Propositional Reasoning by Model?

Luca Bonatti

Two theories of propositional deductive reasoning are considered: Johnson-Laird's mental models and Braine's mental logic. The model theory is said to account for practically all of the known phenomena of deductive propositional reasoning, offer a general theory of conditionals, account for the most important aspects of Braine's theory, and predict new phenomena that rule theories cannot explain. I argue that (a) the model theory is flawed in a way that is difficult to overcome, (b) conditionals are seriously misrepresented, (c) the algorithms proposed to implement it either allow invalid inferences or are psychologically useless, (d) Braine's theory accounts for all of the new phenomena worth considering, and (e) the model theory can predict Braine's results only at the cost of self-refutation. I conclude that the mental model theory of propositional reasoning offers no reason to reject the program of mental logic.

In the last 10 years, an impressive literature has grown under the title of mental models. Most of it has a common theme: It contraposes a theory of deductive reasoning in terms of rules to a theory in terms of models, and then it shows the superiority of the latter. Yet much of this literature has an original sin. It makes very little sense to compare two theories in the absence of the second one, and in most domains there is practically no theory of mental logic that at the same time is psychologically motivated and sufficiently developed to hold scrutiny. The psychological value of a mental logic theory of a given domain of reasoning depends on how its details are spelled out. Specifying a set of rules is only a first step for it to be psychologically testable: The rules themselves must have independent psychological plausibility. This is still not enough. Because a sentence can have infinite proofs from a given set of axioms, the theory must also indicate the procedures implementing the rules, their order of accessibility, and their relative difficulty; define some measure of complexity over proofs that correlates with psychological measures; and test them against types of problems that are likely to really engage reasoning competencies.

To my knowledge, the only theory of mental logic that addresses all these questions and can approach a psychologically real model of human reasoning is the one developed by Braine and his collaborators (Braine, 1990, 1994; Braine & O'Brien, 1991; Braine, Reiser, & Rumain, 1984; Lea, O'Brien, Fisch, Noveck, & Braine, 1990). Although extremely rich, it only concerns propositional reasoning, that is, a minimal fragment of reasoning abilities.

Johnson-Laird, Byrne, and Schaeken (1992) and Johnson-Laird and Byrne (1991) developed an alternative treatment of propositional reasoning with mental models, thus allowing, for the first time, a direct comparison of two existing and competing theories of deductive reasoning. They presented the overall outcome of the comparison with mental logic as a complete victory: "The evidence for our experiments has enabled us to compare formal rules and mental models as theories of human reasoning. There will be no prize for guessing the outcome (Johnson-Laird & Byrne, 1991, p. x). In a more modest manner, for propositional reasoning they claimed that the model theory accounts for practically all the known phenomena of deductive propositional reasoning; that it offers a general theory of conditionals; that it accounts for the most important aspects of Braine's theory; and that as a further advantage, it predicts new phenomena that rule theories cannot explain.

In this article, I claim that the model theory is flawed in a way that is difficult to overcome; that the "new predictions" should not be taken too seriously; that anyhow, Braine's theory accounts for practically all of them, although for different reasons; that the model theory can predict Braine et al. (1984) results only at the cost of self-refutation; and that even if its present defects were overcome, it is still likely to be false. I assume the reader's familiarity with mental models in general and with Braine's system as presented in Braine et al. (1984) and recall the aspects of propositional mental models relevant to my arguments.

Structure of the Theory

The mental model theory of propositional reasoning aims to be "a theory that reconciles the semantics of truth tables with the constraints on mental processing, and does so in a way that explains human performance" (Johnson-Laird & Byrne, 1991, p. 43). It thus intends to be a psychologically plausible version of the truth table theory. The latter holds that when one understands propositional premises, a data structure is built that corresponds to the truth value assignments of a truth table, and, notoriously, has many problems. There is no evidence that reasoners compute truth tables (Braine & Rumain, 1983; Osherson, 1975). More important, truth tables for complex sentences grow exponentially with respect to the number of atomic sen-

Luca Bonatti, Laboratoire des Sciences Cognitives et Psycholinguistique, Paris, France, and Philosophy Department, Rutgers University.

I am grateful to Emmanuel Dupoux, Jacques Mehler, and Christophe Pallier for comments on the first draft of this article, and to P. N. Johnson-Laird for his constructive criticisms. Martin Braine gave me important suggestions for improvements, and I am indebted to him for many of the points I have made. Usual attributions and disclaimers of responsibilities follow.

Correspondence concerning this article should be addressed to Luca Bonatti, Laboratoire des Sciences Cognitives et Psycholinguistique, 54 Boulevard Raspail, Paris 75006 France.
ternal sentences they contain; that is, they grow too fast to be psychologically plausible. The mental model theory would avoid such problems by assuming that, instead of computing full truth tables, reasoners represent the minimal amount of information required; in practice, only those value assignments that make the premises true, and only the minimal amount of those value assignments. I briefly recall how it handles disjunction, negation, and conditionals.

When given ‘A or B’, reasoners first form a representation containing the two models:

A
B

which is consistent with both the inclusive and exclusive readings of the disjunction. Thus, mental models allow one to explain why people are often not aware of the ambiguity.

When other factors force a reasoner to choose one interpretation, information previously kept implicit can be made explicit, issuing the correct reading. This may occur in two ways: First, all the information can be explicitly brought to memory. Such a process of “fleshing out” information can be conceived to occur in stages: New models that can be constructed out of the input sentences, may be progressively added until no new model can be built. Second, an exhaustivity operator (indicated here with square brackets) can be used. This is best seen as an operator over types of symbols in a model that, in its restricted basic meaning, has the following effect: If in a collection of models the tokens of a type are bracketed, when new information is added to it no new token of that type can appear either in the old or in the newly added models. For example, exclusive disjunction is obtained by exhausting the items represented in previous models:

[A]  
[B]

whereas, when the disjunction has to be interpreted inclusively, a new model is added (thus models are “fleshed out”) and types are exhausted:

[A]  
[B]  
[A]

To clarify the relation with a truth table, notice that if one maps the models into it, they correspond to the description of the truth values of the atoms in the rows in which the inclusive (incl), or exclusive (excl), or indeterminate (indet) disjunctions are true:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A \lor \text{incl B}</th>
<th>A \lor \text{excl B}</th>
<th>A \lor \text{indet B}</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>*</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Thus, basically propositional models modify the classical truth table theory by allowing truth value gaps and assuming that subjects minimally represent those truth assignments for which a partial function gives the value True, and maximally represent those truth assignments for which the corresponding total functions give the value True.

Negation is somewhat special. When an atom is negated, it is represented directly into the models as

\neg A,

where ‘\neg’ is a “propositional-like tag” (Johnson-Laird & Byrne, 1991, p. 44). So, for atoms, negation is represented in its propositional form and is not an operation over models. However, for a complex sentence, a negation should return the complement set of the models for that sentence; in that case, negation is the operation of complementation.

Also, a sentence containing conditional expressions is ambiguous. Its initial interpretation gives rise to these models:

[A]  
B  

where the dots stand for an empty or implicit model, possibly to be filled in. Just like for disjunction, the interpretation corresponds to the first row of the truth table of the material conditional, and the initial model is neutral with respect to the material conditional or the biconditional interpretations. Additional information can bend it into a biconditional, either by exhaustively tokening the tokens in the initial model, thusly:

[A]  
[B]  

or by fleshing it out completely, again, according to the relevant rows of the biconditional truth table:

[A]  
[B]  
[\neg A]  
[\neg B]

Alternatively, fleshing out the initial implicit interpretation can give rise to the material implication:

[A]  
[B]  
[\neg A]  
[\neg B]

The representation of the conditional is extremely flexible. Thus, the models for ‘A only if B’ are different from those for ‘If B then A’. The latter yields one initial model, but the former yields two models:

[A]  
\neg A  
\neg B  

which would explain why ‘only if’ problems are harder than ‘if . . . then’ problems, but at the same time with ‘if then’ sentences modus tollens (MT) is more difficult than modus ponens (MP), whereas with ‘only if’ sentences it is not. But also ‘If A then not B’, and ‘If not A then B’ have their own special way to give rise to models. Unlike simple conditionals, they yield more models; in that order,

[A]  
\neg B  
B  

and
This is because "a negation is likely to call to mind the affirmative alternative" (Johnson-Laird & Byrne, 1991, p. 67); such differences allow one to explain, inter alia, various aspects of the selection task (pp. 79–81).

As these examples show, the theory of mental models for connectives is very flexible—in fact, so flexible that it can be made consistent with almost everything one wants: correct performance, incorrect performance, individual differences, ambiguities, and lack of ambiguities, whenever any of these occur. The question is whether or not such flexibility is compatible with coherent predictions.

Finally, I recall that, in general, for the theory the difficulty of problems is (mostly) a function of the number of models they require. However, "number of models" is ambiguous in at least three ways: It may mean the sum of models required by the interpretation of each premise, or the sum of models required by the interpretation of each premise plus the number of models required by the conclusion, or else just the number of models required by the integration of the premises on which the conclusion is evaluated. An analysis of the literature on mental models (e.g., Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1989, 1991; Ehrlich & Johnson-Laird, 1982; Garnham, 1987) reveals that it is the last way of counting models that is relevant. For the integrated models of premises, the limit beyond which reasoning becomes too difficult is also precisely specified: "Once a deduction called for three models, it became almost impossible for our subjects" (Johnson-Laird et al., 1992, p. 434). Should this limit be overcome, the model theory would be refuted: "[The model theory is in principle simple to refute: An easy deduction that depends on many models violates its principal prediction]" (p. 436). Keep this in mind for further use.

Do We Know Enough to Make Predictions?

Johnson-Laird et al. presented a full theory and two partial implementations: an artificial intelligence (AI) and a "psychological" (sic: p. 424) algorithm. The full theory is, in fact, only a general sketch of a theory; "full" only indicates that all of the machinery for models is exploited. It is the full theory that is the actual basis for predictions, and not by chance: All of its devices are needed for them to turn out right. However, little instruction is provided as to how and when crucial tools such as complementation, exhaustion, or fleshing out, are used, and the description of the rules governing them is not sufficient for understanding what to do in the generality of the cases. For example, should one take literally the idea that a negation always involves construction of the complement of the models of a sentence?

The initial interpretation of nonnegated complex sentences is often ambiguous, and this allows one to predict some phenomena; but how do we interpret their negation? Consider disjunction: Its initial models are compatible with both its inclusive and exclusive readings. However, their complement is not uniquely defined; to compute it, one should first flesh out the models and then process the negation. But how should that be done? How does one know whether disjunction is to be disambiguated in one way or the other? And what will the claim be? That disjunction is perceived as undifferentiated when not negated, but that people know its precise meaning when it is negated? If so, why should they not know it for the nonnegated case? All the subtle differences between the multiple interpretations are lost if one literally applies the rule for negation. Or, we are told, a negated token "calls to mind" another model including its affirmative alternative, and this allows one to make yet other predictions. However, is one always supposed to build an extra model? If so, the representation of 'Not A' would be the same as the initial representation of 'Either not A or A'; and one should be entitled to predict that MP from 'If A then not B', which requires two models, is more difficult than MP from 'If A then B'. Do we really want that? And what do we do with the extra model when new sentences are to be integrated? Do we eliminate it, or do we carry it with us? Or, when are implicit models filled in by new information, and when are they left empty? Finding the general rules governing these operations is a hard task: The full theory is just not well defined enough.

Where does one look for further light? The obvious places are the algorithms implementing the theory.

The AI Algorithm Is Psychologically Useless

One of the algorithms, the AI algorithm, is meant to reproduce "a mark of human intelligence" (Johnson-Laird et al., 1992, p. 426), namely, people's ability to draw parsimonious conclusions. In it, connectives correspond to the usual set-theoretic operations: "and" is associated to the Cartesian product, negation returns the complement of a set of models, and "the meanings of the other propositional connectives are defined in terms of [Cartesian product] and [complementation]" (Johnson-Laird & Byrne, 1991, p. 176). Thus, this algorithm significantly differs from the full theory. There are no ambiguities of interpretation, no gaps, no empty models, no fleshing out, and no exhaustivity operators: only truth tables with some optimization. Thus, the problem of finding out what happens in the dubious cases does not arise. Connectives are interdefinable; disjunction corresponds to the usual inclusive disjunction; and implication corresponds to material implications, with all its paradoxes. All of the fine explanations of most experimental findings (including the ones "discovered" by Johnson-Laird et al., 1992) are lost. Worst of all, even if the algorithm avoids exponential growth, models for complex sentences including implications or disjunctions still grow too fast. To conclude 'If A then D' from 'If A then B', 'If B then C', and 'If C then D', subjects should compute three models for each premise, or a total of five for the composition of the premises—far too much for a theory that claims that three models give a floor effect. Thus, the AI algorithm does not provide a psychologically plausible model of reasoning. The authors would agree with this conclusion; however, the trouble is that their more psychologically plausible algorithm is mistaken, and fixing it would make it very close to the psychologically useless algorithm. I now turn to show this point.

---

1 Paradoxes of material implications are, indeed, a feature of the "full" theory as well. This, in itself, should cast doubts on a theory that claims to explain the role of content in reasoning.
The Psychological Algorithm Licenses Many Invalid Inferences

The second, "psychological," algorithm maintains the basic structure of the AI algorithm patterned on classical logic while trying to introduce at least some of the psychologically motivated features of mental models, notably, truth value gaps and the principle that the least possible amount of information is made explicit. Johnson-Laird et al. (1992, p. 425) specified the following rules of composition:

1. Every new premise is added to the model(s) by Cartesian product, with the following conventions:
   - explicit (set of) model(s) × explicit (set of) model(s) = one single (set of) model(s) joining the information from the two (sets of) models;
   - implicit model × implicit model = implicit model;
   - implicit model × explicit model = nil.
2. Models containing a token and its negation are eliminated.
3. Models containing repeated information are reformulated by eliminating redundancies.

Notice, first, that the psychological algorithm does not contain either an exhaustivity operator or fleshing-out procedures, and so it cannot elucidate many aspects of reasoning that, according to the "full" theory, require such mechanisms. Furthermore, for the reasons outlined earlier, without fleshing out it is not clear how the complement of models for negative complex sentences should be computed. But, more important, the given rules license many invalid inferences. Just apply them to:

\[
\text{If there is an } A, \text{ then there is a } B
\]

\[
\text{There is a } C
\]

They will generate:

\[
A \times B \times C \text{ yields } ABC
\]

\[
\cdots \times C \text{ yields } \text{nil};
\]

that is, after some cleaning up:

\[
A \times B \times C,
\]

that is, the model of a conjunction. And the procedure to generate conclusions (by following the principle of parsimony) will output 'There is an A' and 'There is a B'. Thus, a conditional is tantamount to a conjunction whenever it occurs next to a categorical statement, and one completely irrelevant to the conditional! Notice, too, that by substituting 'B' for 'C' in (2) we obtain the affirmation of the consequent fallacy as a particular case.

Consider also what happens if one substitutes 'B and not B' for 'B' in (1). By straightforward application of the rules, one gets a contradiction. This latter accident is blocked not because the algorithm is coherent but because it cannot express contradictory sentences, and a fortiori cannot express conditionals with a contradictory consequent, such as 'If A, then B and not B'. Worse, this is not a shortcoming of the psychological algorithm, but of the current full theory when models are left implicit (see Appendix). Indeed, two issues must be kept separate: one is what a system can prove; another is what a system can express. The mental model theory cannot prove contradictions, and correctly so, just as Braine's mental logic theory cannot. However, mental logic can express everything, contradictions included. Mental models, on the other hand, cannot because they do not represent truth conditions of sentences, but (as opposed to truth tables), at most, those situations in which a sentence is true. Thus there is no mental model for a contradiction, which is never true (but has perfectly well-specified truth conditions: truth conditions that never obtain. We have a dilemma: if the theory could express a sentence it should be able to express, then the rules for the algorithm would be inconsistent.

The problem with the given rules is general. They contain a mistake concerning implicit models: Their role is annihilated because the rules forbid new information to be added to them. As a result, the meaning of conditional statements is hopelessly distorted. However, the full theory does not seem in better shape. When they informally illustrated how models are composed in the full theory, Johnson-Laird et al. (1992) did not use the rules they specified for their psychological algorithm (in which case the theory would issue the wrong predictions) but the examples from which the reader is supposed to extract the rules governing implicit models (see Johnson-Laird & Byrne, 1991, pp. 47, 48, and 51; Johnson-Laird et al., 1992, pp. 422–423) seem to lead to the same shortcoming as the psychological algorithm. The underlying rule seems to state that when a new (nonimplicit) model can be incorporated into an already existing collection including an implicit model, the implicit model is eliminated; when it cannot (because that would lead to a contradiction), then inconsistent models are eliminated and the new information is added to the implicit model. But then composing (1) and (2) according to the full theory will give the same results of the psychological algorithm, and also the full theory will be invalid. Because we are not in the business of founding mathematics, proposing invalid reasoning procedures is not, in principle, wrong; the question is whether the theory wants to claim that in point of competence our reasoning system is invalid. And this is just not in the official agenda of mental modelers.

Now, it is interesting to ask at what costs the algorithm could be amended. Two ways out could be attempted. Fleshing out of the conditional before computing the Cartesian product would be necessary to block the unwanted conclusions, but the algorithm does not include fleshing-out procedures. Thus, the first possibility would be to introduce them and require models to be fleshed out before composing premises. However, for this procedure not to be arbitrary, one should suppose that it applies to any conditional, at least when combined with nonconditional sentences. As a result, first, the theory would lose the possibility of exploiting such a vast explanatory space; for example, all of the above explanations of how one reasons using different kinds of conditionals would be lost. Second, the algorithm would converge toward the AI algorithm and would therefore lose its psychological plausibility: The number of models in problems including conditionals would grow too fast to support the right predictions for minimally complex problems.

The other solution would be to devise a rule that allows information to be added to empty (implicit) models, such as the following: 'Implicit model × explicit model = explicit model'. However, once again, the theory would issue the wrong predictions. Each composition of conditional statements containing n atoms would have at least n models, excluding the implicit one. Thus, subjects should not be able to solve problems including chains of conditionals such as 'If A then B, if B then C, if C then
D, if D then E. But A. Therefore E’, but such simple chains are clearly within subjects’ reasoning competencies.

I conclude that the “full” theory is not so full; that, of the two algorithms that are supposed to elucidate it, the first one is psychologically implausible and the second is deeply flawed; and that fixing the latter one is likely to be hard and unlikely to improve the chances of coming up with a psychologically plausible theory.

The Evidence for Propositional Mental Models: What It Does Not Show, and Why Braine’s Theory Predicts What It Does Show

I now turn to the supposed new evidence for the full theory. I will make the following two points: I argue that the clearly interpretable cases are explained by Braine’s theory as well; and I also show that the same mechanisms generating the new evidence allow one to set up cases that refute the full theory.

“A new theory should suggest new phenomena. The present theory does indeed lead to some novel predictions” (Johnson-Laird et al., 1992, p. 430). The authors presented four experiments to show such predictions. Before considering them, I need to make two observations.

The first observation concerns the type of task involved. In all of the experiments, subjects were asked to freely generate conclusions out of given premises, whereas in Braine et al. (1984), subjects had to judge whether a conclusion followed from premises or not. The first task tests production, the second one recognition. As in other cases in which production and recognition are compared, one should not expect the two tasks to cover the same ground. Suppose we asked a group of subjects to freely continue a given story, and another group to judge whether a given continuation of the same story makes sense. We should not expect them to perform equally well. The story would have the same difficulty, so either subjects understand it or they do not, in both cases; however, although subjects may be able to judge the consistency of a given continuation, the story may be just too unconstrained for them to suggest a continuation on their own. Failure to do it would not indicate a deficit in their “story understanding” abilities, but just their lack of fantasy. In the same way, one has to be very careful in judging what results about free production of conclusions exactly indicate. Failure to reach a conclusion on a problem does not necessarily mean that the problem is too difficult. To be certain of that, one should show that subjects also err when required to judge a given conclusion.

The second observation concerns a general feature of Braine’s theory. The theory distinguishes between two general kinds of reasoning processes: direct and indirect. Direct processes are supposed to be easily available and almost automatically applied. Indirect processes are not universally available, all have a different cost, and incorporate strategies of reasoning that are not easily deployed even by the subjects who abstractly master them. Thus, for free production of conclusions, the theory will distinguish two main situations. If the conclusions can be obtained by application of direct reasoning processes alone, subjects should be able to generate them rather easily. However, if no conclusion can be generated by direct schemata, and indirect processes are necessary, then subjects should reach a conclusion only if it is easy to apply an easy indirect reasoning strategy. When the search space for what strategy to apply is too unconstrained, then they should tend to say that nothing follows, or answer incorrectly. In this latter case, however, when asked to judge a conclusion that can be obtained by application of an easy indirect strategy, performance should improve. Once again, from a “nothing follows” result alone, nothing follows.

Now, I direct attention back to the experiments. The first one showed that deductions with conditionals (in fact, just MP) are easier than deductions with exclusive disjunctions. In the experiments, subjects saw premises of the following form:

- If Linda is in Amsterdam, Cathy is in Majorca.
- Linda is in Amsterdam.

They had to freely generate conclusions by answering the question, “What follows?” In this case, subjects produced 91% correct conclusions. However, when they were given the following.

- Linda is in Amsterdam or Cathy is in Majorca, but not both.
- Linda is in Amsterdam.

What follows?

they produced only 41% correct conclusions.

The result is explained thus: A conditional requires only one initial model, whereas exclusive disjunction needs two; hence, the difficulty is increased.

I make three observations. First, for later use, notice that the theory seems to predict sharp changes of performance in an extremely fine way: A difference of just one model is sufficient to halve performance. Second, the experiment is based on a comparison between two forms of reasoning used in different contexts, and different in structure, frequency of occurrence, length, and naturalness. All of these factors, for which there is no control, may have an effect, and there is no reason to attribute the difference in complexity to mental models. Third, Braine’s theory would issue roughly the same prediction, although for different reasons. In both cases, only direct reasoning schemata are needed to reach a conclusion, so most subjects should reach one. However, the representation of an exclusive disjunction in logical form is more complex because it requires an extra clause, and although straight MP with a conditional requires a single application of a basic schema, an exclusive disjunction requires more steps. Thus, Braine’s theory predicts a difference precisely in the observed sense; the size of the difference, on the other hand, would probably be attributed to the factors mentioned earlier. However, I will show that the mental model explanation is hardly believable. Hence, to explain the size of the difference, also mental models will probably have to invoke other factors external to reasoning proper.

The second and third experiments concerned reasoning with conditionals and biconditionals. According to the model theory, MP from a conditional and from a biconditional should be equally easy; however, MT should be harder from a conditional than from a biconditional because it requires fleshing out, and although a fleshed-out conditional has three models, a biconditional requires only two. In both experiments, two clear results 2

2 In fact, it is not so clear that MT with a conditional requires three models. If subjects flesh out one model at the time, as it is reasonable to suppose, then two models may suffice if the second model subjects build is “A, B.” To get the three models, and thus the difference, one should suppose that this is never the case. But hypotheses on the order in which fleshing out occurs are not spelled out.
emerged: MP is easier than MT, and MP is not significantly more difficult from a conditional than from a biconditional. However, in the second experiment Johnson-Laird et al. (1992) failed to find the predicted interaction between type of deduction and type of conditional; in the third experiment, a replication of the second with more subjects, the predicted interaction occurred.

Braine’s theory does not contain a primitive connective for a biconditional, but by treating it as a conjunction of two conditionals one would expect the following: MP both from conditionals and biconditionals requires only direct reasoning procedures; thus, most subjects should reach a correct conclusion. Furthermore, a deduction from a biconditional involves only one very low-difficulty rule more than the corresponding deduction from a conditional: simplification of conjunction. Because in Braine’s system it is not the length of derivations per se that determines the difficulty of problems, but the degree of difficulty of the rules involved, such differences should not be significant. Thus MP should be easy, and negligibly more difficult from biconditionals than from conditionals. On the other hand, MT involves application of indirect strategies, notably reductio, and therefore it should be more difficult than MP. These predictions were verified in both experiments. The interaction is not predicted by Braine’s theory, but it presents problems for the mental model theory as well: There is no explanation of why, although in the previous experiment adding one model halved performance, in this case the same quantitative difference was barely detectable.

The final experiment tested the prediction that problems with exclusive disjunctions are easier than problems with inclusive disjunctions. As Johnson-Laird and Byrne (1991) put it, when reasoning with inclusive disjunctions, “people indeed appear to be overwhelmed by the possibilities” (p. 57). Both negative and positive double disjunctions were tested. For positive disjunctions, subjects were given inclusive premises of the form:

June is in Wales or Charles is in Scotland, or both.
Charles is in Scotland or Kate is in Ireland, or both.
What, if anything, follows?

and premises of the form

June is in Wales or Charles is in Scotland, but not both.
Charles is in Scotland or Kate is in Ireland, but not both.
What, if anything, follows?

In the first case, the model theory predicts that people should construct three models per premise, which can be integrated to yield five models. This should be too difficult a task. On the other hand, in the second case subjects should construct only two models per premise, which can be integrated in just two final models. Indeed, there were only 6% correct conclusions for inclusive disjunctions, and 21% for exclusive disjunctions.

Negative inclusive disjunctions were problems of the form:

June is in Wales or Charles is in Scotland, or both.
Charles is in England or Kate is in Ireland, or both.
What, if anything, follows?

and the respective exclusive case

June is in Wales or Charles is in Scotland, but not both.
Charles is in England or Kate is in Ireland, but not both.
What, if anything, follows?

One may wonder where the negation comes in, and the answer is that “of course, if Charles is in Scotland, then he is not in England” (Johnson-Laird et al., 1992, p. 433).

Negative disjunctions should be harder, according to model theory, because they “call for the detection of the inconsistency between the contrary constituents” (p. 433). Indeed, there were 8% correct conclusions for exclusive negative disjunctions, and 2% for inclusive negative disjunctions. In all cases, the most frequent answer was “nothing follows.”

One wonders why people should draw any correct conclusion from premises like these. The space search is too unconstrained for any correct response to be expected, but not because the problem is difficult. As this discussion suggests, errors or responses like “nothing follows” in free production tasks do not allow one to draw any general conclusion about reasoning abilities if there are no other controls that people do not answer or answer illogically because they follow normal, pragmatically justifiable response strategies. The right test would be to see whether people can judge the validity of given conclusions rather than produce their own. To see that this is a proper comment, consider the kinds of correct conclusions that subjects should have drawn. For inclusive disjunctions the authors suggested “June is in Wales, and Kate is in Ireland, or Charles is in Scotland, or both”; as a parallel example of a conclusion from a double exclusive disjunction, they offered “June is in Wales, and Kate is in Ireland, or Charles is in Scotland, but not both.”

However, these are barely less complex than the premises themselves, and there is no reason to suppose that subjects would go to such trouble to undertake a complex reasoning for no advantage. Thus, normal economy principles are sufficient to explain why most subjects should conclude that nothing follows.

Now look at what Braine’s theory would predict. First, by application of direct reasoning routines alone, both exclusive and inclusive premises lead to no conclusion; hence, “nothing follows” conclusions are predicted as frequent responses. As for the significant difference in correct responses between positive exclusive and positive inclusive disjunctions, notice that there are much more natural conclusions from exclusive disjunctions than the one suggested by Johnson-Laird et al.; for example:

If June is in Wales, then Kate is in Ireland.

(6)

or

If June is not in Wales, then Kate is not in Ireland.

(7)

A double exclusive disjunction can lead to conclusions such as (6) or (7) by repeated application of an indirect reasoning strategy—lemma generation—which is known to be significantly more difficult than direct reasoning strategies but relatively easy if compared with other indirect strategies (Braine et al., 1984, pp. 339–341). However, for a derivation starting with double inclusive disjunctions, even starting to prove lemmas would get nowhere. Other, more difficult strategies must be deployed, and even so, looking for a conclusion is more a blind search than a reasoning process. In such a case, no derivation can be directly reached if a conclusion does not indicate which assumptions to make or where to look for a derivation: People just do not know how to continue the story. Hence, the difference in correct responses between inclusive and exclusive disjunctions. In sum, Braine’s theory predicts both results about free production of
conclusions with double positive disjunctions: that they should elicit mostly "nothing follows" answers, and that, among the few correct answers, exclusive disjunctions should be easier than inclusive disjunctions. However, the phenomena have nothing to do with being "overwhelmed by the possibilities."

What about negative double disjunctions? According to Johnson-Laird and Byrne (1991), people might draw the following conclusions from an inclusive negative disjunction:

[June] is in [Wales] or [Kate] is in [Ireland],
or both, or equivalently,

If [June] is not in [Wales], then [Kate] is in [Ireland] (p. 58).

However, if people are really reasoning propositionally, then they are right to say that nothing follows: Propositionally, the suggested conclusions do not follow from the given premises. In fact, the authors exclude "those cases where [Charles] is in [Scotland] and [Charles] is in [England], because one person cannot be in two places at the same time" (Johnson-Laird & Byrne, 1991, p. 57). This makes the argument valid, but adds an extra, nonpropositional, premise. Thus, to get to any conclusion, extra premises and inferences are needed; in both cases, derivations will be longer and, especially, will involve more difficult reductio strategies. Thus, in both cases, Braine's theory predicts that fewer valid inferences should be drawn from negative than from positive disjunctions.

The Mental Models Theory Predicts the Wrong Results

I now argue for my second point: Even if these data were in favor of the mental model theory, there still would be reasons to doubt that models offer the right metric for reasoning complexity.

The full theory does not just speak of differences in problem solving abilities, but also of the size of these differences. If building models is costly, such a cost is to be quantifiable. Indeed, mental models venture to say that any reasoning involving more than three models will generate a floor effect. However, it is easy to find propositional problems that do not comply with that scale. Consider sentences including disjunctions. A simple conditional requires one model; however, if its antecedent (or consequent) is a disjunction even in its minimal interpretation, it will generate at least the number of models required by the disjunction in the antecedent (consequent) multiplied by the number of models in the consequent (antecedent). For example, the easy 'If A or B, then C' will give

A  C
B  C
...

However, if both antecedent and consequent happen to be disjunctive, the result will already overflow the maximum limit of models that working memory should be able to contain. A sentence such as 'If A or B, then C or D' will require at least four explicit models:

A  C
B  C
A  D
B  D
...

and a sentence such as 'If A or B, then C or D or E' will ask for six, and so on. Yet there should not be need of experiments to figure out that people can draw conclusions from two premises such as 'If A or B, then C or D or E; A or B'.

One can create infinite cases like these. Because for complex sentences negation is complementation, another class of cases will be created by conditionals with negative conjunctive antecedents or consequents. The theory will issue the wrong predictions for all of them. Just as the few cases presented count in favor of the theory, all these other cases should count against it.

How Mental Modelers Explain Braine's Results

I turn, finally, to the last support offered in favor of the model theory: It can explain Braine et al.'s (1984) results. I argue that if one takes such a claim seriously, then the mental model theory is self-refuting.

For Johnson-Laird et al. (1992) and Braine et al. (1984) "The most striking finding was that the rated difficulty of the problems was predicted by a regression equation based on two parameters: the length of the problem and the number of steps in its derivation" (p. 428). However, Braine et al. tested two equations defined over their theory. The first one assumed that each reasoning step has equal psychological cost and therefore that (besides other constant factors) derivation length alone should correlate with problem difficulty. The second one assumed that each reasoning step has a different psychological cost. In fact, the second equation is a much better predictor than the first one. When reasoning steps are weighted, as in the second equation, a .92 correlation can be obtained; when only numbers of steps are considered, as in the first equation, a .79 correlation obtains. The latter figure, however, is closer to what models can do: Johnson-Laird et al. (1992) found that models correlate .80 with problem difficulty in Braine et al.'s (1984) list when the presence of double negations is considered; brute models, on the other hand, account for 53% of variance.

The trouble is that once the proposed explanation is considered in the details, it looks more like a refutation than a confirmation of the model theory. Models were counted by using the invalid psychological algorithm, and this would be sufficient to discredit the result. However, there is another, more problematic aspect of the way the correlation is obtained. Whereas generally mental modelers explain their predictions on the basis of the final number of models of the integrated premises of a problem, in this case they compute the correlation on the sum of the models required by all of the premises. If the same criterion were used for other experiments adduced as evidence for mental models, the favorable evidence would disappear.

However, if one grants the exception and looks at what the correlation says, it turns out that very simple problems require very many models for the correlation to obtain. So in the same article, we find that

a. Correct conclusions drop from 91% for conditionals to 41% for exclusive disjunctions; one extra model is sufficient to halve performance;
b. Correct conclusions drop to 2% for a double negative disjunction, which requires five models, because "people appear to be overwhelmed by the possibilities";

However,
c. A problem in Braine et al.'s list that subjects solve easily and rate at a 3.86 difficulty over a scale of 9 requires seven (!) models.

Why are people not "overwhelmed by the possibilities" also in this case? To be consistent, Johnson-Laird et al. should predict that for Braine's problems, far from being good problem solvers, subjects massively fail to give correct answers.

Changing the way models are counted will not pull the theory out of the problem. If one looks at the correlation of problem difficulty with the number of models required by the integrated premises, as one should do to be consistent with the predictive strategy of model theorists, then models turn out to be a useless predictive factor. Almost all of the problems in Braine's list require either one or two composed models. A regression equation specifying only the final number of models as an independent variable is a very bad predictor of problem difficulty and accounts very poorly for variance ($r^2 = .034$; $F$ ratio of the regression: $2.066; p = .15$). A more complete equation, including double negations as an independent variable, is still a very poor predictor (multiple $r^2 = .154$), and models are a negligible factor in it. The only minimally reasonable equation includes total number of models, number of words, and double negations ($r^2 = .65$), but models have no predictive value in it at all (partial $r^2 = 0$). Finally, even if it is wrong, for this particular list of problems the psychological algorithm does give a correct idea of the integrated number of models that the rules of thumb for the full theory would require. Thus, even if one resorted to the full theory, predictions would be equally poor.

I conclude that either the authors were right in their experiments and in the way they usually count models, in which case they just have to drop the claim that they account for Braine's data, or they were right in accounting for Braine's data, in which case they have to drop their theory.

Conclusion

I have shown that the model theory of propositional reasoning is ill-defined; that the algorithms proposed to explain it better are either psychologically useless or untenable; that even if they were amended along envisionable lines, the mental model theory would issue the wrong predictions for an infinite class of cases; that Braine's theory accounts for the acceptable evidence offered; and that if one accepts Johnson-Laird et al.'s explanation of Braine's results, then the mental model theory is self-refuting.

Although my conclusions extend beyond the domain of propositional reasoning in that some of the mechanisms and explanatory strategies discussed are used by the general mental model theory, I do not want to deny that, in some domains, people represent knowledge with nonpropositional structures. Notably, this may be so for comprehension of spatial text (Bryant, Tversky, & Franklin, 1992; Johnson-Laird & Byrne, 1989; Ehrlich & Johnson-Laird, 1982; Franklin & Tversky, 1990; Mani & Johnson-Laird, 1982; Taylor & Tversky, 1992), but we only have a rough idea of what these internal structures are like, and propositional mental models do not help to provide a clearer understanding of them.

The idea that people reason with a mental logic has floated around for centuries, but its transformation into a psychologically plausible hypothesis is extremely recent. There is no reason to abandon it now—certainly, no reason coming from propositional mental models. We know very little about reasoning; it seems pointless to eliminate whole research programs, especially on the ground of ill-defined alternatives. It is not the right time for anybody to sing victory, probably not even the right century.

References


THEORETICAL NOTES

Appendix

Consider what the mental model of 'If A, then B and not B' would be. A first possibility is to construct its implicit model. This would give the following:

\[ A \quad B \quad \neg B. \]

Then, a procedure for eliminating contradictory models would erase the first model and leave only the second, implicit model. This will not do. Thus one has to suppose that, in this case, all models must be explicitly constructed. In this second case, one would again have different possibilities. One possibility is to suppose that the whole truth table is directly constructed, including the rows in which the sentence is false. Then subjects select the rows in which the sentence is true and construct the corresponding models. This would give the following:

\[
\begin{align*}
\neg A & \quad B \\
\neg A & \quad \neg B.
\end{align*}
\]

However, there is no reason why sentences including contradictions should behave in a completely different way with respect to all the other sentences. Another possibility is that subjects use synonyms of the given sentence. For example, the previous models are also the models of the sentence 'Either not A and B, or not A and not B, but not both.' However, the transformation would then require a mental logic. Another possibility is that subjects do not construct the conditional directly but resort to its definition in terms of Cartesian product and complementation. This makes it possible to get the correct results, but then one is obliged to assume that the AI algorithm correctly describes the full theory, with all the unwanted consequences for its explanatory power for psychological purposes.

Instructions to Authors

Authors should prepare manuscripts according to the Publication Manual of the American Psychological Association (3rd ed.). All manuscripts must include an abstract containing a maximum of 150 characters and spaces (which is approximately 120 words) typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the Manual. Also, all manuscripts are subject to editing for sexist language.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 6.24). As this journal is a primary journal that publishes original material only, APA policy prohibits as well publication of any manuscript that has already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data upon which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 6.25). APA expects authors submitting to this journal to adhere to these standards. Specifically, authors of manuscripts submitted to APA journals are expected to have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242.

Masked reviews are optional, and authors who wish masked reviews must specifically request them when submitting their manuscripts. Each copy of a manuscript to be masked reviewed should include a separate title page with authors' names and affiliations, and these should not appear anywhere else on the manuscript. Footnotes that identify the authors should be typed on a separate page. Authors should make every effort to see that the manuscript itself contains no clues to their identities.

Manuscripts should be submitted in quadruplicate. All copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typewriter is acceptable only if it is clear and legible. Dittoed and mimeographed copies will not be considered. In addition to addresses and phone numbers, authors should supply electronic mail addresses and fax numbers, if available, for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. Effective in January 1994, the Incoming Editor is receiving all submissions to the journal. Submissions that are accepted will be published beginning in the 1995 volume. Mail manuscripts to the Incoming Editor, Robert A. Bjork, Psychological Review; Department of Psychology, University of California, Los Angeles, California 90024-1563, according to the instructions provided above.

Received June 18, 1993
Revision received December 21, 1993
Accepted January 11, 1994