Cognition, 50 (1994) 17–39
0010-0277/94/$07.00 © 1994 – Elsevier Science B.V. All rights reserved.

Why should we abandon the mental logic hypothesis?

Luca Bonatti*
Laboratoire de Sciences Cognitives et Psycholinguistique, 54, Boulevard Raspail, 75006 Paris, France
Philosophy Department, Rutgers University, PO Box 270, New Brunswick, NJ 08903-0270, USA

Abstract

Two hypotheses on deductive reasoning are under development: mental logic and mental models. It is often accepted that there are overwhelming arguments to reject the mental logic hypothesis. I revise these arguments and claim that they are either not conclusive, or point at problems which are troublesome for the mental model hypothesis as well.

1. Introduction

An old and venerable idea holds that logic is concerned with discovering or illuminating the laws of thought. Its psychological corollary is that a system of logic in the mind underlines our thinking processes. This thesis fits very well with representational views of the mind according to which cognitive processes are largely proof-theoretical. Within such a framework, it is a thesis about the structure of the vehicle of internal representations. In a nutshell, it holds that reasoning consists of operations on mental representations, according to logical rules implemented in procedures activated by the forms of the mental representations. Even if the thesis loomed around for centuries, there is still little convincing psychological evidence of the existence of a mental logic. Such evidence has mostly been accumulated in the last few years, and almost exclusively concerns propositional reasoning (Braine, Reiser & Rumain, 1984; Lea, O'Brien, Fisch, Noveck & Braine, 1990; Rips, 1983).

**Correspondence to:* L. Bonatti, Laboratoire de Sciences Cognitives et Psycholinguistique, 54, Boulevard Raspail, 75006 Paris, France.

The author is indebted to Martin Braine, Emmanuel Dupoux, Jerry Fodor, Jacques Mehler, and Christophe Pallier for comments on a first draft of this paper.

SSDI 0010-0277(93)00613-C

In the same years in which some results were beginning to appear, mental logic has been seriously challenged by an alternative – mental models – mostly due to the work of Johnson-Laird and his collaborators. Both hypotheses share the basic geography of cognition: also the mental models hypothesis is (*inter alia*) about the nature of the internal representations of deductive processes. They differ, however, on their supposed nature. Roughly, the mental model hypothesis claims that understanding a text consists of the manipulation of tokens representing concrete samples of entities in the world, and reasoning consists of the construction of alternative arrangements of tokens. No abstract rules should be needed to accomplish deduction. Thus, at least at first blush, while mental logic seems naturally to require a language of thought on whose formulas abstract rules apply, mental models seem to be able to dispense with it and substitute analog simulations for discrete manipulation of propositional-like objects (McGinn, 1989).

Originally, crucial aspects of the new hypothesis were left vague, and both its exact status and the feasibility of its claims were a puzzle (Boolos, 1984; Rips, 1986). What precisely a mental model is seemed to be a question of secondary importance, if compared to the big revolution introduced by the theory. Only recently has a substantial effort of formal clarification been undertaken (especially in Johnson-Laird & Byrne, 1991 and Johnson-Laird, Byrne, & Schaeken, 1992), but the task is still far from being accomplished (Bonatti, in press; Hodges, 1993). Nevertheless, the hypothesis had an enormous success, to the point that probably the words "mental models" are second only to "generative grammar" for their consequences within the cognitive science community. In a very short time, among psychologists an almost unanimous consensus has been reached on the death of mental logic and on the fact that reasoning is carried out by constructing mental models; nowadays the group of psychologists who doubt of the truth of the mental model theory is on the verge of extinction.

A good part of this sweeping success, vagueness notwithstanding, is due to the impressive list of problems the new hypothesis promised to solve. Let me list them. Mental models would:

(1) provide a general theory of deductive reasoning (Johnson-Laird, 1983a; Johnson-Laird & Bara, 1984, p. 3; Johnson-Laird & Byrne, 1991, p. x), and, in particular

(1a) explain propositional reasoning (Johnson-Laird & Byrne, 1991, 1993; Johnson-Laird, Byrne, & Schaeken, 1992);

(1b) explain relational reasoning (Johnson-Laird, 1983b; Johnson-Laird & Byrne, 1989, 1991, 1993);

(1c) explain the figural effect in reasoning (Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991, Ch. 6);

(1d) explain syllogistic reasoning (Johnson-Laird, 1983a, Ch. 5; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991), including individual differences

(Johnson-Laird, 1983a, pp. 117–121) and the belief bias effect (Johnson-Laird & Byrne, 1991, pp. 125–126; Oakill, Johnson-Laird, & Garnham, 1989);

(1e) explain reasoning with single and multiple quantifiers (Johnson-Laird, 1983a; Johnson-Laird & Byrne, 1991; Johnson-Laird, Byrne, & Tabossi, 1989);

(2) explain how logical reasoning is performed without logic (Byrne, 1991; Johnson-Laird; 1983a, Ch. 6, 1983b; Johnson-Laird & Byrne, 1991);

(3) account for a vast series of linguistic phenomena, such as anaphors, definite and indefinite descriptions, pronouns and plausibility effects in language processing (Johnson-Laird, 1983a; Garnham, 1987);

(4) offer a theory of the structure of discourse (Johnson-Laird, 1983a, pp. 370–371; Garnham, 1987);

(5) explain the difference between implicit and explicit inferences (Johnson-Laird, 1983a, Ch. 6);

(6) "solve the central paradox of how children learn to reason" (Johnson-Laird, 1983a, p. 45);

(7) explain content effects in reasoning (Byrne, 1991, p. 77);

(8) offer an explanation of meaning (Johnson-Laird, 1983a, p. 397; McGinn, 1989);

(9) "readily cope with the semantics of propositional attitudes" (Johnson-Laird, 1983a, p. 430) and solve the problems presented by them (Johnson-Laird, 1983a, pp. 430–436);

(10) provide a solution to the controversy on the problem of human rationality (Johnson-Laird & Byrne, 1993, p. 332);

(11) solve the problem of how words relate to the world (Johnson-Laird, 1983a, p. 402, 1989, pp. 473–474, 489; Garnham, 1987; McGinn, 1989);

(12) elucidate the nature of self-awareness and consciousness (Johnson-Laird, 1983a, pp. xi; Ch. 16).

Even the most benevolent reader, when confronted with a theory so rich in both philosophical consequences and empirical power, should have at least felt inclined to raise her critical eyebrows. Nevertheless, critical voices were confined to a "small chorus of dissenters", almost all tied to the "ardent advocates of rule theories" (Johnson-Laird & Byrne, 1991, p. ix). In fact, with some patience and time, I think it can be shown that all the philosophical advantages claimed for mental models are unsupported propaganda, and that most of the psychological evidence is much less firm than generally admitted. But showing it is quite a long task.

Another source of support for the mental model hypothesis came from a parallel series of arguments to the conclusion that the mental logic hypothesis is doomed to failure. In this paper, I will confine myself to a modest task. I will plainly go through the list of this second class of arguments and show that either they are not conclusive or, when reasonable, they point at problems which are troublesome for the mental model theory as well. The arguments follow in no particular order of importance.

2. Mental logic doesn't have the machinery to deal with meaning and cannot explain the role of content and context in understanding and reasoning

This is one of the major complaints against a mental logic. How could a formal theory enlighten us on such a clearly content-driven process as reasoning? In fact, as mental logic theorists recognize, one should distinguish two separate processes involved in problem solving. The first one is comprehension; the second one is reasoning proper. Accordingly, for mental logic theories a comprehension mechanism sensible to pragmatic information drives input analysis (Braine et al., 1984; Braine & O'Brien, 1991; O'Brien, 1993). Though the comprehension principles guiding it are only sketched, there is a hypothesis on their role in the time course of reasoning. After a first processing roughly delivering a syntactic analysis of a linguistic signal, the identification of its logical form and a first semantic analysis retrieving literal meaning, pragmatics and general knowledge aid to select a *particular logical form* for the input signal. Afterwards, representations possibly sharply different from the first semantic analysis are passed onto a processor blind to content and pragmatics. The general picture suggested, with some integration, looks like the diagram in Fig. 1.

So a theory of mental logic cannot, and does not intend to, explain the role of content in reasoning, though it may help to locate how and when content and pragmatics interact with reasoning proper. From this point of view, the complaint is correct.

However, models are no improvement; the thesis that "in contrast [to mental logic], the model theory has the machinery to deal with meaning" (Byrne, 1991, p. 77) is false. Models are supposed to be constructed either directly from perception, or indirectly from language. In the first case, no detailed account on how perception should generate models has been given.[1] For linguistic models, a sketch of the procedures for their constructions exists. According to it, models are constructed from propositional representations via a set of procedures sometimes

[1] Sometimes it looks as if perceptual models in Marr's sense are considered to be equivalent to mental models in Johnson-Laird's sense (see Johnson-Laird, 1983a; Johnson-Laird et al., 1992, p. 421), but there are structural differences between the two constructs which make it difficult to accept the identification. To mention the most apparent one, perceptual models don't contain negation, but mental models do. For this reason, for each perceptual model there is an infinite number of mental models corresponding to it. A perceptual model of John scratching his head is a mental model of John scratching his head, but also of John not scratching his leg, of John not running the New York Marathon, of Mary being late for a date, and so on.

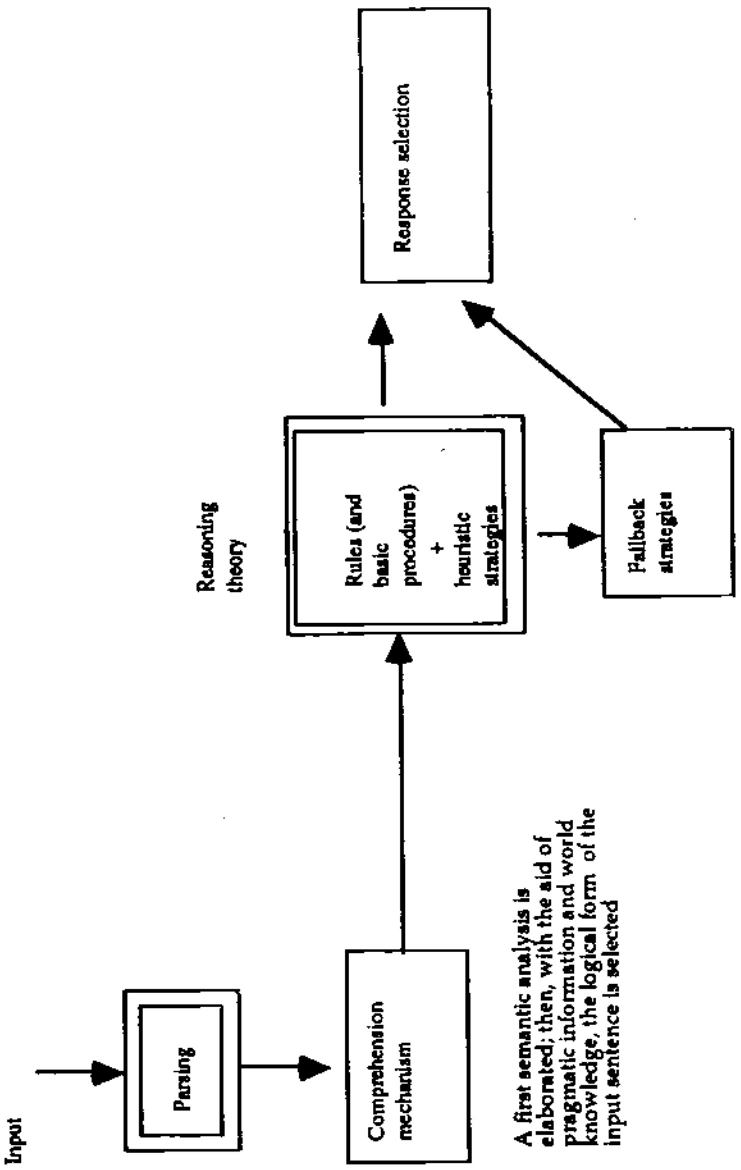
Input
Parsing
Comprehension mechanism
A first semantic analysis is elaborated; then, with the aid of pragmatic information and world knowledge, the logical form of the input sentence is selected
Reasoning theory
Rules (and basic procedures) + heuristic strategies
Fallback strategies
Response selection

Figure 1. *The place of pragmatics, comprehension mechanisms and reasoning proper according to mental logic. (The plausibility modular processes are in double squares.)*

called procedural semantics. For example (Johnson-Laird & Byrne, 1991, p. 170 ff.), when given as input a sentence like

The circle is on the right of the triangle

a parser will start working and after some crunching the following information will be placed on the top of its stack:

$$(\text{The-circle-is}\ldots) \rightarrow \text{Sentence}\ ((1,0,0)(\triangle)(\bigcirc))$$

The output of the parser is a couple containing *both* the grammatical description of the input ("sentence") *and* its semantical evaluation (in this case, an array containing numerical coordinates specifying the interpretation of the spatial relation, and the interpretations of the definite descriptions). Only at this point will procedural semantics take over and construct a model out of the propositional representation of the sentence; in this case, the model will be:

△ ○

that is, an image of a triangle to the left of the circle.

Now, notice the following points. First, the procedures that construct models do not operate properly on natural language sentences, but on the logical forms of propositional representations. Thus procedural semantics presupposes logical forms. By the same token, procedural semantics presupposes the literal meaning of words and sentences, which have to be received as its input. As Johnson-Laird himself writes, "The reader should bear in mind that the present theory uses a procedural semantics to relate language, not to the world, but to mental models" (1983a, p. 248). Procedural semantics is essentially translation from mental representations to mental representations, not a function from mental representations to the world. But, then, if procedural semantics is not about literal meaning and logical forms, neither are mental models.

Second, procedural semantics can work only if the output of the parser is not ambiguous: for example, scope relations must be already straightened out. The sentence

(1) Every man loves a woman

must be parsed to yield either

(2) For all men *x* there is some woman *y* such that (*x* love *y*)

or

(3) For some woman y all men x are such that (x loves y)

Only on the basis of *one* of them can procedural semantics yield a mental model. Thus the input to procedural semantics must be *clear*.

Third, the possibility to construct the appropriate models of a text strictly depends on the expression power of the logical forms on which procedural semantics operates. To continue with the previous example, there are interpretations of (1) which don't correspond to either (2) or (3), involving a generic reading of the indefinite description. While it is not clear that a mental model can express the difference between a generic woman and a specific woman, this much is clear: *if the logical form is not rich enough to articulate such a distinction*, then mental models cannot represent it either, since they come from expressible logical forms. Thus the input to procedural semantics must be *rich*.

Fourth, while the programs implementing the mental model theory described in Johnson-Laird (1983a) and Johnson-Laird et al. (1992) assume that the syntactic analysis of the input sentence plus word meaning is sufficient to determine its propositional content and logical forms, in a more natural setting the propositional content needed to construct the relevant mental models cannot be the first semantic analysis of the input, but *the propositional content and the logical forms of the message it conveys*. Now, by standard Gricean reasons the message conveyed in "Luca is a nice guy" in a text like

Q: Is Luca a good philosopher?
A: Well, let's say that Luca is a nice guy

has something to do with my ability as a philosopher, and not with how much people like me. So if we take seriously the proposal that mental models are the kind of structure we build when comprehending a text, it is this contextual message that they must retain. A similar point can be made for metaphors, analogies, and all the cases in which the hearer/reader gathers information from an utterance aided by her general world knowledge, her understanding of relevance in communication, and other pragmatic factors. Now, since procedural semantics is proposed as a set of procedures extracting models from propositional representations, clearly the propositional representations on which it has to act in order to build the right mental models *are not* the results of a first semantic analysis of input sentences retrieving their literal meaning, but *the analysis of their message in context*, which, therefore, has to be retrieved *before* models are constructed. Procedural semantics works once all the disambiguations due to context, scope phenomena and retrieval of the speaker's intentions have taken place.

To sum up, the input to procedural semantics presupposes both the literal meaning of the text and its logical form, and must be rich, clear, free from

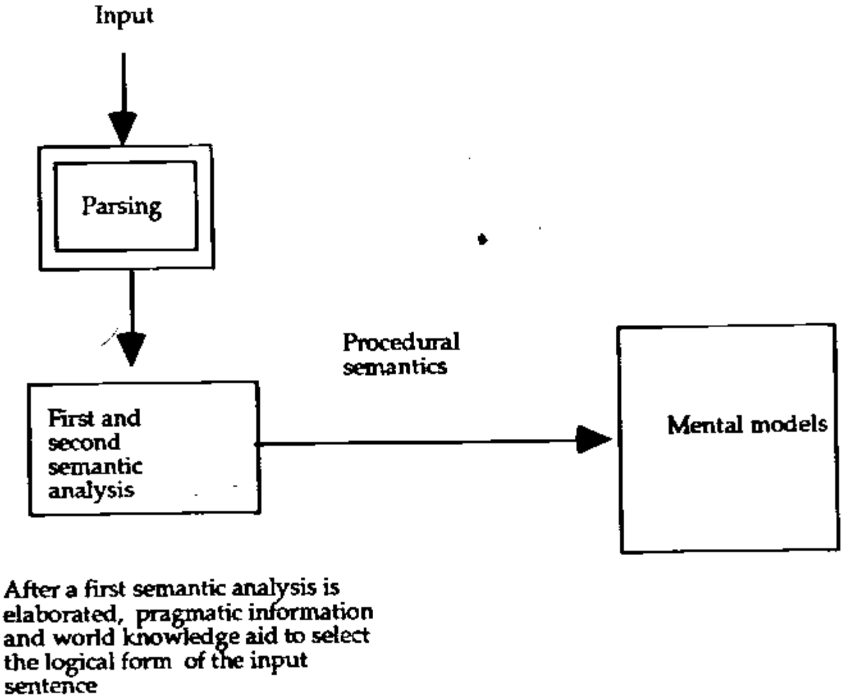
Input
Parsing
First and second semantic analysis
Procedural semantics
Mental models
After a first semantic analysis is elaborated, pragmatic information and world knowledge aid to select the logical form of the input sentence

Figure 2. *The place of pragmatic and comprehension mechanisms in the mental model hypothesis.*

structural ambiguities, and post-pragmatic. Thus when we begin to fill in the details, we come up with a sophisticated input analysis and we get the overall picture presented in Fig. 2, which for what concerns the role of pragmatics and meaning, *has no difference* from the mental logic picture – just as mental logic, procedural semantics and mental models *presuppose*, and do not explain, a theory of how pragmatics affects the selection of the correct message a set of utterances carries in the relevant situation.

It could be objected that I am presenting a misleading picture, based on the algorithms implementing a small fraction of the mental model theory rather than on the theory itself. Algorithms are only a part of the story; with time, the rest will come. So Johnson-Laird et al. (1992) write:

> The process of constructing models of the premises is, in theory, informed by any relevant general knowledge, but we have not implemented this assumption. (p. 425)

But such "assumption" amounts to the solution to the frame problem, and the suspicion that it won't be implemented is more than warranted (Fodor, 1983). In any case, if the problem were solvable, it would still be the case that the retrieval of the relevant message would occur in the pre-modelic construction processes selecting the right logical forms and propositional contents which are input to procedural semantics.

In fact, there is a litmus paper to test sensibility to content. The natural understanding of entailment seems to require a connection in content between antecedent and consequent. But the paradoxes of material implication allow false arbitrary antecedents to imply arbitrary consequents, regardless of their contents and even of their truth values. So if a theory of reasoning licenses them, it surely can't be advertised as the model to imitate for sensibility to content. Now, while in Braine and O'Brien's (1991) logical theory of implication the paradoxes are not available as theorems, mental models allow one to derive them as valid inferences (Johnson-Laird & Byrne, 1991). The reason is pretty clear. The mental model theory of connectives mainly consists of a variation on truth tables, and *truth tables are only sensible to truth values*, not to content connections or relevance.

Thus besides their name, models have no advantage over mental logic to explain the role of content in reasoning, in any of the relevant senses of "content". They cannot explain literal meaning, nor meaning in situation, nor how pragmatics and general knowledge affect interpretation, and they don't seem to have the adequate structure to do it.

3. There is no mental logic because people make fallacious inferences

People often reach conclusions which, if judged according to the canons of standard logic, are fallacious. And this should be a problem for a mental logic.

> The most glaring problem is that people make mistakes. They draw invalid conclusions, which should not occur if deduction is guided by a mental logic. (Johnson-Laird, 1983a, p. 25)

In less sophisticated versions, the argument notices that undergraduates make mistakes, and, worst of all, they show reiterate resistance to the teacher's efforts to correct them (Bechtel & Abrahansen, 1991, p. 168 ff.), or that they make more mistakes than what the average individual should innately know according to the logical competence mental logic attributes to people (Churchland, 1990, p. 283).

In fact, mistakes come in different classes. They may be due to cognitive components not engaging reasoning proper, such as the comprehension stage or strategies of response selection; to performance failures; or to faulty competence. Any errors due to pre-deductive, comprehension mechanisms, or post-deductive, response selection strategies, can be accommodated by the two hypotheses roughly in the same way: the existence of such errors doesn't count against mental logic any more than it counts against mental models. Performance mistakes are explained away by mental models by indicating how models are built and handled by mechanisms non-proprietary of reasoning – mostly, mechanisms of working memory storage and retrieval. A system based on mental logic can account for them in the same way.

Errors of competence – as it were, directly generated by how the reasoning box *is* – are a more delicate matter. The question is to decide with respect to which point of reference they are errors. Does failure to apply excluded middle count as an error? Does the absence of reasoning schemata corresponding to material implication count? Classical logic – or, for that matter, any alternative logics – cannot be a favored point of reference without further justifications.

One major task of a psychological theory of deductive reasoning is to characterize *what people take the right implications to be* starting from certain premises, under ideal conditions. What could count as a systematic error in this context? Previous assumptions on the nature of rationality must be exploited. It can be argued, for example, that it is rational to proceed from truths to truths. On this basis, invalid reasoning processes could count as mistakes. If it could be shown that under ideal conditions people respond erratically to identical problems, or embody a rule which brings about a systematic loss of truths, then it may be said that subjects make mistakes in point of competence *regardless* of the compliance of natural logical consequence to classical, or other, logics. But if this were the case, mental models would be in a worse position than mental logic. It is possible (though not desirable) to account for systematic errors within a mental logic framework by indicating which rules (if any) induce systematic violations of the selected normative model. As of today the algorithms proposed to implement logical reasoning by models are either psychologically useless or ill defined (Bonatti, in press; O'Brien, Braine & Yang, in press), so it is difficult to give a definite judgement on this issue, but the tentative set of rules proposed for model construction is meant to be truth preserving *in principle*. Thus it is puzzling to figure out how models might account for purported systematic violations: errors in point of competence would be an even deeper mystery for the mental model hypothesis.

4. There is no mental logic because higher-order quantifiers are not representable in first-order logic, and yet we reason with them

This argument has been considered "the final and decisive blow" to the doctrine of mental logic (Johnson-Laird, 1983a, p. 141). According to Barwise and Cooper (1981), expressions such as "More than half of" or "Most" are sets of sets, and therefore an adequate logic for natural language needs to extend beyond first order. The argument from this proposal to the rejection of mental logic runs as follows:

> [Higher-order calculus] is not complete. If there can be no formal logic that captures all the valid

> deductions, then *a fortiori* there can be no mental logic that does either. It is a remarkable fact that natural language contains terms with an implicit "logic" that is so powerful that it cannot be completely encompassed by formal rules of inference. It follows, *of course*, that any theory that assumes that the logical properties of expressions derive directly from a mental logic cannot give an adequate account of those that call for a higher-order predicate calculus. This failure is a *final and decisive blow* to the doctrine of mental logic. (Johnson-Laird, 1983a, pp. 140–141, italics mine)

The argument has often been repeated (see, for example, Johnson-Laird & Bara, 1984, p. 6; Johnson-Laird & Byrne, 1990, p. 81; Johnson-Laird & Byrne, 1991, p. 15); so it must be attached a certain importance. The question is to figure out why.

The nature of the representational device in which mental processes are carried out is an empirical question, and if patterns of inference are required that can be better formalized in second-order logic, so be it. So what can possibly be wrong in using higher-order logic? We are told, it is not complete. Such objection makes sense only if one presupposes that a mental calculus must be complete. But an argument is needed to ask for completeness as a *constraint* over a mental logic, and it is difficult to see what it would look like. We may impose constraints on a logical system by requiring that it possesses certain logical properties such as consistency, or completeness, because we can decide what we want from it. But finding out how people reason is an empirical enterprise. It would be a very interesting empirical discovery to find out that, say, a subject's system for propositional reasoning is complete, but it's not enough that we *want* it to be so. Even more basic logical properties cannot be granted *a priori*. It would be desirable that subjects reason consistently, as everybody hopes to discover that under ideal conditions they do, but, again, to presuppose that our reasoning system is consistent requires an argument. Barring such arguments, the "final and decisive blow against mental logic" blows up.

In fact, it may backfire. Johnson-Laird et al. blame the incompleteness of a higher-order mental logic system *as if* the mental model counterproposal *were* complete. But the only fragment for which a psychological implementation has been proposed – propositional reasoning – is not even valid. Models have no advantage over mental logic on the issue of completeness. Neither should they: such an advantage, in the absence of evidence that natural reasoning is complete, would be irrelevant.

5. There is no evolutionary explanation of the origin of mental logic

Another alleged argument against mental logic concerns its origin. A bland version of it simply claims that there is no evolutionary explanation of mental logic, and this is enough to reject the theory (Cosmides, 1989). A richer version runs as follows. To accept that there is a mental logic seems to lead to the admission that most of our reasoning abilities are innate. Nativism, in general, cannot be a problem: everybody has to live with it, and the only issue is whether you like it weaker or stronger. But there should be something specifically wrong with nativism about mental logic: there is no evolutionary explanation for its origin:

> By default, it seems that our logical apparatus must be inborn, though there is no account of how it could have become innately determined (Johnson-Laird, 1983a, p. 40).

> The moral that Fodor drew is an extreme version of nativism – no concept is invented; all concepts are innate. Alas, any argument that purports to explain the origins of all intellectual abilities by postulating that they are innate merely replaces one problem by another. No one knows how deductive competence could have evolved according to the principles of neo-Darwinism. (Johnson-Laird, 1983a, pp. 142–143)

> So intractable is the problem for formal rules that many theorists suppose that deductive ability is not learned at all. It is innate. Fodor (1980) has even argued that, in principle, logic could not be learned. The difficulty with this argument is not that it is wrong, although it may be, but that it is too strong. It is hard to construct a case against the learning of logic that is not also a case against its evolution. If it could not be acquired by trial-and-error and reinforcement, then how could it be acquired by neo-Darwinian mechanisms? (Johnson-Laird & Byrne, 1991, p. 204)

It is first worth noticing that the argument is meant to apply to cognition, and only to very restricted kinds of cognitive abilities. If you try to generalize it beyond this domain, it becomes flatly absurd. For the given premise is that Darwinian mechanisms are a sort of trial-and-error and reinforcement mechanisms applied to the species. Its generalization says: for any x, if x cannot be acquired by trial-and-error and reinforcement, then how could it be acquired by a neo-Darwinian mechanism? Now take a non-cognitive phenomenon and substitute it for x; breathing cannot be acquired by trial-and-error and reinforcement, so how did the species acquire the ability to breathe? That doesn't work. And neither does it work for most innate cognitive abilities. Try with colors, or perceptual primitives: the ability to recognize colors (or any perceptual primitive) cannot be acquired by trial-and-error and reinforcement, so how could the ability to recognize colors be acquired by neo-Darwinian mechanisms? This doesn't work either. So I assume that the argument is really targeted against mental logic.

Second, even restricting its field of application, notice that there are at least three different questions one may raise. What is the logical syntax of mental processes? What logical system underlies reasoning abilities? What concepts is the mind able to entertain, whether innately or by experience? The above argument does not keep them separate, yet they may have radically different answers. For example, an organism may be innately endowed with the syntax of first-order logic, but it may keep changing its logical system (for simplicity, the set of its axioms) by flip-flopping an axiom, and at the same time may need to learn any concept by experience. Such an organism would have an innate logical syntax, but no innate logic or innate concepts. Or else, an organism may be endowed with an innate logical syntax and an innate logic, but may need experience to acquire contentful concepts. The arguments for or against nativism are quite different in the three cases.

I will assume that the above argument is really targeted against nativism of a *system* of logic. Then, it can be reconstructed in the following way. If there is a mental logic, an account is due of how it is acquired. Since there is no theory of its acquisition, it must be assumed that the logical system – not just its syntax – is innate. But, alas, this claim is unsupported because there is no evolutionary story on how such a system gets fixated. Thus the doctrine of mental logic has to be rejected.

The short answer to such an argument (in its bland and its rich forms) is: too bad for evolutionary explanations. The long answer requires a reflection on the state of evolutionary explanations of cognitive mechanisms. The argument presupposes that *there must be* an evolutionary explanation of how deductive abilities are fixated. What would it look like? For the much clearer case of language, evolutionary explanations are uninformative. Whether a mutation endowing humans with linguistic abilities concerns the structures of the organism or in its functions; whether language has been a direct mutation, or a byproduct of another mutation; under what metric it turned out to be advantageous: these are unanswered questions. This is a general problem concerning the application of evolutionary concepts to cognition. The quest for a Darwinian explanation of cognitive evolution is founded at best on an analogy with biological evolution, and analogies may be misleading. Lewontin specifically makes this point for problem solving:

> . . . generalized problem solving and linguistic competence might seem obviously to give a selective advantage to their possessors. But there are several difficulties, First, . . . human cognition may have developed as the purely epiphenomenal consequence of the major increase in brain size, which, in turn, may have been selected for quite other reasons. . . . Second, even if it were true that selection operated directly on cognition, we have no way of measuring the actual reproductive advantages. . . . Fourth, the claim that greater rationality and linguistic ability lead to greater offspring production is largely a modern prejudice, culture – and history – bound. . . . The problem is that we do not know and never will. We should not confuse plausible stories with demonstrated truth. There is no end to plausible story telling. (Lewontin, 1990, pp. 244–245)

And there is no reason to ask for mental logic what does not exist and might not exist for other, better-known, cognitive domains.

But let us suppose that one should seriously worry for the lack of a Darwinian explanation of how innate logic has been selected. Again, here one should sense the kind of comparative advantage that the mental model hypothesis gains. The argument seems to presuppose that, as opposed to the case of mental logic, either (a) the ability of building mental models is not innate but learned, and thus Darwinian worries don't arise, or (b) if it is not learned, there is an evolutionist explanation of its origin.

Alternative (a) is empty. There is no learning theory for models and it is unlikely that any future such theory will bring about substantial economies in nativism, since most of the structures needed for problem solving are the same regardless of which theory turns out to be correct. Without (a), alternative (b) assumes the following form: an innate mechanism for building mental models gives an evolutionary advantage that an innate mental logic doesn't give. But evolutionary explanations are not so fine-grained to discriminate between our capacity to construct models *as opposed to derivations*. If there is any such explanation, it will work for both; if there isn't one for mental logic, there isn't one for mental models either.

6. Mental logic cannot explain reasoning because people follow extra-logical heuristics

Often heuristics of various sorts guide human responses even in deduction. But it is unclear how this counts against mental logic. Models need heuristics as much as logical rules do. For example, if a premise has different possible interpretations, an order is needed to constrain the sequence of constructed models (Galotti, 1989). Such an order too may depend on heuristics having nothing to do with models proper, such as reliance on the most frequent interpretation, or on previous experience, or on previously held beliefs.

But there may be something more to the argument. It may be argued that heuristics don't pose any special problem to model-based theories of reasoning, whereas they do for logic-based theories. Just like Dennett's queen moving out early, heuristics can be an epiphenomenon of the structure of models, whereas rule-based systems must express them explicitly. For example, a model for the sentences "*a* is to the right of *b*" and "*b* is to the right of *c*" allows us to derive "*a* is to the right of *c*" with no explicit rule to that effect (see Johnson-Laird, 1983a; Johnson-Laird & Byrne, 1991). In this case, transitivity is an emerging feature of the structure of the model. Analogously, it may be argued that also other apparent rule-following behaviors such as strategies are emerging features of models. However, often subjects reason by following heuristics that they can perfectly spell out and that are not accounted for by the structure of models (see, for example, Galotti, Baron & Sabini, 1986), and this squares very badly with a radical rule epiphenomenalism. At least in principle, models may help to solve the problem of implicitness: certain processes may be externally described by explicit rules which nevertheless are not explicitly represented in the mental life of an organism. Solution: the rules supervene to the structure of models. But the other side of the coin is the problem of explicitness: how could a system represent the information that *is* explicitly represented? This is no difficulty for mental logic, but how could a heuristic be explicitly represented within models? Tokens and possibly some of their logical relations are explicit in models, but not performatives. Models don't contain information specifying the order in which certain operations have to be executed, but only *the result* of such operations. So while a propositional-like system doesn't have the problem of explicitness, models may have it.

7. Mental logic cannot offer a theory of meaning for connectives

> In fact, the formal rules for propositional connectives are consistent with more than one possible semantics . . . Hence, although it is sometimes suggested that the meaning of a term derives from, implicitly reflects, or is nothing more than the roles of inference for it, this idea is unworkable . . . (Johnson-Laird et al., 1992, p. 420)

But truth tables (and thus models) don't have such a problem, since they "are merely a systematic way of spelling out a knowledge of the meanings of connectives" (Johnson-Laird et al., 1992, p. 420).

Johnson-Laird et al. refer to an argument presented in Prior (1960, 1964). But an aspect of it has been forgotten. Prior argued that rules of inference cannot analytically define the meaning of the connectives they govern. If there were nothing more to the meaning of a connective than the inferences associated to it, then the connective *tonk* could be defined, with the meaning specified by the following rules:

(1) From P, derive P tonk Q
(2) From P tonk Q, derive Q

and with *tonk* we could obtain the following derivation:

2 and 2 are 4
Therefore, 2 and 2 are 4 tonk 2 and 2 are 5
Therefore, 2 and 2 are 5.

Prior's argument is a challenge to a conceptual role semantics. If meaning is inferential role, how to avoid *tonk*? According to Prior, *tonk* shows that explicit definitions cannot give the meaning to a term on the ground of the analytical tie between the *definiens* and the *definiendum*, but can at most correspond to a previously possessed meaning: we *see* that certain rules of inferences are adequate for "and" because we know its meaning and judge the adequacy of the rules with respect to it. We can perfectly introduce a sign for *tonk* governed by the above rules and have a purely symbolic game running. But games with rules and transformations of symbols don't generate meaning: "to believe that anything of this sort can take us beyond the symbols to their meaning, is to believe in magic" (Prior, 1964, p. 191). The difference between "and" and *tonk* is that in the first case the rules correspond to the (previously held) sense of the word "and": they don't confer it its meaning, but are "indirect and informal ways" (Prior, 1964, p. 192) to clarify it. But in the second case there is no prior sense to appeal to. We *can* define a class of signs standing for conjunction, and a class of signs standing for contonktion, but the latter is an empty class. There are conjunction-forming signs, because there is a conjunction. There are no contonktion-forming signs, because there is no contonktion and the explicit introduction of a sign for it does not give life to a new connective.

So Prior's argument goes. One way to read it is that rules can't give a symbol its meaning, but something else can: namely, truth tables in a metalanguage. This seems to be the interpretation adopted by Johnson-Laird et al. (1992) when they claim that mental logic cannot explain the meaning of connectives, but truth tables can.

In fact, Prior (1964) remarked that explicitly defining connectives in terms of truth tables did not change the point of his criticism. In his view, there was "no difference in principle between [rules of inferences and truth tables]" (Prior, 1964, p. 192). Instead of using rules, he argued, we can define a conjunction-forming sign by using the familiar truth table, but this will not give conjunction its meaning; any formula of arbitrary length with the same truth table will turn out to be a conjunction-forming sign; so will formulas involving non-logical conceptions such as "P ett Q", which is the abbreviation for "Either P and Q, or Oxford is the capital of Scotland" (Prior, 1964, p. 194).

The point of this further facet of the argument is that truth tables identify a much broader class of signs than conjunction, and moreover, signs that are understood on the basis of the understanding of conjunction (see Usberti, 1991). We might try to eliminate all the unwanted signs which would be defined by the truth table for conjunction by saying that the table defines the meaning of the *shortest* possible sign for conjunction. We would probably be happy with this solution. But, Prior noticed, we would accept it because we understand that such a characterization captures the meaning of the conjunction, and not of the other signs.

Thus, truth tables are in no better position than rules to generate meanings. If they apparently don't suffer from tonkitis, they suffer from another equally worrisome disease. And if we wanted to resort again to formal games, then tonkitis would reappear, since a (symbolic) truth table game defining a contonktion-forming sign is easy to find: *tonk* "is a sign such that when placed between two signs for the propositions P and Q, it forms a sign that is true if P is true and false if Q is false (and therefore, of course, both true and false if P is true and Q is false)" (Prior, 1964, p. 193).

We can now leave Prior and touch on the real problem. If we grant that explicit rules, or truth tables, don't define the meaning of the logical symbols, but are accepted on the basis of their correspondence to some pre-existent meaning we attach to connectives and quantifiers, we still have to explain what the source of our intuitions about the meaning of connectives and quantifiers is, because if thinking that a game of symbols can take us beyond the symbols to their meanings is magic, as Prior said, it is equally magic to think that the meaning of logical symbols comes from nowhere.

For Johnson-Laird, Byrne, and Schaeken, truth tables are merely "a systematic way of spelling out a knowledge of the meanings of connectives". But in general this is false. There are 16 binary truth tables: only *some* of them do, or seem to, spell out the meaning of binary connectives; *others* clearly don't. Why is it so? Why do we feel that the truth table for the conjunction reflects the meaning of the conjunction, whereas the classical truth table for the implication doesn't reflect the meaning of natural implication, and the anomalous truth table for *tonk* can't reflect the meaning of a new connective?

Nothing seems to block the following possibility. When I see somebody who reminds me of my brother, one of the possibilities is that it *is* my brother. So when I see a set of rules for the conjunction and I think that it adequately expresses what I mean by a conjunction, one of the possibilities is that I find that resemblance because the rules are the *exact* expression of the patterns of inferences of a logical connective in the mind. In this case, *there is nothing more to the meaning of the term than the rules themselves*. At the same time, when I see the truth table of material implication I realize that it does *not* spell out the meaning of natural implication because the rules governing natural implications are not reflected in it, and when I see the rules of inference – or the truth table – for *tonk*, I have no intuition about their adequacy because there is no logical connective for *tonk* in the mind, from which the explicit rules are a clone copy. Contonktion cured.

Intuitions, however, are not good guides. It is not enough to say that conjunctions have a meaning because they seem to correspond to rules in the mind but contonktions don't because they don't titillate our intuitions. There are lots of logical operators that may not have any straightforward correspondence with natural language, and yet are computed in retrieving the truth conditions of natural language sentences – consider, for example, focus, or quantifiers over events. If a semanticist presented us with a set of rules for them, we would not probably have the same immediate intuition we feel for conjunction. This is where a theory of mental logic comes in. A developed theory of mental logic offers empirical reasons to show that conjunctions are in the mind, while contonktions are not. If such a theory can be worked out (and a tiny part of it already exists), then mental logic can be the basis of a theory of meaning for natural connectives. For the moment, we are very far from having such a complete theory. The present point is simply that no argument exists to hamper its development.

8. There is no mental logic because valid inferences can be suppressed

This recent argument is based on the so called "suppression of valid inferences" paradigm. By modifying a paradigm used by Rumain, Connell, and Braine (1983), Byrne (1989) set up an experiment in which to premises such as

If she meets her friend she will go to a play
She meets her friend

an extra premise was added, transforming the argument in

If she meets her friend she will go to a play
If she has enough money she will go to a play
She meets her friend

and she showed that in this case the percentage of subjects applying *modus ponens* drops from 96% to 38%.

Mental model theorists attributed a considerable importance to this result. It shows, they claimed, that also valid deductions as strong as *modus ponens* can be blocked:

> Models can be interrelated by a common referent or by general knowledge. Byrne (1989) demonstrated that these relations in turn can block *modus ponens*. . . . The suppression of the deduction shows that people do not have a secure intuition that *modus ponens* applies equally to any content. Yet, this intuition is a criterion for the existence of formal rules in the mind. (Johnson-Laird et al., 1992, p. 326)

and as a consequence that

> by their own argument, rule theorists ought to claim that there cannot be inference rules for (valid deduction). (Johnson-Laird & Byrne, 1991, p. 83)

But no argument is offered to ensure that *modus ponens* is really violated, or to justify the claim that this result supports the mental models hypothesis. If we assume that deductive rules apply *not* to the surface form of a text, but to its integrated representation, then subjects may be led by pragmatic reasons to construe the two premises

If she meets her friend she will go to a play
If she has enough money she will go to a play

as a single

If (she meets her friend *and* she has enough money) she will go to a play and therefore when provided only with the premise "She meets her friend", they don't know about the truth of the conjunctive antecedent and correctly refuse to use *modus ponens*.

In other cases, also studied by Byrne, when subjects are given arguments such as

If she meets her friend she will go to a play
If she meets her mother she will go to a play
She meets her friend

they do conclude that she will go to a play. This may be because subjects compose the premises "If A then B" and "If C then B" as a single "If (A *or* C), then B", and knowing one of the disjuncts of the composed antecedent suffices to correctly apply *modus ponens*. Thus, under this interpretation, there is no suppression of valid inferences: simply, people tend to construct a unified representation of a text which may itself be governed by formal rules of composition.

It may be replied that my response to the suppression argument puts the weight of the explanation on pre-logical comprehension processes, rather than on deduction proper, and that mental logic theorists have no account of such processes. This wouldn't be necessary for models, because they "have the machinery to deal with meaning". But I have shown that such a claim is false. Models too rely on pragmatic comprehension mechanisms, and don't explain them. If model theorists want to explain why people draw the inference in one case and not in the other, they have to say that in one case a model licensing the inference is constructed, and in the other a model not licensing the inference is constructed. To account for why it is so, they offer no explanation.

9. Conclusions

"Yes, but mental logic has had its shot. It has been around for centuries and nothing good came out of it. It's time to change." Often the contrast between the long history of mental logic and its scarce psychological productivity is taken as a proof of its sterility. In fact, this impression derives from a mistake of historical perspective. The idea is very ancient, but the conceptual tools needed to transform it into the basis for testable empirical hypotheses are very recent. For centuries, logic too remained substantially unchanged, to the point that Kant considered it a completed discipline (1965, pp. 17–18). So there was no reason to change the conventional wisdom on the relations between logic and psychology: the former was stable because considered complete and the latter was stable because non-existent. When, with Frege, Russell and the neopositivists, logic as we mean it started being developed, the routes of logic and psychology separated.

Well beyond the 1930s, among the large majority of philosophically minded logicians, showing interest in psychological processes became a sort of behavior that well-mannered people should avoid. No substantial argument against the psychological feasibility of mental logic motivated this change of view. Rather, its roots have to be looked for in the general spirit of rebellion against German and English idealism from which twentieth-century analytic philosophy stemmed. Nevertheless, for independent reasons, the same conclusion became popular among experimental psychologists and was generally held until the early 1960s, both by behaviorists and by the new-look psychologists. There was, indeed, the Piagetian exception, but it does not count: Piaget's flirting with mental logic was never clear enough to become a serious empirical program (Braine & Rumain, 1983), and recent Piagetian-oriented investigations on mental logic (see Overton, 1990) have not helped towards a clarification.

It was again an impulse coming from logicians – not from psychologists – that put logic back in the psychological ballpark. Hilbert first directly expressed a connection between symbols and thought which could serve as a psychological underpinning for mental logic. For him, the fundamental idea of proof theory was "none other than to describe the activity of our understanding, to make a protocol of the rules according to which our thinking actually proceeds. Thinking, it so happens, parallels speaking and writing: we form statements and place them one behind another" (1927, p. 475). Yet Hilbert's intuition was not enough. Formal systems, as conceived by the axiomatic school, were the least possible attractive tool to investigate the psychology of reasoning. What was still missing to render logic ready for psychological investigation was on the one side a more intuitive presentation of formal systems, and on the other side a model of how a physical structure can use a formal system to carry out derivations. The first was provided by Gentzen, and the second by Turing.

However, once again, the distance between Gentzen's and Turing's ideas and a real psychological program should not be underestimated. Gentzen did introduce the systems of natural deduction with the aim to "set up a formal system which comes as close as possible to actual reasoning" (Gentzen, 1969, p. 68), but his reference to "actual reasoning" was merely intuitive. And Turing did offer the abstract model of how a physical mechanism could perform operations once considered mental along the lines suggested by Hilbert, but Turing's real breakthrough consisted of the realization that *a computer can be a mind*, namely, that certain kinds of properties once attributable only to humans can also be appropriately predicated of other physical configurations. Such insight, however, leaves the mechanisms and procedures by which the mind itself operates underspecified. It says that mental processes can be simulated, but it leaves it undetermined whether the *simulandum* and the *simulans* share the same *psychology*. The further step necessary to the formulation of a psychological notion of mental logic came when functionalism advanced the explicit thesis that the psychological vocabulary *is* computational vocabulary, and that the natural kinds described by psychology are not organisms, but computational devices. The change leading to this second step was gradual, and required a lot of philosophical work to be digested.

We are now beyond the 1960s, and not in Aristotle's age. Only then had logic and philosophy come to the right point of development to take the mental logic hypothesis seriously. And another decade or more had to go before experimental techniques were sufficiently developed to begin asking nature the right questions in the right way. The works by Braine, Rips and their collaborators are the first attempts at elaborating mental logic in a regimented psychological setting.

Thus the *psychological history* of mental logic is very recent. It is, in fact, roughly contemporary with the psychological history of the mental model hypothesis. This shouldn't come as a surprise: both needed largely the same conceptual tools to be conceived. Mental models are not the inevitable revolution after millennia of mental logic domination.

So, contrary to widespread assumptions, there are no good arguments against mental logic, be it point of principle, or in point of history. If a case against it and in favor of mental models can be made, it cannot rest on principled reasons, but on the formal and empirical development of the two theories. Indeed, extending the mental logic hypothesis beyond propositional reasoning engenders formidable problems connected with the choice of an appropriate language to express the logical forms of sentences on which rules apply, the choice of psychologically plausible rules to test, and the choice of appropriate means to test them. Approaching these problems requires the close collaboration of psychologists, natural language semanticists and syntacticians. But these are problems, however hard, and not mysteries. Most psychologists have abandoned the program and married the mental models alternative, both for its supposed superiority in handling empirical data and for the overwhelmingly convincing arguments against mental logic. In fact, the case for mental models has been overstated under both counts. Given how little we know about the mind and reasoning, conclusions on research programs that only began to be adequately developed a few years ago are premature. Psychologists should keep playing the mental logic game.

References

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

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind*. Oxford: Basil Blackwell.

Bonatti, L. (in press). Propositional reasoning by model? *Psychological Review*.

Boolos, G. (1984). On 'Syllogistic inference'. *Cognition*, *17* 181–182.

Braine, M.D. (1979). If then and strict implication; A response to Grandy's note. *Psychological Review*, 86, 158–160.

Braine, M.D., & O'Brien, D.P. (1991). A theory of *if*: Lexical entry, reasoning program, and pragmatic principles. *Psychological Review*, *98*, 182–203.

Braine, M.D., Reiser, B.J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. *The psychology of learning and motivation* (Vol. 18, pp. 313–371). San Diego, CA: Academic Press.

Braine, M.D., & Rumain. B. (1983). Logical Reasoning. In J. Flavell (Ed.), *Caarmichael's handbook of child psychology* (Vol. III, pp. 263–340). New York: Wiley.

Byrne, R. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61–83.

Byrne, R. (1991). Can valid inferences be suppressed? *Cognition*, *39*, 71–78.

Churchland, P.M. (1990). On the nature of explanation: a PDP approach, reprinted in S. Forrest (1991), *Emergent computation* (pp. 281–292). Cambridge, MA: MIT Press.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Ehrlich, K., & Johnson-Laird, P.N. (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior*, *21*, 296–306.

Fodor, J.A. (1980). On the impossibility of acquiring 'more powerful' structures. In M. Piatteli-Palmarini (Ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky* (pp. 142–163). Cambridge MA: Harvard University Press.

Fodor, J.A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Galotti, K.M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin*, *105*, 331–351.

Galotti, K.M., Baron, J., & Sabini, J.P. (1986). Individual differences in syllogistic reasoning: deduction, rules or mental models? *Journal of Experimental Psychology: General*, *115*, 16–25.

Garnham, A. (1987). *Mental models as representations of discourse and text*. Chichester: Ellis Horwood Ltd.

Gentzen, G. (1969). Investigations into logical deduction. In M.E. Szabo (Ed.), *The collected papers of Gehrard Gentzen* (pp. 68–128). Amsterdam: North Holland.

Hilbert, D. (1927). The foundations of mathematics. In J. van Heijenoort (Ed.), *From Frege to Gödel* (pp. 464–469). Cambridge, MA: Harvard University Press 1967.

Hodges, J. (1993). The logical content of theories of deduction. *Behavioral and Brain Sciences*, *16* 353–354.

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

Johnson-Laird, P.N. (1983b). Thinking as a skill. In J.B. Evans (Ed.), *Thinking and reasoning: psychological approaches* (pp. 44–75). London: Routledge & Keegan.

Johnson-Laird, P.N. (1989). Mental models. In M. Posner (Ed.), *Foundations of cognitive science* (pp. 469–499). Cambridge, MA: MIT Press.

Johnson-Laird, P.N., & Bara, B. (1984). Syllogistic inference. *Cognition*, *16*, 1–61.

Johnson-Laird, P.N., & Byrne, R. (1989). Spatial reasoning. *Journal of Memory and Language*, *28*, 565–575.

Johnson-Laird, P.N., & Byrne, R. (1990). Meta-logical problems: Knights, knaves and Rips, *Cognition*, *36*, 69–84.

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

Johnson-Laird, P.N., & Byrne, R.M. (1993). Précis of *Deduction*. *Behavioral and Brain Sciences*, *16*, 323–380.

Johnson-Laird, P.N., Byrne, R., & Schaeken, W. (1992) Propositional reasoning by model. *Psychological Review*, *99*, 418–439.

Johnson-Laird, P.N., Byrne, R.M., & Tabossi, P. (1989). Reasoning by model: the case of multiple quantification. *Psychological Review*, *96*, 658–673.

Kant, I. (1965). *Critique of pure reason*. New York: St. Martin Press.

Lea, R.B., O'Brien, D.P., Fisch, S., Noveck, I., & Braine, M. (1990). Predicting propositional logic inferences in text comprehension. *Journal of Memory and Language*, *29*, 361–387.

Lewontin, R.C. (1990). The evolution of cognition. In D. Osherson and E. Smith (Eds.), *Thinking: an invitation to cognitive science* (vol. 3, pp. 229–246). Cambridge, MA: MIT Press.

McGinn, C. (1989). *Mental content*. Oxford: Basil Blackwell.

Oakill, J., Johnson-Laird, P.N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, *31*, 117–140.

O'Brien, D. (1993). Mental logic and irrationality: we can put a man on the moon, so why can't we solve those logical reasoning problems? In K.I. Manktelow & D.E. Over (Eds.), *Rationality* (pp. 110–135). London: Routledge.

O'Brien, D., Braine, M.D., & Yang, Y. (in press). Proportional reasoning by mental models? Simple to refute in principle and in practice. *Psychological Review*.

Overton, W. (Ed.) (1990). *Reasoning, necessity and logic: Developmental perspectives*. Hillsdale, NJ. Erlbaum.

Prior, A.N. (1960). The runabout inference ticket. *Analysis*, *21*, 38–39.

Prior, A.N. (1964). Conjunction and contonktion revisited. *Analysis*, *24*, 191–195.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*, 38–71.

Rips, L.J. (1986). Mental muddles. In M. Brand & R. Harnish (Eds.), *The representation of knowledge and belief* (pp. 258–286). Tucson: University of Arizona Press.

Rumain, B., Connell, J., & Braine, M.D. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: *It* is not the biconditional. *Developmental Psychology*, *19*, 471–481.

Usberti, G. (1991). Prior's disease, *Teoria*, *2*, 131–138.