Models Rule, OK? A Reply to Fetzer

P. N. JOHNSON-LAIRD[1]
[1]*Department of Psychology, Princeton University, Princeton, NJ 08544, U.S.A.*
E-mail: phil@clarity.princeton.edu

RUTH M. J. BYRNE[2]
[2]*Department of Psychology, Trinity College, University of Dublin, Dublin 2, Ireland*
E-mail: rmbyrne@tcd.ie

How do logically-untrained individuals make deductions? A prevalent view in the psychology of reasoning is that they rely on tacit rules of inference akin to those of a formal logic. In *Deduction* (Johnson-Laird and Byrne, 1991), we argued instead that the untrained mind is not equipped with formal rules of inference, but relies on the general semantic principle of validity: A deduction is valid if the conclusion must be true given that the premises are true. Reasoners put this principle into practice in the following way. They construct *mental models* of the situations described by the premises, formulate a conclusion that holds in these models – if none is provided by a helpful interlocutor – and check its validity by ensuring that it holds in all possible models of the premises. This account has theoretical advantages. It dovetails with other parts of mental life – perception delivers models of the world (Marr, 1982), and comprehension of discourse delivers models of what is described (Garnham and Oakhill, 1996). And it 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. The account also has empirical advantages. It predicts robust phenomena. Reasoners are faster and make fewer errors with deductions that require them to construct only one model than with deductions that require them to construct multiple models. And they characteristically err by drawing conclusions that are supported by one model of the premises.

James H. Fetzer has been kind enough to review *Deduction* twice (Fetzer, 1993, 1998), which is going well beyond the bounds of duty, especially as he does not seem to have a good opinion of it. We are grateful to him for the chance of replying to his latest salvo. It is always tricky for individuals in one discipline to review work in another discipline (or indeed to reply to such reviews). We hope that this reply will help cross-disciplinary understanding.

Fetzer argues that the tenability of our theory depends on the meaning of the phrase, ‘mental model’. We agree. And that is why much of our book is devoted to characterizing mental models and why much of our research is devoted to developing computer implementations of mental models. We stress that our polemical stance concerns only the merits of current psychological theories, not the merits of

Minds and Machines **9:** 111–117, 1999.

different ways of doing logic, because at times Fetzer appears to have logic in mind rather than psychology. We also stress that current psychological theories based on formal rules are odd because they are "incomplete", i.e., there are valid inferences that cannot be proved within them (Rips, 1994),[1] and because they do not make use of certain well-known rules of inference. Fetzer writes that the formal rules of inference in these theories include *modus ponens*, *modus tollens*, and other principles of logical systems of so-called "natural deduction" in which there are rules for each connective. In fact, none of the current psychological theories includes the rule of *modus tollens*. *Modus tollens* inferences are difficult, and these theories account for the phenomenon by dropping its rule from the mind.

As Fetzer remarks, we contend that formal rules are syntactic – indeed, the proponents of these theories make this claim (see, e.g., Braine and O'Brien, 1991; Rips 1994) – whereas mental models are semantic. And he goes on to draw a cogent parallel between making deductions by mental models and justifying formal rules. The parallel is no accident, at least in the case of inferences that hinge on sentential connectives, such as "if", "or", and "and". Logicians have developed both syntactic calculi and semantic systems, such as truth tables, for these inferences. Psychologists used to wonder whether logically-untrained individuals relied on a tacit system of truth tables, but Daniel Osherson (1974–1976) was able to refute the hypothesis. If one adds a new atomic proposition to an argument, the size of its truth table doubles; yet its psychological difficulty does not double. When psychologists learned of this result, they abandoned the idea that reasoning was a semantic process. They turned instead to formal rules. The mental-model theory, however, is a semantic theory, but it does not rely on truth tables, and so it is not embarrassed by Osherson's result.

A fundamental principle of the model theory is that reasoners normally represent only what is true. In this way, they minimize the load on their short-term memory. This idea, which we have recently baptized as the "principle of truth", is subtle because it applies at two levels. First, reasoners represent only true possibilities. Second, within the true possibilities, they represent only those literal propositions (affirmative or negative) in the premises that are true. Thus, given an exclusive disjunction about a hand of cards, such as:

There isn't a king in the hand or else there is an ace in the hand,

reasoners construct two alternative models, which we show here on separate lines:

```
¬king
      ace
```

where '¬' denotes negation. Each model corresponds to a true possibility, and each model represents only those literal propositions in the disjunction that are true within the possibility. Hence, the first model does not represent explicitly that it is false that there is an ace in this case, and the second model does not represent explicitly that it is false there there is not a king in this case. Reasoners make mental "footnotes" to keep track of this false information, and in *Deduction* we introduced

square brackets as a notation for these footnotes (see pp. 45 *et seq.*). However, these mental footnotes are soon forgotten. Only fully explicit models of what is possible given the disjunction represent both its true and false literals in each model:

−king −ace

king ace

But, according to the principle of truth, reasoners do not normally construct fully explicit models, and can do so only for simple premises. It should not be lost on readers that fully explicit models correspond to the true rows in a truth table, whereas mental models are partial representations of the true rows in a truth table. It follows that the number of models, unlike the number of rows in a truth table, does not double as a result of adding a new atomic proposition to an argument.

The theory of mental models is not only an account of sentential reasoning. It has been successfully applied to spatial and temporal reasoning, and to reasoning with quantified assertions. And, since the publication of our book, it has also been successfully applied to counterfactual reasoning (e.g., Byrne, 1996), reasoning from suppositions (e.g., Byrne and Handley, in press), modal reasoning (e.g., Johnson-Laird and Bell, 1997), and to extensional reasoning about probabilities (e.g., Johnson-Laird, Legrenzi, et al. 1997).

Life is not a laboratory in which helpful experimenters draw conclusions for you. You often have to draw them for yourself. Hence, unlike logicians, psychologists need to account for which particular valid conclusions individuals tend to infer. We described this aspect of reasoning in *Deduction* (p. 22):

1. to deduce is to maintain semantic information; i.e., people tend not to throw semantic information away by drawing conclusions that hold for more possibilities than the premises;
2. to deduce is to simplify; i.e., people tend to draw conclusions that are more parsimonious than the premises;
3. to deduce is to reach a new conclusion; i.e., people do not merely restate a premise.

Fetzer notes (as we did) that none of these principles is a matter of logic: Valid conclusions can throw semantic information away, they can be unparsimonious, and they can be old hat. The moral (for us) is simple: Logic alone cannot give a complete account of human deductive competence (*pace* Piaget).

> Other students of logic may share my dismay at discovering that the essence of the theory of mental models can be captured even more adequately by those "rules of thumb" known to every instructor, such as that . . . *disjunctions are only false when both disjuncts are false*, and that *conditionals are only false when their antecedents are true and their consequents are false (together)*.

So Fetzer writes (1998). We can spare him his dismay. His discovery about mental models could not be further from the truth or, more precisely, from the principle of truth. The essence of the model theory has nothing to do with the conditions in which assertions are false, and everything to do with the conditions in which

they are true. Fetzer may counter that this essence is still banal. It certainly seems innocuous. But, lurking within it, as we have recently discovered, is a surprising phenomenon, which cannot be predicted by the current psychological theories based on formal rules. What the model theory predicts is that certain inferences should be illusory, that is, they should have conclusions that seem obvious, that most people draw, and yet that are totally wrong. Here is one of many examples (for another, see Johnson-Laird, 1997):

Only one of the following premises is true:

> There is a king in the hand or there is an ace, or both.
> There is a queen in the hand or there is an ace, or both.
> There is a jack in the hand or there is a ten, or both.

Is it possible that there is an ace in the hand?

Nearly everyone responds: "yes" (Johnson-Laird and Goldvarg, 1997). The theory predicts this response because reasoners fail to consider the false cases. But, the response is an illusion: If there were an ace in the hand, then two of the premises would be true, contrary to the opening claim that only one of them is true. Illusory inferences have so far been demonstrated in deductive reasoning (Johnson-Laird and Savary, 1997), probabilistic reasoning (Johnson-Laird and Savary, 1996), and, as the previous example shows, modal reasoning about possibilities (Johnson-Laird and Goldvarg, 1997). "There is", Fetzer (1998) remarks, referring to mental models, "nothing distinctively semantical about their approach . . . ". But what could be more distinctively semantical than a theory that draws so sharp a contrast between the representation of what is true and the representation of what is false?

Fetzer pooh-poohs our claim that mental models are finite. We wrote (*Deduction*, p. 36):

> The theory [of mental models] is compatible with the way in which logicians formulate a semantics for a calculus . . . . But, logical accounts depend on assigning an infinite number of models to each proposition, and an infinite set is far too big to fit inside anyone's head (Partee, 1979). The psychological theory therefore assumes that people construct a minimum of models . . . .

We are here contrasting mental models with "possible worlds" semantics (not formal rules, as Fetzer implies). He adds that, if our argument were well-founded, it is difficult to imagine how people could add and subtract, since there are infinitely many sets of possible numbers. He goes on: "More importantly, the authors apparently have no understanding of the nature of metatheoretical results, which apply to infinite domains without having to provide a separate proof for each of their instances" (Fetzer 1998). But this argument is a non-sequitur. Our point – Barbara Partee's (1979) point, originally – is that an assertion such as "The boy stood on the burning deck" has infinitely many models in a possible-worlds semantics, each corresponding to the different possible ways in which the sentence could be true; e.g., the boy could be standing at the prow, a fraction of an inch back from it, and so

on *ad infinitum*. No human mind can accommodate all these models. Fortunately, in order to add or subtract, humans are not obliged to have in mind the infinitely many sets of possible numbers at one and the same time.

Fetzer argues that the alleged differences between mental models and formal rules are more apparent than real. Yet he goes on to claim the following difference: "Indeed, the superiority of formal rules over mental models (within sentential logic, for example) can be demonstrated relative to *the desideratum of provability*" (Fetzer, 1998) Here he seems to have forgotten that our concern is how logically-untrained individuals reason, not the best methods of proof, and that the principal difference between the formal-rule theories and the mental-model theory is that they make different predictions. Ironically, the logician Jon Barwise (1993) has shown that a method of proof based on models *is* superior to formal rules. Quine (1974, p. 75) had pointed out that syntactic methods of proof have no general way of establishing invalidity. As he wrote, "failure to discover a proof for a schema can mean either invalidity or mere bad luck". Barwise (1993) shows that the same problem vitiates psychological theories based on formal rules. They, too, can propose only that reasoners search for a proof and, if they fail to find one, judge that an argument is invalid. But this procedure, as Barwise points out, "gives one at best an educated, correct guess that something does not follow" (Barwise 1993, p. 338). Fetzer argues, in effect, that models have the mirror-image deficit when it comes to establishing validity, which calls for showing that there is no model of the premises in which the conclusion is false. The theory founders, he says, because it fails to distinguish between an unsuccessful search for counterexamples and the non-existence of counterexamples. He concludes: "Unless Johnson-Laird and Byrne are prepared to deny the difference between *merely believing that an argument is valid* and *that argument's being valid*, they must admit that their method does not yield an effective decision procedure even for sentential logic" (Fetzer, 1988). In fact, as Barwise shows, we can have our logic and eat it. If we treat each model as representing an indefinite number of possible worlds, then we can establish the validity of an inference by examining a finite number of models.

With Barwise's argument in mind, we will try to elucidate some intricate matters, particularly the distinction between mental models and fully explicit models. Fetzer is right that mental models are not an effective decision procedure for sentential reasoning, nor are they intended to be. Why not? Because people make systematic errors, such as the illusory inferences that we described earlier. So, logically-untrained individuals may believe that an argument is valid, and yet be wrong. In contrast, fully explicit models yield an effective decision procedure that is more efficient than truth tables. (Neither method is tractable, because sentential reasoning itself is not tractable.) The AI algorithm that we described in Chapter 9 of *Deduction* generates all possible models for each premise, and combines them with the models of the previous premises. It evaluates given conclusions as valid or invalid. It also generates the most parsimonious conclusion capturing all the information in the premises – a procedure that is equivalent to minimizing the com-

ponents in a Boolean circuit. Hence, contrary to Fetzer's claim, the mental-model methodology does yield a decision procedure for sentential reasoning.

There is one further clarification to be made. We pointed out that human reasoners can draw a tentative or probabilistic conclusion. Fetzer suggests that we must be studying an alternative conception of reasoning. "Tentative and probabilistic reasoning," he remarks, "exemplify inconclusive inductive reasoning . . . " (Fetzer 1998). Not necessarily. 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.

Likewise, there are many valid arguments yielding probabilistic conclusions. Suppose, for example, a binomial test yields 0.01 as the conditional probability of some data given the null hypothesis. This result is deductively valid granted the data and the assumptions of the test. What is inductive, of course, is to make the further step of rejecting the null hypothesis. A simpler example of a valid inference yielding a probabilistic conclusion is as follows:

> If there is a red or a green marble in the box, then there is a blue marble in the box.
> ∴ It is more probable that the blue marble is in the box than that the red marble is in the box.

Reasoners appear to make this inference by examining the proportions of models in which the two events occur (Johnson-Laird, Legrenzi, et al. 1997).

In his conclusion, Fetzer writes that humans may use different types of reasoning to achieve different purposes. We agree: They may. The issue is an empirical one. He goes on to argue that it would be a mistake to think that mental models could or should displace formal rules. The tenability of this proposal depends on the meaning of the phrase "formal rules". If, as we suspect, Fetzer means formal systems in logic, then we agree. Mental models are a psychological theory, not a rival logic. But, if he means current psychological theories of reasoning based on formal rules, then we disagree. Mental models could replace formal rules; the robust psychological data suggest that they should replace them, too.

1. Notes

[1] Editor's Note: See the Discussion Exchange between Johnson-Laird and Rips: Johnson-Laird, Philip N. (1997), 'Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof*', *Minds and Machines* 7: 387–407.

[2] Rips, Lance J. (1997), 'Goals for a Theory of Deduction: Reply to Johnson-Laird', *Minds and Machines* 7: 409–424.

[3] Johnson-Laird, Philip N. (1997), 'An End to the Controversy? A Reply to Rips', *Minds and Machines* 7: 425–432.

References

Barwise, Jon (1993), 'Everyday Reasoning and Logical Inference', *Behavioral and Brain Sciences* 16, pp. 337–338.

Braine, Martin D. S., and O'Brien, David P. (1991), 'A Theory of If: A Lexical Entry, Reasoning Program, and Pragmatic Principles', *Psychological Review* 98, pp. 182–203.

Byrne, Ruth M. J. (1996), 'Towards a Model Theory of Imaginary Thinking', in Jane Oakhill and Alan Garnham, (eds.), *Mental Models in Cognitive Science: Essays in Honour of Phil Johnson-Laird*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 155–174.

Byrne, Ruth M. J., and Handley, Simon J. (in press), 'Reasoning Strategies for Suppositional Deductions', *Cognition*

Fetzer, James H. (1993), 'The Argument for Mental Models Is Unsound', *Behavioral and Brain Sciences* 16, pp. 347–348.

Fetzer, James H. (1998), 'Deduction and Mental Models', *Minds and Machines* 8: 000–000.

Garnham, Alan, and Oakhill, Jane V. (1996), 'The Mental Models Theory of Language Comprehension', in Bruce K. Britton and Arthur C. Graesser (eds.), *Models of Understanding Text*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 313–339.

Johnson-Laird, P. N. (1997), 'Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof'*, *Minds and Machines* 7: 387–407.

Johnson-Laird, P. N., and Bell, Victoria A. (1997), 'A Model Theory of Modal Reasoning', *Proceedings of the 19th Annual Conference of the Cognitive Science Society (Stanford University)*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 349–353.

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

Johnson-Laird, P. N., and Goldvarg, Yevgeniya (1997), 'How to Make the Impossible Seem Possible', *Proceedings of the 19th Annual Conference of the Cognitive Science Society (Stanford University)*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 354–357.

Johnson-Laird, P. N.; Legrenzi, Paolo; Girotto, Vittorio; Legrenzi, Maria S.; and Caverni, Jean-Paul (1997), 'Naive Probability: A Model Theory of Extensional Reasoning', under submission.

Johnson-Laird, P.N., and Savary, Fabien (1996), 'Illusory Inferences about Probabilities', *Acta Psychologica*, 93, pp. 69–90.

Johnson-Laird, P.N., and Savary, Fabien (1997), 'Truth and Illusory Deductions', under submission.

Marr, David (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.

Osherson, Daniel N. (1974–1976), *Logical Ability in Children*, Volumes 1–4, Hillsdale, NJ: Lawrence Erlbaum Associates.

Partee, Barbara H. (1979), 'Semantics – Mathematics or Psychology?', in Rainer Bauerle, Urs Egli, and Arnim von Stechow (eds.), *Semantics from Different Points of View*, Berlin: Springer-Verlag.

Quine, Willard Van Orman (1974), *Methods of Logic; 3rd edition*. London: Routledge.

Rips, Lance (1994), *The Psychology of Proof*. Cambridge, MA: MIT Press.