Cognition, 10 (1981) 139–143

Cognition, computers, and mental models

P. N. JOHNSON-LAIRD*
University of Sussex

Two questions ought to haunt any student of cognition. First, is it possible to achieve a scientific understanding of the mind? It is to be hoped, of course, that a complete science is impossible since it would probably destroy the consciousness of free will. Second, are there profound uniformities in the ways in which the mind works? For example, if we understood how speech is perceived, would we thereby advance our understanding of, say, the visual perception of shapes? These two questions are presumably related in that a positive answer to the second is likely to lead to a positive answer to the first. My own research certainly inclines me towards supposing that indeed there are underlying uniformities in thought. For many years, I worked alternately on reasoning and comprehension. If I became stuck in one area, then I would switch to the other. Unfortunately, this strategy has recently been denied to me by the discovery of an underlying communality in the two areas. It concerns the role of mental models.

My first inkling that mental models might be important in comprehension is reflected in the following:

> It is possible that from the meanings of sentences in a connected discourse, the listener implicitly sets up a much abbreviated and not especially linguistic model of the narrative, and that recall is very much an active reconstruction based on what remains of this model. Where the model is incomplete, material may even be unwittingly invented to render the memory more meaningful or more plausible—a process which has its parallel in the initial construction of the model. A good writer or raconteur perhaps has the power to initiate a process very similar to the one that occurs when we are actually perceiving (or imagining) events instead of merely reading or hearing about them. (Johnson-Laird, 1970).

Bransford and his colleagues, of course, advanced a similar 'constructive' theory of comprehension and gathered convincing evidence in its support (see e.g., Bransford and McCarrell, 1975).

My second inkling about mental models arose from studying syllogisms (Wason and Johnson-Laird, 1972). These are simple deductive inferences of

*Reprint requests should be sent to P. N. Johnson-Laird, Centre for Research on Perception and Cognition, Laboratory of Experimental Psychology, University of Sussex, Brighton, BN1 9QG, England.

the form:

Some of the scientists are parents
All of the parents are drivers
Some of the scientists are drivers

Syllogisms are a nice test case for the feasibility of a cognitive science. There are only 64 possible forms for their premises, and if we are ever to understand anything about mental processes, we ought to be able to understand how people draw conclusions from them. They were first studied experimentally at the turn of the century, yet we still have no complete understanding of how human beings cope with them. A seductive hypothesis is that there is some sort of mental logic, perhaps based on representations akin to Euler circles; for a time, I certainly subscribed to such a doctrine. A major problem with it, however, is that it gives no very ready account of either the 'figural' effect—subjects tend to draw conclusions like the one illustrated in the example above rather than its equally valid converse—or the systematic errors that they make (see Johnson-Laird and Steedman, 1978).

How do people mentally represent syllogistic premises? No psychologist was ever able to tell me, but several subjects reported that they formed images of the states of affairs described in premises. The pattern of errors that one observes is certainly compatible with the idea that subjects in general construct mental models of the premises, whether or not they take the form of images. For example, suppose you present your subjects with the task of drawing a conclusion from the following premises:

All of the artists are beekeepers
All of the chemists are beekeepers

They can form a mental model of the first premise by imagining an arbitrary number of artists and identifying each of them as a beekeeper. Since there may be beekeepers who are not artists, they, too, must be represented in the model. Its structure must accordingly take the following form:

a = b
a = b
a = b
(b)
(b)
(b)

where each 'a' represents an artist, each 'b' represents a beekeeper, and the parentheses indicate that an individual may, or may not, exist. In adding the information from the second premise to this model, a logically prudent subject should consider all the different ways in which it can be combined. Some subjects evidently consider only this combination:

a = b = c
a = b = c
a = b = c
(b)
(b)
(b)

They draw the invalid conclusion that *all of the artists are chemists,* or its invalid converse, *all of the chemists are artists.* Other subjects also consider the combination:

a = b = c
a = b = c
a = b
b = c
(b)
(b)

They refrain from the previous conclusions, but draw the equally invalid conclusion that *some of the artists are chemists,* or its invalid converse, *some of the chemists are artists.* Fortunately for the rational reputation of the human race, about half the subjects that we have tested evidently consider the further combination:

a = b
a = b
a = b
b = c
b = c
b = c

They correctly reply that there is no valid conclusion interrelating the artists and chemists. The figural effect remains something of a mystery. It could reflect an inherent directional bias in the structure of mental models, or alternatively, the process of forming an integrated model in working memory.

My current conception of comprehension is that there is an initial rapid translation of an utterance into its superficial linguistic form, followed by an optional process in which this representation is used in the construction of a mental model. The process of model building goes beyond the literal content of the utterance since it relies on inferences based on general and specific knowledge. It also relies on a 'procedural semantics' (see Miller and Johnson-Laird, 1976; Johnson-Laird, 1977) in which the meanings of words play an appropriate part in constructing models *ab initio,* adding information to them from subsequent utterances, verifying sentences with respect to them, and recursively manipulating them in order to check whether there is any way in which a sentence could be consistent (or inconsistent) with the prior discourse.

The claim that sentences can be verified with respect to mental models should not be taken as a species of 'verificationism': it is one thing to compare a sentence with a mental model, quite another to verify it in reality.

Most of these processes occur in inference, too. Indeed, the logical properties of many words emerge directly from the semantics that is required to construct models from them. There is no need to postulate rules of inference governing their behaviour. There is no need to postulate a mental logic: the same principle applies equally to sentential connectives such as *and* and *or*, and, as we have seen, to quantifiers such as *some* and *all*. The advantage of this approach is that it solves at a stroke the problems of which particular logic or logics are in the mind, how they are mentally specified, and how children acquire them. These issues are cut off without a source, because logic is banished from the mind. When one considers the follies and horrors of the human predicament, it may be tempting to suppose that rationality is thereby banished, too. However, the fundamental semantic principle governing both the truth of a general assertion and the validity of an inference is that there should be no counterexamples. Some people at least are aware of this principle, and the experimental evidence suggests that they search, in a more or less haphazard way, for models of premises that are inconsistent with the putative conclusions that they have drawn. It is important to emphasize that the search appears to be neither systematic nor exhaustive, because the absence of these characteristics is the best evidence we have that deductive thinking is not guided by mental logic.

The theories that my colleagues and I have developed in order to account for these phenomena have often been modelled in the form of computer programs. There are obvious analogies between the operations of the mind and the execution of a computer program—a relation that was not lost on Kenneth Craik (1943), who was the first psychologist to suggest that reasoning might consist of the manipulation of models of reality, and he was writing several years before the invention of the programmable digital computer. However, there is another more important reason for computer modelling. Theoretical intuitions are very valuable (to those that have them) but, if they are needed to work out what a theory predicts, there is a strong possibility that they are responsible for the predictions, and that the theory itself has no explanatory value. It is not a signpost, but a crutch on which the theorist leans in order to point the way. A simple criterion that avoids this danger is to check that those components of the theory that give rise to predictions are describable in the form of an *effective procedure*, i.e. they can be expressed in the form of a working computer program. This criterion does not imply that the mind is nothing but a computer. It may turn out that the mind uses functions that are not computable—it is easy enough to prove the existence

of non-computable functions. But this phenomenon would, of course, place a strong limitation on the possibility of a scientific psychology. Likewise, it may turn out that someone will succeed in refuting the thesis that all effective procedures are computable. At present, however, any scientific theory of the mind should certainly be restricted to an effective procedure. To abandon this criterion is to allow that theories can be vague, confused, and, like mystical doctrines, only properly understood by their proponents. Nevertheless, although I have sketched an optimistic answer to my initial question about possible uniformities in mental processes, the answer to my first question may be negative: there may be certain aspects of human mentality that cannot be captured in any theory that can be modelled by a computer program.

References

Bransford, J. D. and McCarrell, N. S. (1975) A sketch of a cognitive approach to comprehension: some thoughts about what it means to comprehend. In W. B. Weimar and D. S. Palermo (eds.), *Cognition and the Symbolic Processes.* Hillsdale, NJ, Erlbaum.

Craik, K. (1943) *The Nature of Explanation.* Cambridge, Cambridge University Press.

Johnson-Laird, P. N. (1970) The perception and memory of sentences. In J. Lyons (ed.), *New Horizons in Linguistics.* Harmondsworth, Middx., Penguin.

Johnson-Laird, P. N. (1977) Procedural semantics. *Cog.*, *7*, 189–214.

Johnson-Laird, P.N. and Steedman, M. J. (1978) The psychology of syliogisms. *Cog. Psychol.*, *10*, 64–99.

Miller, G. A. and Johnson-Laird, P. N. (1976) *Language and Perception.* Cambridge, Cambridge University Press; Cambridge, Mass., Harvard University Press.

Wason, P. C. and Johnson-Laird, P. N. (1972) *The Psychology of Reasoning.* London, Batsford.