Deduction and Mental Models

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

P. N. Johnson-Laird and Ruth M. J. Byrne, *Deduction*, Essays in Cognitive Psychology, Hillsdale, NJ: Lawrence Erlbaum Associates, 1991, xii + 243 pp., $49.95 (cloth), ISBN 0-86377-149-1.

Some theories have such appealing names that their truth virtually appears to be self-evident. The thesis that humans perform deductions by using mental models, for example, invites the generalization that humans may not only reason but also think by means of mental models. Indeed, the notion of a model is sufficiently ambiguous that, in one or another sense, both these theses might be true. But they might not be true in other senses. Indeed, when "models" are identified with *signs* as things that stand for other things, for example, it may be the case that reasoning involves models because all kinds of thinking require signs (Fetzer 1988). Yet when "models" are identified with *physically isomorphic counterparts*, such as scale models of battleships, it would be false.

The tenability of the theory advanced by Johnson-Laird and Byrne, namely, that humans perform deductions by means of mental models, thus depends in large measure on the meaning of that phrase. In its general sense, the process of deduction is supposed to involve three kinds of thinking: *comprehension*, during which reasoners use their understanding of language and background knowledge to construct an internal model of the premises of an argument; *description*, during which reasoners attempt the parsimonious re-description of the contents of those premises, making some previously implicit content explicit; and *validation*, during which reasoners search for alternative models in which their putative conclusion is false, where a conclusion follows from its premises so long as no such counterexamples exist (pp. 35–36).

The authors maintain that their account ought to be preferred over prevalent conceptions of deductive reasoning, according to which deduction involves following formal rules of inference or applying content-specific generalizations. The formal rules of inference they have in mind include *modus ponens*, *modus tollens*, and other principles of systems of natural deduction, while the content-specific generalizations they cite include inferences from antecedents to consequences or from conditions to actions of kinds familiar within various contexts of AI. An expert system such as DENDRAL, for example, analyzes mass spectrograms on the basis of content-specific production rules of such a kind: IF there is a high peak at 43 *amu*s, a high peak at 71 *amu*s, a high peak at 86 *amu*s, and any peak at 58 *amu*s, THEN there is an N-PROPYL-KETONE3 substructure (p. 33; cf.

Barr and Feigenbaum 1982, pp. 106–110; '*amu*' stands for "atomic mass unit").

From the perspective of epistemology, the differences between formal rules of inference (such as *modus ponens*) and content-specific generalizations (used by DENDRAL) are obvious, since the former can be justified on logical grounds, while the latter require empirical justification as constant conjunctions, statistical correlations, or natural laws. Students of logic should readily discern that the differences that Johnson-Laird and Byrne allege to obtain between formal rules of inference and the use of mental models are not equally apparent. They contend that formal rules of inference are syntactic, while mental models are semantic, but their difference may be far less significant than they suggest, because the adoption of syntactic rules requires justification on semantic grounds.

Consider, for example, *modus ponens* itself. This rule, properly understood, maintains that, from two lines of the forms '$p \rightarrow q$' and 'p', respectively, a new line of the form 'q' may be obtained. Within sentential calculus, for example, this syntactical rule warrants adoption on semantic grounds, namely, that no inference from premises having these forms can yield a conclusion of counterpart form that is false when those premises are true. The adoption of rules of this kind depends upon their invulnerability to counterexamples, which can be demonstrated by truth tables. It is well known that an argument (argument form) is valid (in sentential logic) when and only when its corresponding conditional – formed by taking the conjunction of its premises as antecedent and its conclusion as consequent – is a tautology (tautologous) (Gustason and Ulrich 1973, pp. 58–59).

The similarities between the processes of justifying formal syntactic rules of inference and of performing deductions by means of mental models are striking. Both require the construction of models of the premises of arguments (argument forms) by reasoners on the basis of their understanding of language and background knowledge, which is the stage of *comprehension*. (The internal/external difference does not distinguish between them, since models of both kinds can be thought through or written down.) Both require searching for alternative models (or truth-table assignments) in which putative conclusions are false when their premises are true, where a conclusion follows from its premises (as a valid conclusion) provided no counterexamples exist, which is the stage of *validation*.

Such differences as may obtain between them, therefore, appear to arise at the stage of *description*, during which reasoners attempt the parsimonious re-description of the contents of those premises, making some previously implicit content explicit. Johnson-Laird and Byrne accent this aspect of reasoning due to their concern with *the psychology of deduction*. Much of what appears to be going on here revolves about traditional distinctions within philosophy between "the context of discovery" and "the context of justification", where the source of an idea (where it came from, how it was discovered) is a psychological activity that does not have to satisfy normative standards, while the acceptance of an idea (whether it should be taken to be true or acted upon) is a logical activity that does have to satisfy normative standards (Fetzer and Almeder 1993, p. 29).

At least three issues arise here that need to be carefully disentangled. First, the assumption that usually prevails in relation to the evaluation of arguments within philosophical analyses takes for granted that their conclusions as well as their premises are specified. The purpose of deductive logic is therefore that of certifying (validating) whether or not a conclusion that has already been drawn follows from its premises rather than that of establishing guidelines or rules of reasoning that would cause a reasoner to draw a specific conclusion because s/he has accepted specific premises. The methodology that Johnson-Laird and Byrne adopt – of inviting subjects to draw inferences in an "open-ended" fashion without specifying targeted conclusions - thus implies a rather different conception.

Second, nothing about deductive reasoning as a logical activity motivates a desideratum of making some previously implicit content explicit. Arguments in which the same sentence consisting of the same words in the same sequence occurs in both premises and conclusion are no less valid than others in which conclusions consist of words in sequences that do not appear in the premises. From the perspective of psychology, no doubt reasoners typically, if not always, draw conclusions that make explicit previously implicit content. It ought to be observed, however, that infinitely many conclusions are validly derivable from even the simplest premises (since 'p' implies 'p or q') and that inductive arguments are never valid, yet may have conclusions that are perfectly proper.

Third, the re-description of part or all of the content of the premises of arguments does not need be "parsimonious" for such arguments to qualify as valid. It is characteristic of valid arguments that their conclusions recapitulate (part or all of) the content of their premises, which supplies an explanation for why their conclusions cannot be false when their premises are true (Fetzer 1990, pp. 101–102). But Johnson-Laird and Byrne consider "the essence of the theory" of mental models to be that "people use models that make explicit as little information as possible, and in this way, they overcome the unwieldy bulk of truth tables" (p. 52). They illustrate what they have in mind with a set of models for disjunction, for example, where only lines for true disjuncts are explicitly given.

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 *conjunctions are only true when both conjuncts are true*, 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)*. Moreover, it should be evident that these models are syntactical, where '$\bigcirc$'s and '$\triangle$'s stand in for 'p's and 'q's within more familiar systems (p. 52). There is nothing distinctively semantical about their approach, which becomes evident from the use of '$-$' as a sign for negation. '$\bigcirc$' and '$\triangle$' function as variables in the same way as do 'p' and 'q'.

Another benefit alleged to derived from mental-model methodology is said to be that, unlike formal rules, it only depends upon finite numbers of models:

> ... logical accounts depend on assigning an infinite number of models to each

> proposition, and an infinite set is too big to fit inside anyone's head .... The psychological theory therefore assumes that people construct a minimum of models: they try to work with just a single representative sample from the set of possible models, until they are forced to consider alternatives. (p. 36.)

If such arguments were well-founded, of course, it is difficult to imagine how people could add and subtract, since there are infinitely many sets of possible numbers. 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. The advantages of this approach are enormous, of course, and encompass the use of variables.

The alleged differences between "mental models" and "formal rules", therefore, are more apparent than real. Indeed, the superiority of formal rules over mental models (within sentential logic, for example) can be demonstrated relative to *the desideratum of provability*. When a proof is understood to be a sequence of lines where every line is either given as a premise or obtained from preceding lines using the rules, and the last line is the desired conclusion, then any argument that satisfies these conditions in relation to an acceptable set of formal rules must be valid. When arguments are appraised using formal rules, proofs can establish their validity. There can still be valid arguments for which proofs are unavailable, but arguments for which proofs are available are valid.

When a proof is understood to be a mental model for which reasoners have searched for alternative models of their premises in which their putative conclusions are false, however, it should be obvious that the existence of counterexamples is perfectly compatible with the failure to discover them (Fetzer 1993). The methodology of mental models thus founders upon a crucial equivocation, namely, the difference between an unsuccessful search for counterexamples and the nonexistence of counterexamples. 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.

The application of truth-table procedures, by contrast, provides an effective decision procedure, which can be used to determine the validity of arguments within sentential logic. These procedures are routine (or mechanical), completable (in a finite sequence of steps), and conclusive (their solutions are definitive). Because human beings differ greatly in both logical acumen and capacity for imagination, the methodology of mental models cannot satisfy these conditions. It requires the use of imagination to consider possible counterexamples, where even thinking long and hard may not exhaust them, and where failure to discover them does not mean that they do not exist. The application of mental-model methodology, even when diligently pursued, cannot guarantee validity.

The implications of this limitation would be devastating were these authors attempting a normative analysis; instead, they casually observe, "If it is uncertain whether there is an [additional yet undiscovered] alternative model of the premises,

then the conclusion can be drawn in a tentative or probabilistic way" (p. 36). Since valid deductive reasoning is conclusive, where the conclusion of a valid deductive argument cannot be false when its premises are true, Johnson-Laird and Byrne are studying an alternative conception, where reasoning can be *supposed to be* deductive even when the possibility remains that the conclusion can still be false. Tentative and probabilistic reasoning exemplify inconclusive inductive reasoning rather than conclusive deductive reasoning.

Thus, the core of their empirical research consists in presenting premises to subjects in order to ascertain what conclusions they will draw. One example they use is intended to measure the use of exclusive disjunction (p. 54):

> Linda is in Amsterdam or Cathy is in Majorca, but not both.
> Linda is in Amsterdam.
> What follows?

The appropriate inference to draw, of course, is that Cathy is not in Majorca. Another is intended to measure the use of negation and disjunction (p. 55):

> Either Steven is in Donegal or Jenny is in Princeton, but not both.
> Jenny is in London.
> What follows?

The appropriate inference to draw is that, since Jenny is in London, Jenny is not in Princeton; since Jenny is not in Princeton, it follows that Steven is in Donegal.

Studies of this kind conducted involving various forms of deductive reasoning for *modus ponens*, *modus tollens*, affirmative disjunction, and negative disjunction (above) have displayed these tendencies to derive valid conclusions (p. 55):

Modus ponens:	91% correct
Modus tollens:	64% correct
Affirmative disjunction:	48% correct
Negative disjunction:	30% correct

The relative frequency of the occurrence of valid deductive reasoning in the form of correct answers to such questions, therefore, provides an empirical measure of the extent to which human beings are skillful in their deductive reasoning, on the assumption that these questions test deductive reasoning. But they obviously presuppose the availability of normative standards for determining the validity of deductive arguments on independent grounds.

Such considerations tend to remove the presumptive tension that the authors suggest obtains between formal rules and mental models. The purpose of formal rules is to codify the conditions under which deductive reasoning is properly qualified as "valid". The purpose of mental models is to represent the extent to which human beings satisfy those standards in their ordinary reasoning. Even if human reasoning typically involves the stages of comprehension, re-description, and validation, therefore, that does not mean that the conclusions they infer have to be

true. The distinction between believing that an argument is valid and its actually being valid is thus presupposed and not explained by this methodology, which does not specify conditions of validity.

When these issues are placed in perspective, it becomes apparent that humans may use different types of reasoning to achieve different purposes. When we theorize about reasoning, we tend to use formal rules for specifying sentence and argument forms, which supply the apparatus required to make assertions about infinite domains that are precise, concise, and general (Fetzer 1996, pp. 109–110).[1] Because mental models do not supply a metatheory for normative reasoning, and because formal rules are not intended to describe ordinary reasoning, it should be evident that they do not qualify as alternatives for understanding the same domain and are not meant to fulfill the same function. Whether or not mental models capture the elements of ordinary reasoning, it would be a mistake to think that mental models could or should displace formal rules.

NOTES

[1]Editor's note: See Leiber, Justin (forthcoming), Review of Fetzer 1996, *Minds and Machines*.

References

Barr, Avron, and Feigenbaum, Edward A., eds. (1982), *The Handbook of Artificial Intelligence*, Vol. II, Reading, MA: Addison-Wesley.

Fetzer, James H. (1988), 'Signs and Minds: An Introduction to the Theory of Semiotic Systems', in James H. Fetzer, ed., *Aspects of Artificial Intelligence*, Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 133–161.

Fetzer, James H. (1990), *Artificial Intelligence: Its Scope and Limits*, Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Fetzer, James H. (1996), *Philosophy and Cognitive Science, 2nd edition*, Minneapolis, MN: Paragon House.

Fetzer, James H. and Almeder, Robert F. (1993), *Glossary of Epistemology/Philosophy of Science*, New York: Paragon House.

Gustason, William, and Ulrich, Dolph E. (1973), *Elementary Symbolic Logic*, New York: Holt, Rinehart and Winston.