a possibility. For example, the conditional, 'If they played a game then it wasn't soccer', is compatible with only two possibilities: in one, they played a game that wasn't soccer, and in the other they didn't play a game.

Another effect of modulation is to establish various relations between the situations described in a conditional. For example, the conditional, 'If she put the ball onto table then it rolled off' is compatible with a temporal and spatial scenario in which she put the ball on the table and then it rolled off on to the surface below the table. The upshot is that the system for interpreting conditionals and other connectives, such as *and* and *or*, must take into account the meanings of the clauses that they interconnect, the entities that the clauses refer to, and general knowledge. The system cannot work in the way in which logic assigns interpretations to connectives, which concerns only whether propositions are true or false.

An alternative account of conditionals is that the if-clause invites us to make a supposition, and that we evaluate the then-clause in this hypothetical case (e.g. Evans and Over 2004). Some conditionals elicit suppositions, but not all do. As someone once said to me in Manhattan, 'If it's as hot as this now [in April] then it will be even hotter in the summer.' There was nothing hypothetical about the situation described in the if-clause. The model theory allows that individuals often make suppositions about clauses in conditionals and in other sorts of assertion (van der Henst *et al.* 2002). But suppositions cannot be the whole story. They fail to explain why individuals list three possibilities for simple conditionals, and why, as Ormerod and his colleagues have shown, they paraphrase a conditional of the form: if not p then q, as a disjunction, p or q, and vice versa (e.g. Ormerod and Richardson 2003).

Logic is built on Boolean connectives and on quantifiers such as 'all' and 'none', and so once the model theory had an account of reasoning based on these terms researchers began to investigate how it might be extended to other sorts of reasoning. It is impossible to describe all these developments, and so I mention only three diverse examples: reasoning about what is permissible (Bucciarelli and Johnson-Laird 2005); reasoning about relations, including those that appear to be transitive but are not, such as 'is a blood relative of' (Goodwin and Johnson-Laird 2005); and reasoning in psychological illnesses (Johnson-Laird *et al.* 2006).

Fifty years ago, cognitive psychology was in a nascent state. The story that I have told here is about a single strand in its subsequent development. I have focused on one approach to reasoning—the idea that it depends on constructing mental models of situations, from either perception or discourse. During the past decade, the theory has burgeoned, although it remains controversial. It began as a theory of deductive reasoning, but it now offers explanations of other sorts of reasoning—inductive reasoning, probabilistic reasoning, the detection and resolution of inconsistent beliefs, and the reasoning that underlies our ability to solve problems. The theory could be wrong. But, it has two

strong empirical supports. The first is the ability of individuals to list what is possible given a description—this simple task lies beyond the scope of most alternative theories. The second is the consequences of the principle of truth: reasoning on the basis of mental models leads to systematic illusions. Finally, what makes the selection task difficult? It may be the lack of familiar counter-examples. The participants in the postal experiment were familiar with what violated the regulation. Those who don't know the regulation tend to err.

## References

- Barrouillet, P., Grosset, N., and Leças, J. F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition* 75: 237–66.
- Boole, G. (1854). *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*. London: Walton and Maberley.
- Bucciarelli, M. and Johnson-Laird, P. N. (2005). Naive deontics: a theory of meaning, representation, and reasoning. *Cognitive Psychology*, 50, 159–193.
- Byrne, R. M. J. (2005). *The rational imagination: how people create alternatives to reality*. Cambridge, MA: MIT Press.
- Byrne, R. M. J. and Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language* 28: 564–75.
- Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
- Evans, J. St. B. T. and Over, D. E. (2004). *If*. Oxford: Oxford University Press.
- García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, E., and Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology* 54A: 613–32.
- Garnham, A. (2001). *Mental models and the interpretation of anaphora*. Hove: Psychology Press.
- Goodwin, G. and Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review* 112: 468–93.
- Inhelder, B. and Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge & Kegan Paul.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (2006). *How we reason*. Oxford: Oxford University Press.
- Johnson-Laird, P. N. and Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Johnson-Laird, P. N. and Byrne, R. M. J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review* 109: 646–78.
- Johnson-Laird, P. N. and Savary, F. (1999). Illusory inferences: a novel class of erroneous deductions. *Cognition* 71: 191–229.
- Johnson-Laird, P. N., Legrenzi, P., and Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology* 63: 395–400.
- Johnson-Laird, P. N., Mancini, F., and Gangemi, A. (2006). A hyper emotion theory of psychological illnesses. *Psychological Review* 113: 822–41.
- Miller, G. A. and Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.
- Ormerod, T. C. and Richardson, J. (2003). On the generation and evaluation of inferences from single premises. *Memory & Cognition* 31: 467–78.
- Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- Schaeken, W. S., Johnson-Laird, P. N., and d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition* 60: 205–34.
- van der Henst, J.-B., Yang, Y., and Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science* 26: 425–68.
- Van Dijk, T. A. and Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
- Wason, P. C. (1966). Reasoning. In: Foss, B. M. (ed.) *New horizons in psychology*, pp. 135–51. Harmondsworth: Penguin.
- Wason, P. C. and Johnson-Laird, P. N. (1972). *The psychology of deduction: structure and content*. Cambridge, MA: Harvard University Press.