Mental Models: Reasoning without Rules

James H. Fetzer
Department of Philosophy, University of Minnesota, Duluth, MN 55812, U.S.A.
E-mail: jfetzer@d.umn.edu

According to P. N. Johnson-Laird and Ruth M. J. Byrne (1998), their theory of mental models, when properly understood, affords an appropriate framework for understanding the psychology of reasoning: "Mental models are a psychological theory, not a rival logic". Thus, in their work, *Deduction* (1991), they report that, given problems involving standard forms of argument, sample subjects were less than completely successful in deriving valid conclusions, where, for example, a success rate of 91% was achieved involving *modus ponens*, a rate of 64% using *modus tollens*, 48% for *affirmative disjunction*, and so forth (p. 55). Of course, these success rates reflect corresponding failure rates that, apart from *modus ponens*, fall below 70%, and therefore would not qualify for average (or even passing) grades in most conventional courses on logic.

According to Johnson-Laird and Byrne, human beings actually reason by means of mental models, in which they construct (in their minds) models of premises and then consider (in their minds) whether specific conclusions do or do not hold in those models. This approach, they claim, has many virtues:

> [I]t provides a unitary explanation of inferences yielding necessary, probable, and possible conclusions. A necessary conclusion holds in all the models of the premises, a probable conclusion holds in most of them, and a possible conclusion holds in at least one of them. (Johnson-Laird and Byrne 1998.)

They emphasize that their focus is upon the psychology of reasoning rather than the logic of deduction, which suggests that their theory is intended to describe how people do reason rather than how they should reason, which is compatible with arbitrary deviations from appropriate standards. When sample subjects achieve success rates of 64% with *modus tollens* arguments, for example, that means that they fail to satisfy the appropriate standards of validity 36% of the time, but they may still regard the "conclusions" they draw as "holding in all the models of the premises" they mentally construct!

[In response to Fetzer 1998, Johnson-Laird and Byrne (1998) observe regarding *modus tollens* that "none of the current psychological theories includes the rule of *modus tollens*" because these "inferences are difficult, and these theories account for the phenomenon by dropping its rule from the mind." Previous studies of the psychology of reasoning, however, have not found *modus tollens* so problematic. Peter Wason and Johnson-Laird (1972), for example, reported a remarkable capacity for valid inference, even among six- and eight-year-olds. Thus, a child is asked, "If that boy is John's brother, then he is ten years old. He is not ten years old. Is he John's brother?". The responses of their subjects were usually correct (Wason

and Johnson-Laird 1972, p. 41). This problem, of course, involves a *modus tollens* inference. If problems that six- and eight-year-olds can solve are "too difficult" for current psychological theories, then why should they be taken seriously?]

There thus seems to be a crucial ambiguity at the heart of the theory of mental models, namely: the difference between conclusions that *hold in all (most, some) models of the premises* and conclusions that *subjects believe* hold in all (most, some) models of the premises! Subjects who believe that specific conclusions hold in all (most, some) models of the premises, no doubt believe that those conclusions are necessary (probable, possible) in suitable counterpart senses. But that does not mean that those specific conclusions actually do hold in all (most, some) models of the premises, when content relations between premises and conclusions are properly appraised! That would be expected only if every member of the subject sample had incorporated appropriate principles of necessary (probable, possible) reasoning among their "habits of mind", which Johnson-Laird and Byrne's studies establish is not the case.

It should therefore be evident that a purely descriptive theory of mental models could support a theory of deductive reasoning only if it impurely presupposed suitable standards of validity. It would not be possible to measure relative frequencies of success in deriving conclusions from premises *in the absence of suitable standards that specify which conclusions are valid consequences of those premises*. There therefore seems to be no sense in which the study of mental models (as applied reasoning) could possibly displace the study of deductive logic (as pure reasoning), insofar as applied deductive reasoning can only be appraised by means of the extent to which it conforms – implicitly or explicitly – to the principles of deductive validity.

The pretensions of the theory of mental models to account for necessary, probable, and possible inference, moreover, appear fraught with problems. The kinds of examples they advance are actually ones that trade in logical necessities, probabilities, and possibilities. Johnson-Laird and Byrne seek to rebut my remark (Fetzer 1998) that "[t]entative and probabilistic reasoning exemplify inconclusive inductive reasoning" by advancing a specific example of "tentative reasoning", where,

> **Given the premise:**
>
> > **The flaw is in the dynamo or the turbine, or both.**
>
> **the following tentative conclusion is valid:**
>
> > **Possibly, the flaw is in the dynamo.**
>
> (Johnson-Laird and Byrne 1998.)

But as a counterexample this really will not do. According to their own definitions, probable and possible conclusions are ones that hold in most or in some models of the premises. But the conclusion, "Possibly, the flaw is in the dynamo", is one that holds in *every* appropriate model of the premises. It would be a *tentative*

conclusion to infer, "The flaw is in the dynamo", of course, but equally obviously such a conclusion could not possibly be *valid*.

Moreover, as this example itself reflects, Johnson-Laird and Byrne often import normative considerations that have no standing within the theory of mental models as a psychological account of human practice. That the conclusion "Possibly, the flaw is in the dynamo" may be valid in the sense that it must hold in every model in which its premises hold, when normatively considered, establishes no basis for inferring that every, most, or even any of us would actually draw that specific inference in our ordinary reasoning. Not only is there no reason to expect the right answer from different human beings within this theory, there is also no reason to expect the same answer as a function of life history and mental agility.

Thus, in response to my assertion that mental-model theory cannot provide an effective decision procedure for sentential or any other mode of reasoning, Johnson-Laird and Byrne (1998) initially seem to agree, "Because people make systematic errors". But the errors people make need not be "systematic"; they could be *highly idiosyncratic* as a result of personal experience, background knowledge, and power of imagination! If mental-model theory really is the study of how humans reason, and if their reasoning is frequently flawed (as they report), how could it possibly provide an effective decision procedure?

Moreover, that an AI implementation of some algorithms for reasoning "generates all possible models for each premise" and "evaluates given conclusions as valid or invalid" (Johnson-Laird and Byrne 1998) should yield a decision procedure for sentential reasoning has nothing to do with human reasoning or with the theory of mental models – unless Johnson-Laird and Byrne are suggesting that people are computers. And, if they are suggesting this, then it would be interesting to know how authors who tout the supremacy of semantics over syntax envision the operation of an AI mechanism that functions on the basis of Turing-machine principles as a syntax-processing system in defense of mental-model theory!

A variant theory of human reasoning that might be worth exploring would be to study the relative frequency m/n with which specific premises are associated with specific conclusions. Consider, for example, the following premise:

(P1) This coin is being tossed.

in relation to various possible conclusions, such as, say, the following three:

(C1) This coin comes up heads.

(C2) This coin comes up tails.

(C3) A nuclear explosion occurs.

Presumably, about half the time (C1) will occur, and about half the time (C2). Yet there are possible models of the premises in which (C3) might hold, namely, where the tossing of a coin would be used to trigger a nuclear explosion!

The strength of the inference to a specific conclusion, given specific premises, therefore, might be measured by the relative frequency with which that conclusion is taken to be true in relation to the truth of those premises. Thus,

(X0) p
- - - $[m/n]$
q

would represent the pattern of association between specific premises p and specific conclusions q.[1] Even if most sample subjects were to assign values of 1/2 to (C1) in relation to (P1), say,

(X1) This coin is tossed.
- - - - - - - - - - - - - - - - - - - [1/2]
This coin comes up heads.

and 1/2 to (C2) in relation to (P1),

(X2) This coin is tossed.
- - - - - - - - - - - - - - - - - - - [1/2]
This coin comes up tails.

others might consider models involving more complex scenarios, where, say,

(X3) This coin is tossed.
- - - - - - - - - - - - - - - - - - - [1/100]
A nuclear explosion occurs.

Indeed, there is no reason why the same subject could not endorse all three! Insofar as the psychology of reasoning incorporates no normative dimension, it could turn out to be the case that various people assign various strengths of support to various conclusions on the basis of their private mental models without concern for whether they conform to the mathematics of probability.

Johnson-Laird and Byrne, I suspect, would object that patterns of reasoning such as these are not instances of deductive reasoning, since their conclusions could be false even when their premises are true! That, of course, is correct, where these patterns of reasoning are examples of ordinary thinking for which the conclusions are tentative and inductive. An approach of this kind, it should be observed, also has the capacity to apply to deductive and inductive reasoning when general principles, such as that m/n As are Bs, are introduced. Consider, for example, the following argument concerning a medical procedure:

(X4) 70% of patients survive this procedure.
Mary Smith will undergo this procedure.
- [0.70]
Mary Smith will survive this procedure.

where the specified premises presumably confer a degree of evidential support of 0.70 upon the conclusion, indicating that in 70% of the mental models in which the premises are true, the conclusion is also true. Now it should be obvious that a counterpart argument could be constructed about that patient's non-survival:

(X5) 30% of patients do not survive this procedure.
Mary Smith will undergo this procedure.
- [0.30]
Mary Smith will not survive this procedure.

where the same person might be capable of entertaining (X4) as a mental model yet be incapable of entertaining (X5) as a mental model, even though (X5) must be true if (X4) is true! Consider, for example, Mary Smith, who might be incapable of contemplating any mental model in which she does not survive; hence,

(X6) 30% of patients do not survive this procedure.
Mary Smith will undergo this procedure.
- [0.99]
Mary Smith will survive this procedure.

where Mary Smith, as a psychological phenomenon, considers herself to be an exception, such that, no matter what may have been true of patients undergoing this procedure in the past, there is virtually no chance she will not survive! Thus, as I have already explained, there is no reason at all why different people should not display arbitrary differences in their reasoning with mental models.

Suppose, for example, that the premises declared that 100% of the patients who undergo this procedure do not survive; indeed, even assume it is a natural law that no one who undergoes this procedure could possibly survive. It might still be the case that Mary Smith engages in the following pattern of reasoning:

(X7) 100% of patients do not survive this procedure.
Mary Smith will undergo this procedure.
- [1.00]
Mary Smith will survive this procedure.

There are simply no constraints within the theory of mental models that force anyone's reasoning to conform to the normative constraints of deductive validity, inductive propriety, or logical coherence. Requirements of this kind might be conditions of rational reasoning, but there is nothing inherent in the theory of mental models that suggests or implies that human reasoning has to be rational!

This alternative conception would have several advantages relative to the the theory of Johnson-Laird and Byrne. Assuming that the basic pattern for reasoning involving generalizations has the form exemplified by (X4), namely:

(X8) m/n As are Bs.
This is an A.

- - - - - - - - - - - - $[m/n]$
This is a B.

it would permit the differentiation between forms of reasoning that are deductive and conclusive (where the conclusion cannot be false, if the premises are true) when $m = n$ (and therefore the value of $m/n = 1$) and inductive and inconclusive (where the conclusion can still be false, even if the premises are true) when $m \neq n$. It also provides a normative standard against which degrees of irrationality of belief may be measured by magnitudes of departure from m/n. In the case of (X6), for example, that degree of irrationality would equal 29%.

Other major flaws in mental-model theory are not difficult to discern. Thus, Johnson-Laird and Byrne (1998) assert that "A fundamental principle of the [mental-]model theory is that reasoners normally represent only what is true. In this way, they minimize the load on their short-term memory." But if this were actually the case, it would severely constrain the range of human reasoning and largely defeat its very rationale. This contention, after all, confounds the difference between validity and soundness. The principles of reasoning are intended to apply to arguments by virtue of their form with respect to what must be true (necessarily, probably, possibly) *given fixed premises*. They are hypothetical and frequently applied to premises not known to be true that are often false.

Johnson-Laird and Byrne systematically conflate the difference between *what is true* and *what might be true* by using the language of "true possibilities". Thus, in their reply (Johnson-Laird and Byrne 1998), they offer the representation of the specific sentence,

(P2) There isn't a king in the hand or else there is an ace in the hand

which they claim to represent by two alternative models on separate lines:

(M1) −king

(M2) ace

where '−' denotes negation. According to Johnson-Laird and Byrne (1998), "Each model corresponds to a true possibility, and each model represents only those literal propositions in the disjunction that are true within the possibility". So, according to them, it follows from (P2) that (M1) is possible and that (M2) is possible but not both, assuming an exclusive disjunction.

Johnson-Laird and Byrne provide this instance of reasoning as an example of reasoning to a possible conclusion. But possible conclusions may or may not be true. Within conventional sentential calculus, of course, the same disjunctive sentence (using obvious sentence letters 'K' and 'A') would be represented by

(SL) $-K \vee A$.

Here the wedge '∨' represents disjunction, but spacing on separate lines serves the same purpose within the theory of mental models. It is not even correct to claim that one is syntactical and the other is semantical: As representations of (P2), the mental model pair-of-models (M1) and (M2) would appear to have no obvious benefits with respect to their significance in comparison to (SL). Such differences as Johnson-Laird and Byrne claim for themselves appear illusory.

Not the least disturbing methodological aspect of their study, moreover, is their apparent lack of attention to argumentative intentions. In ordinary discourse, these intentions are signaled by the use of premise indicators ('given', 'since', 'assuming', and so on) and conclusion indicators, where inductive conclusion indicators (such as 'probably', 'possibly', 'suggests') differ from deductive conclusion indicators (including 'therefore', 'consequently', and 'proves'). Since every proper inductive argument is deductively invalid, only arguments that are intended to be deductive are correctly appraised by deductive standards. Without knowing the intent of those who offer them, we cannot know the standards by which they should be appraised and cannot non-arbitrarily ascertain whether invalidity makes them defective.

Philosophers commonly deal with phrases or expressions that are ambiguous and vague in attempts to clarify and illuminate their meaning by offering proposals or recommendations as to how they should best be understood. In the case of the theory of mental models, the seductive sound of the language employed creates the impression that the position thereby represented must be both significant and true. We have discovered, however, that this approach possess no normative content whatever for how human beings should reason. As a theory for how human beings actually do reason, it qualifies as no more than a framework, because it cannot even claim that human beings reason in conformity with a single pattern. Arbitrary deviations from normative standards are not only permitted but, given their own evidence, are to be expected from person to person.

Although the authors promise "a unitary explanation of inferences yielding necessary, probable, and possible conclusions" (Johnson-Laird and Byrne 1998), what they offer to support such claims, when properly analyzed, cannot withstand critical scrutiny. The idea of "reasoning without rules" sounds appealing, but it trades upon an ambiguity between formal rules and normative standards. Without normative standards, this approach is no more than a proposal for the accumulation of information about performance. To the extent to which mental-model theory concerns the psychology of reasoning, it has nothing to do with logical competence; and to the extent to which it concerns logical competence, it is not a psychological theory.

NOTES

[1]The lines between premises and conclusion are broken lines, rather than solid lines (which would indicate a deductive argument) or double lines (which would indicate an inductive argument); I want these to be non-committal.

References

Fetzer, James H. (1998), 'Deduction and Mental Models', *Minds and Machines* 0, pp. 000–000.

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

Wason, Peter, and Johnson-Laird, P. N. (1972), *The Psychology of Reasoning*, Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., and Byrne, Ruth M. J. (1998), 'Models Rule, OK? A Reply to Fetzer', *Minds and Machines* 0, pp. 000–000.