Topoi (2009) 28:75–80
DOI 10.1007/s11245-008-9049-9

The Mental Model Theory of Conditionals: A Reply to Guy Politzer

Philip N. Johnson-Laird · Ruth M. J. Byrne · Vittorio Girotto

Published online: 20 December 2008

Abstract This paper replies to Politzer's (2007) criticisms of the mental model theory of conditionals. It argues that the theory provides a correct account of negation of conditionals, that it does not provide a truth-functional account of their meaning, though it predicts that certain interpretations of conditionals yield acceptable versions of the 'paradoxes' of material implication, and that it postulates three main strategies for estimating the probabilities of conditionals.

Keywords Human reasoning · Conditionals · Mental models · Probabilistic reasoning

1 Introduction

The model theory accounts for deductive, inductive, and abductive reasoning. For deduction, it postulates that individuals use the meanings of propositions and general knowledge to construct mental models of the possibilities consistent with the premises (Johnson-Laird 2006). If a conclusion holds in all these possibilities then individuals infer that it is valid. They can show that an inference is invalid, not by searching in vain for its proof, but by constructing a counterexample, i.e., a model of a possibility consistent with the premises but not with the conclusion. Johnson-Laird and Byrne (2002) (henceforth JL&B) showed how the theory elucidates factual, deontic, and counterfactual conditionals (see also Byrne 2005). Politzer (2007), however, has criticized JL&B, and our aim in what follows is to outline enough of our theory to enable us to rebut his four main criticisms.

P. N. Johnson-Laird (✉)
Department of Psychology, Princeton University, Princeton, NJ 08540, USA
e-mail: phil@princeton.edu

R. M. J. Byrne
School of Psychology and Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland
e-mail: Ruth.Byrne@tcd.ie

V. Girotto
University IUAV Venice, Convento delle Terese, 30123 Venice, Italy
e-mail: vgirotto@iuav.it

2 The Model Theory of Conditionals

A conditional such as *if she played a game then she didn't play music* is consistent with three possibilities:

She played a game.	She didn't play music.
She didn't play a game.	She didn't play music.
She didn't play a game.	She played music.

It rules out the remaining case as impossible. These three possibilities are the 'core' meaning of conditionals. Barrouillet et al. (2000) showed that adults list such possibilities.

Mental models follow a principle of 'truth': they do not represent the possibilities consistent with assertions in a fully explicit way, but instead represent clauses, whether affirmative or negative, only when they are true in a possibility. The conditional above has mental models that represent explicitly only the possibility in which the antecedent (the if-clause) and the consequent (the then-clause) are true:

She played a game.	She didn't play music.
	. . .

The ellipsis denotes a model that has no explicit content, but that allows for the possibilities in which the antecedent is false. When a task is not demanding, individuals can flesh out mental models into fully explicit models.

Consider a modus ponens deduction, e.g.:

If she played a game then she didn't play music.
She played a game.
Therefore, she didn't play music.

The deduction can be made by forming a conjunction of the mental models of the conditional with the mental model of the categorical premise. This process eliminates the implicit model to leave only a single model:

She played a game. She didn't play music.

which yields the conclusion *she didn't play music*.

A conclusion follows validly provided that it holds in all the possibilities consistent with the premises. Conversely, the premises of some valid deductions hold in all the possibilities consistent with their conclusions. But, other valid deductions have premises that are more informative than their conclusions, e.g.:

She played a game.
Therefore, she played a game or she played music, or both.

The inference is valid, but the premise does not hold in a possibility consistent with the conclusion—the possibility that she didn't play a game and she played music. Individuals balk at such inferences, which throw away information by adding disjunctive alternatives to the possibilities consistent with the premises (JL&B, p. 652).

The principle of truth reduces the load on working memory, but, as a computer program implementing the theory predicted, it has a devastating effect on some inferences. Consider this problem about the cards in a hand:

If there is a king then there is an ace, or else if there isn't a king then there is an ace.
There is a king.
What follows of necessity?

As the mental models predict, nearly everyone infers that there is an ace, but the inference is invalid. One of the conditionals could be false, and the first conditional could be false because there is a king but no ace. Such invalid inferences occur with other sentential connectives, and they occur in all domains of reasoning (Johnson-Laird 2006). Psychological theories based on formal rules of inference have yet to explain, let alone predict, them.

Conditionals have different sorts of meaning. For instance, we can add to the conditional: *if she played a game then she didn't play music*, an explicit rider: *and if she didn't play a game then she may or may not have played music*. But, we can add to the similar conditional: *if she played a game then she didn't play soccer*, only the rider: *and if she didn't play a game then she didn't play soccer*. In fact, conditionals have indefinitely many sorts of meaning. What gives rise to them are the interactions among several simple components, and we now describe two of them from JL&B.

2.1 Syntax

The antecedent of a conditional is a subordinate clause and the consequent is a main clause, whereas 'and' and 'or' can interconnect two main clauses. Sentential operators are often interpreted as though they applied only to main clauses (JL&B p. 650). For example, the assertion:

It is odd that if he served tea then he wore gloves

is taken to mean:

If he served tea then it is odd that he wore gloves.

This pecularity of interpretation is frequent, but strictly speaking erroneous (as we'll see). It is not unique to conditionals, but applies to other sentences with subordinate clauses, e.g., *it is odd that before he served tea he wore gloves*.

2.2 The Modulation of the Core Meaning

The model theory postulates that a conditional is 'basic' if there is no relation between its antecedent and consequent other than their co-occurrence in the same conditional. Such conditionals have the core meaning illustrated earlier. However, other conditionals have a relation between the situations that their clauses describe, because of the meanings of the clauses, co-reference between them, or knowledge. These factors can *modulate* the core meaning in two different ways (JL&B).

On the one hand, modulation can block the construction of a model of a possibility. For instance, the earlier example of *if she played a game then she didn't play soccer* is consistent with only two possibilities because soccer is a game, and so given that she didn't play a game, she can't have played soccer. Modulation can block possibilities, and it yields various sets of possibilities for conditionals (see Table 4 JL&B).

On the other hand, modulation can add temporal, spatial, and other relations between the antecedent and consequent (JL&B, abstract, p. 658, and p. 673). Consider, for instance, this conditional:

If she fell off her bike then she grazed her knee.

It elicits a possibility in which there is a temporal order: she fell off her bike and then grazed her knee. Other cases of modulation elicit more complicated scenarios.

3 A Reply to Politzer

Politzer gives an excellent account of the main psychological theories of conditional reasoning. Some of his criticisms of the model theory are simple to refute. He suggests (p. 86), for example, that the theory doesn't predict the relative difficulties of simple inferences from conditionals. But, as Schroyens et al. (2001) and Oberauer (2006) have shown, it does better than its rivals in Politzer's review. His other criticisms are subtler, and we now attempt to rebut each of them in turn.

3.1 Negation

When individuals are asked to negate a conditional, *If A then C*, they often respond: *If A then not C*. Politzer (p. 86) argues that the model theory fails to predict the response. It is not a matter dealt with in JL&B, but the tendency for sentential operators to be treated as though they apply only to main clauses (see Sect. 2.1) predicts the response. An assertion of the form:

It is not the case that if A then C

tends to be interpreted as:

If A then it is not the case that C.

A proposition and its negation ought to allow no possibilities in common if they are to contradict one another. But, *not-A and not-C* is consistent with both *if A then C* and *if A then not-C*. Hence, individuals have not made the correct negation of a conditional. Their response shows that it is wrong to treat sentential operators applied to conditionals as though they applied only to their consequents. To reveal the correct negation, a better task is to ask individuals to list what is impossible given *If A then C*. They tend to respond: *A and not-C*.

3.2 Truth Functionality

A truth function takes truth values as its input and yields a truth value as its output, and so a connective is truth functional if the truth value of a proposition formed with it depends solely on the truth values of the propositions that it connects. In sentential logic, connectives are truth functional. But, according to the model theory, no connectives in natural language are truth functional, and JL&B (p. 673) assert explicitly that conditionals are not truth functional. Politzer recognizes this point, but he argues (p. 87):

> But they seem to have a special notion of truth-functionality, which makes their claim ambiguous…Here lies the ambiguity for, after modulation, the various interpretations remain defined extensionally. In brief, the authors are right that their conditionals are not (purely) truth-functional, but the semantics which they adopt is nevertheless extensional.

We take him to mean that a semantics in which conditionals refer to sets of possibilities is extensional, and that each such set can be described in a truth-functional way. But, what he and others (e.g., Evans and Over 2004, p. 21) overlook is that modulation can introduce temporal, spatial, and other relations, between the antecedent and the consequent (as stated in JL&B abstract, p. 658, and p. 673). For example, the earlier conditional *if she fell off her bike then she grazed her knee* elicits a temporal interpretation, so that the mere truth of its antecedent and consequent does not guarantee the truth of the conditional: if she grazed her knee and then fell off her bike, the conditional is false. So, the conditional is not truth functional, and the process of interpreting conditionals cannot be, either: it constructs possibilities, not truth values, and it is sensitive to temporal and other relations that are not truth functional (for a program implementing temporal reasoning with models, see Schaeken et al. 1996).

3.3 The 'Paradoxes' of Material Implication

The three fully explicit possibilities of the core meaning of a conditional (see Sect. 2) correspond to the three 'true' entries in the truth table for material implication in the sentential calculus. Hence, the following two inferences are valid for the core meaning:

> She didn't play a game.
> Therefore, if she played a game then she didn't play music.

and:

> She didn't play music.
> Therefore, if she played a game then she didn't play music.

Inferences of this sort are known as the 'paradoxes' of material implication. Some theorists (e.g., Evans and Over 2004, p. 19) argue that the paradoxes are so absurd that any psychological theory permitting them should be rejected.

In contrast, the model theory implies that the paradoxes are valid for some but not all interpretations of conditionals, it explains their counterintuitive nature, and it predicts that they should be acceptable in some cases (JL&B pp. 651–652). They are counterintuitive because they throw away semantic information by adding a disjunctive

alternative to the possibilities consistent with their premises (p. 652; as illustrated in the inference from *she played a game* in Sect. 2). Politzer rejects this account. He argues that the factor of throwing information away is neither necessary nor sufficient to yield oddness. He writes (pp. 86–87):

> To see, for instance, that it is not sufficient, just consider that except when the 'possibilities' in the conclusion of a deductive inference are the same as those in the premises (that is, when the inference expresses an identity) there are always more 'possibilities' in the conclusion of a valid inference (because the 'possibilities' are logical models). Luckily, not all valid inferences are odd.

Here, he may have a different conception of 'possibilities' than JL&B. The conclusions of many valid inferences that are not identities do not introduce disjunctive alternatives. As we showed in Sect. 1, modus ponens is such an inference: it is valid, it is not an identity, and its conclusion holds in the one possibility consistent with its premises. The crux is simple: if the conclusion of a valid inference throws information away by adding a disjunctive alternative to the possibilities consistent with the premises, then it should seem odd, and individuals should balk at it. The challenge to critics is to find a counterexample.

Certain cases of the paradoxes are valid, and so it should be possible to explain their validity to skeptics. The following dialog between an experimenter (E) and a participant (P) may help:

E: Please write down the possibilities consistent with: if she played a game then it wasn't snowing.

P writes (as in Barrouillet et al. 2000):

> She played a game and it wasn't snowing.
> She didn't play a game and it wasn't snowing.
> She didn't play a game and it was snowing.

E: We can describe these three possibilities in this assertion: It wasn't snowing, or, if it was, then she didn't play a game.

P: Correct.

E: If it's actually true that it wasn't snowing, would this assertion be true: it wasn't snowing, or, if it was, then she didn't play a game?

P: Yes.

E: So, if we replace my paraphrase with the original conditional the same relation holds: given that it wasn't snowing, then it's true that if she played a game then it wasn't snowing.

P: Yes.

In what cases, should individuals accept a 'paradox' immediately and without the need for an explanatory dialog? The model theory answers with a clear prediction: when the conclusion does not add a disjunctive alternative in which the premise fails to hold. Here is an example:

> Viv didn't play soccer.
> Therefore, if she played a game then she didn't play soccer.

The inference does not seem absurd; and its premise holds in both the possibilities consistent with the conclusion:

She played a game.	She didn't play soccer.
She didn't play a game.	She didn't play soccer.

Politzer judges as 'rather weak' a second argument that JL&B made about the cause of the paradoxes. We won't dispute this point here: our present account suffices. He concludes (p. 87): 'JL&B do not have a satisfactory explanation of the paradoxes of implication, which suggests that the conditional of ordinary language is not captured by the core meaning assumed by the authors'. In the light of our preceding analysis, we re-iterate three points: (1) the paradoxes are valid for certain meanings of conditionals, (2) their paradoxical nature arises from the fact that their conclusions throw away information by adding a disjunctive alternative to those in which the premise is true, and (3) inferences of the same syntactic form as the paradoxes seem more acceptable when their conclusions do not throw information away in this manner. We conclude that the model theory does have a satisfactory explanation of the paradoxes. Hence, if it fails on other counts then it does at least have a satisfactory explanation of them. And so its core meaning for basic conditionals in everyday language is correct.

3.4 The Probability of Conditionals

What is the probability of a conditional, *if A then C*? Politzer argues that other current theories predict that individuals should assess it as the conditional probability of *C given A* (e.g., Evans and Over 2004). In contrast, he claims (p. 88), the model theory predicts it should equal the probability of material implication for individuals who construct fully explicit models of the conditional, but a different value for those who construct mental models, which contain only an implicit model for cases in which the antecedent is false. This account is only a part of the model theory of the probability of conditionals (see Girotto and Johnson-Laird 2004).

The extension of an assertion is the set of possibilities to which it refers, and, according to the model theory, individuals estimate the probability of a proposition in an extensional way according to this procedure: they construct models of the prior possibilities, they assume that they are equiprobable unless they have evidence to the contrary, and they estimate the probability of the proposition as the

proportion of these possibilities in which the proposition holds (Johnson-Laird et al. 1999). In all but the simplest of cases, naïve individuals do not know at first how to carry out the task. They have to devise a *strategy* for coping with it, and different individuals devise different strategies, just as they do in deductive reasoning (see Van der Henst et al. 2002). The theory proposes three main strategies that individuals are likely to adopt (Girotto and Johnson-Laird 2004), and we illustrate them with an example:

> There are three cards face down on a table: a three, a six and an eight. Pat takes one card at random, and then he takes another at random. Viv says:
>
> If Pat has the eight then he also has the three.
> What is the probability that Viv's assertion is true?

The *equiprobable* strategy relies only on the explicit mental model of the conditional (in which both its antecedent and consequent are true), and compares this case with one in which the conditional is false because its antecedent is true but its consequent is false. This strategy yields the probability of the conditional as ½, which for this conditional, but not others, is the same as the conditional probability of the consequent given the antecedent, i.e., the estimate that the 'conditional probability' hypothesis predicts.

The *conjunctive* strategy also relies on the explicit mental model in which the antecedent and consequent are both true (8 3), but assesses it in relation to the three possible hands that Pat could have:

8 3
6 3
8 6

Hence, individuals treat the conditional as though it were a conjunction—a tendency that has been observed independently (Johnson-Laird et al. 1999). The strategy yields a probability for the conditional of 1/3.

The *complete* strategy treats the conditional as true in all the possibilities that are consistent with it, and false in the one possibility that is inconsistent with it. It yields a probability of 2/3 for the conditional, because only one possible hand out of the three violates the conditional. It is this strategy that is equivalent to treating the conditional as though it were a material implication.

Girotto and Johnson-Laird corroborated the occurrence of the three strategies, but not the conditional-probability hypothesis. Politzer (p. 88) finds our interpretation of these results 'highly debatable'. He doesn't explain why, and so we turn to our misgivings about the conditional probability hypothesis itself. They are threefold.

First, when participants thought aloud, their protocols showed that they transformed:

> What is the probability of if 8 then 3?

into:

> If 8 then what is the probability of 3?

Politzer accepts this evidence, and that the resulting question calls for an estimate of the conditional probability of *3 given 8*. But, he says, 'this [transformation] can be taken as evidence in favour of the conditional probability hypothesis'. Others make the same argument (Over et al. 2006). The transformation certainly occurs, but we have already described how it leads to an erroneous interpretation of negation. There is no reason to suppose that it is any more accurate for estimates of the probability of a conditional. Indeed, individuals are able to paraphrase a basic conditional:

> If 8 then 3.

as:

> Not-8 or 3 (see Richardson and Ormerod 1997).

Yet, no-one is likely to suppose that the probability of the disjunction (2/3) is equal to the conditional probability of *3 given 8* (1/2).

Second, Schroyens et al. (2008) have shown that when a prior task increased the relevance of the possibilities in which the antecedent was false, there was a reliable increase in *complete* estimates of the probability of a conditional. The conditional probability hypothesis offers no obvious explanation for this phenomenon.

Third, conditional probabilities are on the border of naïve individuals' competence: they don't really understand them. For example, in a recent study, participants were presented with a series of conditional probabilities, and asked to estimate the converse conditional probabilities (see Johnson-Laird 2006, p. 203). In fact, the probability of *C given A* tells one almost nothing about the probability of *A given C*. Yet, the participants made a series of these converse inferences and seldom balked at any of them.

4 Conclusion

We are grateful to Politzer for his description and criticisms of the model theory, and for the opportunity to clarify the JL&B theory and to draw attention to aspects of it often overlooked. We have argued that it withstands his four major criticisms. First, its account of the negation of conditionals is correct, and it predicts a common mistake that naïve individuals make in trying to negate conditionals. Second, its account of conditionals is not truth-functional, not even in Politzer's extended sense, because it postulates that among the modulations of the core meaning

of a conditional are cases in which temporal, spatial, and other relations are established between the antecedent and consequent situations. Third, it allows that for some interpretations of conditionals, but not all, the 'paradoxes' of material implication are valid. They seem counterintuitive because their conclusions throw away information by adding a disjunctive alternative to those in which the premise is true. And, perhaps surprisingly, when their conclusions do not throw information away in this manner, they seem to be acceptable inferences. Indeed, readers may not have noticed our use of such a 'paradoxical' inference at the end of Sect. 3.3. Fourth, the model theory proposes three main strategies for extensional estimates of the probability of basic conditionals. Only one of them is an estimate based on the three fully explicit possibilities of the core interpretation of conditionals. Once allowance is made for an inappropriate transformation into questions seeking a conditional probability, the evidence supports this account. Politzer concludes this section of his critique with a call for more research. On this final point we concur.

References

Barrouillet P, Grosset N, Leças J-F (2000) Conditional reasoning by mental models: chronometric and developmental evidence. Cognition 75:237–266

Byrne RMJ (2005) The rational imagination: how people create alternatives to reality. MIT Press, Cambridge

Evans JStBT, Over DE (2004) If. Oxford University Press, Oxford

Girotto V, Johnson-Laird PN (2004) The probability of conditionals. Psychologia 47:207–225

Johnson-Laird PN (2006) How we reason. Oxford University Press, Oxford

Johnson-Laird PN, Byrne RMJ (2002) Conditionals: a theory of meaning, pragmatics, and inference. Psychol Rev 109:646–678

Johnson-Laird PN, Legrenzi P, Girotto V, Legrenzi M, Caverni JP (1999) Naive probability: a model theory of extensional reasoning. Psychol Rev 106:62–88

Oberauer K (2006) Reasoning with conditionals: a test of formal models of four theories. Cogn Psychol 53:238–283

Over DE, Hadjichristidis C, Evans JStBT, Handley SJ, Sloman SA (2006) The probability of causal conditionals. Cogn Psychol 54:62–97

Politzer G (2007) Reasoning with conditionals. Topoi 26:76–95

Richardson J, Ormerod TC (1997) Rephrasing between disjunctives and conditionals: mental models and the effects of thematic content. Quart J Exp Psychol 50A:358–385

Schaeken W, Johnson-Laird PN, d'Ydewalle G (1996) Mental models and temporal reasoning. Cognition 60:205–234

Schroyens W, Schaeken W, d'Ydewalle G (2001) The processing of negations in conditional reasoning: a meta-analytic case study in mental model and/or mental logic theory. Think Reason 7:121–172

Schroyens W, Schaeken W, Dieussaert K (2008) 'The' interpretation(s) of conditionals. Exp Psychol 55:113–120

Van der Henst J-B, Yang Y, Johnson-Laird PN (2002) Strategies in sentential reasoning. Cogn Sci 26:425–468