Counterfactual Reasoning: Inferences from Hypothetical Conditionals

Ruth M.J. Byrne
Psychology Department,
University of Dublin, Trinity College,
Dublin 2, IRELAND
rmbyrne@vax1.tcd.ie

Alessandra Tasso
Psychology Department,
University of Padua,
ITALY
psico06@ipdunivx.unipd.it

Abstract

Hypothetical reasoning -- thinking about what might happen in the future or what might have happened in the past -- enables us to go beyond factual reality. We suggest that human reasoners construct a more explicit mental representation of hypothetical conditionals, such as, *If Linda were in Dublin then Cathy would be in Galway*, than of factual conditionals, such as, *if Linda is in Dublin then Cathy is in Galway*. When people think about the factual conditional, they keep in mind the affirmative situation -- Linda is in Dublin, Cathy is in Galway, and they maintain only an implicit awareness that there may be alternatives to this situation. In contrast, when they think about the hypothetical conditional, they keep in mind not only the affirmative situation, but also the presupposed negative one (Linda is not in Dublin, Cathy is not in Galway). The postulated differences in mental representations lead us to expect differences in the frequency of inferences that people make from the two sorts of conditionals, and we report the results of an experiment that corroborates this prediction. The psychological data have implications for philosophical and linguistic accounts of counterfactual conditionals, and for artificial intelligence programs designed to reason hypothetically.

Hypothetical Conditionals

We stretch our imaginations most, outside of daydreams, when we think about hypothetical possibilities. This capacity ensures that we are not mentally tied to the situation we find ourselves in and we can think about other situations that differ from it. We engage in such hypothetical thinking when we mull over the past, wondering what might have been, e.g.,

1. If John had worn a seat belt, his injuries would have been negligible.

We also do so when we look to the future, thinking about what might be, e.g.,

2. If the ozone layer were replenished, there would be a decrease in the incidence of cancers.

The sorts of hypothetical possibilities that we can think about span a range from situations that are close to the actual course of events in the world's history or potential future, such as 1 and 2, to possibilities that are quite remote, e.g.,

3. If kangaroos had no tails, they would topple over.

(from Lewis, 1973). Hypothetical thinking facilitates political and legal debate, scientific and causal thought, predictions about everyday social and personal activities, learning from mistakes, and experiencing emotions such as surprise and regret. How do people do it? Our attempt to answer this question takes as its starting point an examination of the cognitive processes underlying people's understanding of the logic of hypothetical conditionals.

The Problem of Counterfactuals

Most research on deduction has focused on reasoning with conditionals that are at the factual end of the dimension of hypotheticality, e.g.,

4. If Jane inherited the fortune, she bought the sports car of her dreams.

There is a large body of empirical data on the nature of the inferences that human reasoners make from such conditionals. Reasoners find some sorts of inferences easy and others difficult, and the factors that influence their prowess are many (for a review, see Evans, Newstead, and Byrne, 1993). A primary finding, to which we will return shortly, is that reasoners find some valid inferences from conditionals much more difficult to make than others, and they endorse inferences that, on a strictly logical analysis, they should consider fallacious.

Perhaps hypothetical conditionals can be treated in the same way as factual conditionals? The challenge of extending a general theory of conditionals to encompass hypothetical conditionals (about possibilities) and counterfactual conditionals (about matters which are impossible, or which were once possible but are so no

longer) has been taken up in philosophy, linguistics and artificial intelligence (e.g., Barwise, 1986; Ginsberg, 1986; Jackson, 1991; Pollock, 1986), and recently, in psychology (e.g., Braine and O'Brien, 1991; Johnson-Laird and Byrne, 1991). But, conditionals at the other extreme of hypotheticality from factual conditionals, such as counterfactuals, seem to mean something different. Compare the factual conditional in 4 with its corresponding counterfactual in 5:

5. If Jane had inherited the fortune, she would have bought the sports car of her dreams.

The indicative mood of 4 is a clue to its relative factuality compared to the subjunctive mood of 5 but it is a moot point whether mood and factuality coincide perfectly (e.g., Dudman, 1988). A counterfactual such as 5 carries a presupposition that its antecedent (the first part of the conditional) and its consequent (the second part) are both false. How then is a counterfactual to be evaluated as true or false?

The conditions under which a factual conditional is true can be considered to be a function of the truth of its components. The factual conditional in 4 is true in the situation where Jane inherits the money and buys the car, and false in the situation where she inherits the money but does not buy the car. It is true in two further situations, those in which Jane does not inherit the money, and in these cases, she may buy the car or not. Everyday conditionals can depart from this purely truth-functional account, perhaps because pragmatic factors overlay their meaning (e.g., Grice, 1975), and theories of conditionals need to encompass these vagaries of interpretation (e.g., Johnson-Laird and Byrne, 1991). However, counterfactuals do not yield readily to a similar truth-functional account. The counterfactual in 5 rules out the two possibilities where Jane inherited the money, since it implies that the antecedent is false. On a purely truth-functional account then, the counterfactual is true, but so too is the contrary counterfactual,

6. If Jane had inherited the fortune, she would not have bought the sports car of her dreams.

Clearly, people consider some counterfactuals to be true and others to be false. If we are to account for their understanding of them, we must go beyond an analysis of the truth of the components.

In each of the hypothetical conditionals above, we must suspend our disbelief in the antecedent, a process which requires us to suppose the truth of something we know is currently false or perhaps even impossible, and we must nonetheless maintain consistency between this supposition and the rest of our beliefs. In artificial intelligence research, adding contradictory facts to a database raises non-trivial problems, and yet it is crucial for a system that engages in planning, diagnosis, or the creation of sub-goals to solve problems (e.g., Ginsberg, 1986).

One possibility is that a counterfactual is true if the consequent follows from the antecedent taken together with any relevant premises (e.g., Chisholm, 1946; Goodman, 1973; cf. Kvart, 1986; and for a psychological adaptation, see Braine and O'Brien, 1991), and the problem then is to specify the set of relevant premises. We consider instead that a counterfactual is true if the consequent is true in the models or scenarios constructed by adding the false antecedent to the set of beliefs it recruits about the actual world, and making any necessary adjustments (e.g., Lewis, 1973; Ramsey, 1931; Stalnaker, 1968; and for a psychological adaptation, see Johnson-Laird, 1986; Johnson-Laird and Byrne, 1991). The problem then is to specify the most similar, minimally changed scenario in which to evaluate the consequent. Empirical evidence on the minimal changes people make to mutate a scenario when they wish to mentally undo an outcome suggest the process is influenced by a range of factors, including for example, the temporal order of events, their causal relations, and their exceptionality (e.g., Kahneman and Miller, 1986).

However, there are no psychological studies of the everyday logic of hypothetical conditionals. Psychological analyses have concentrated on assessments of their plausibility (e.g., Miyamoto and Dibble, 1986), and on cross-cultural psycholinguistics (e.g., Kit-Fong Au, 1983). We have carried out a series of experiments to examine the psychological processes underlying people's understanding of the logic of counterfactuals, and in this paper we will consider the results of one of them (see Byrne and Tasso, 1994).

A Sketch of a Theory of Hypothetical Conditionals

In this paper, we attempt to specify the nature of the mental representations that people construct of hypothetical conditionals, and the inferences they make from them, in comparison to factual conditionals. We will sketch a model-based theory of hypothetical reasoning (see Byrne and Tasso, 1994, Johnson-Laird and Byrne, 1991). First, we will outline briefly the model theory of the representation of factual conditionals, and then we will show that hypothetical conditionals, including counterfactuals, can be encompassed in this general theory.

The Representation of Factual Conditionals

The factual conditional,

7. If Linda is in Dublin then Cathy is in Galway.

is consistent with three separate situations, that capture the way the world would be if the conditional were true, which we represent in the following diagram:

8. L C
not-L not-C
not-L C

where "L" represents "Linda is in Dublin", "C" represents "Cathy is in Galway", and "not-L" is a propositional-like tag to represent that Linda is not in Dublin (see Johnson-Laird and Byrne, 1991). Separate models are represented on separate lines, so for example, the first model corresponds to the situation where Linda is in Dublin and Cathy is in Galway. The models may be filled with information about who Linda and Cathy are, where they are usually located, and what the connection between their relative locations is, but these details are not our immediate concern here; the structure of the models is our focus.

We believe that reasoners construct an initial representation that is more economical than this fully fleshed-out set, as illustrated in the following diagram:

9. L C
...

where the three dots represent a model with no explicit content. It may be "fleshed-out" to be explicit, to the three situations above if necessary, and it rules out a conjunctive interpretation. The idea is that reasoners represent explicitly the case mentioned in the conditional, and they keep track of the possibility that there may be alternatives to it.

In fact, the initial representation must record that Linda being in Dublin has been represented *exhaustively* with respect to Cathy being in Galway, i.e., it can occur again in the fleshed-out set only with Cathy in Galway. We capture this notion in our diagrams with square brackets -- a 'mental note' that an individual makes when representing the conditional:

10. [L] C
...

Cathy's being in Galway has not been exhaustively represented and so it may be included in other models without Linda being in Dublin. The model illustrated in the diagram in 10 is the initial economical model that we believe reasoners construct. We have corroborated the theory experimentally (Byrne and Johnson-Laird, 1992; Johnson-Laird, Byrne, and Schaeken, 1992) and modelled it computationally (see Johnson-Laird and Byrne, 1991, for details of the principles by which the psychological algorithm processes models).

The Representation of Hypothetical Conditionals

When reasoners understand a conditional that is hypothetical,

11. If Linda were in Dublin then Cathy would be in Galway.

they engage in a similar process to that for a factual conditional. They construct an economical mental representation based on the information given to them in the premises,

12. hypothetical [L] C
...

They keep track of the epistemic status of their models, making a 'mental note' about whether the models correspond to actual or hypothetical situations, and they tag the models accordingly. The representation of the hypothetical situation will recruit memories that provide further information about the belief reasoners have in the actual status of the antecedent, the status of the consequent, and the connection between them. They will construct models of a different structure in each of these cases (for details see Byrne and Tasso, 1994). Hence, they also represent the actual situation, in so far as they know it, or can induce it from the cues of the mood of the conditional,

13. actual not-L [not-C]
hypothetical [L] C
...

In summary, our suggestion is that the mental representations and processes for factual and hypothetical conditionals differ in how much is made explicit initially.

Inferences from Factual and Hypothetical Conditionals

Inferences based on an initial representation are easier than inferences that require reasoners to flesh-out models (see Johnson-Laird, Byrne, and Schaeken, 1992), as shown in several deductive domains (Byrne 1989a; 1989b; Byrne and Johnson-Laird, 1989; Johnson-Laird and Byrne, 1989; Johnson-Laird, Byrne, and Tabossi, 1989). This observation leads us to expect that certain inferences will be made more readily from a hypothetical conditional than from a factual one, namely, those inferences that require the representation of the negative instance (Linda is not in

Dublin, Cathy is not in Galway). To illustrate this point, we will compare the inferential process from hypothetical and factual conditionals for two sorts of inferences, one that does not require the representation of the negative instance, and one that does.

The premises of the modus ponens inference, that is, the factual conditional in 7 and the minor premise,

14. Linda is in Dublin.

require subjects to construct an initial model of the first premise as illustrated in the diagram in 10, and a model of the second premise:

15. L

The two sets of models can be combined, and the combination eliminates the implicit model, and leaves the first model only:

16. L C

from which it can be concluded that,

17. Therefore, Cathy is in Galway.

A similar process is required for the same inference from the hypothetical conditional. For the hypothetical conditional in 11 and the minor premise in 14, reasoners construct an initial model of the premises, as illustrated in the diagram in 13, and a model of the second premise as illustrated in the diagram in 15. The combination results in one model, whose status is updated to represent an actual situation:

18. actual: L C

and this model supports the conclusion in 17.

Consider now a more difficult inference, modus tollens, from a factual conditional. Given the factual conditional in 7 and the minor premise,

19. Cathy is not in Galway.

reasoners once again construct an initial model of the first premise, as illustrated in 10, and a model of the second premise,

20. not-C

They try to combine them, but this time, the combination fails since the two sets of models seem incompatible. The most common error that reasoners make to this inference is to conclude that nothing follows. A prudent reasoner fleshes out the models to the full set illustrated in 8, and then the combination of the model of the second premise rules out all but the second model:

21. not-L not-C

The valid conclusion can be made,

22. Therefore, Linda is not in Dublin.

The modus tollens inference is difficult, according to the model theory, because reasoners must flesh out their models and keep many alternatives in mind in order to make the deduction.

In contrast, we expect that the modus tollens inference should be easy to make from the hypothetical conditional. To represent the hypothetical conditional in 11 and the minor premise in 19, reasoners first construct an initial model of the premises, as illustrated in 13, and they combine it with the model for the second premise, illustrated in 20. This time the models can be combined directly, with no need to flesh them out. Reasoners can eliminate the hypothetical models, and retain the first model only:

23. actual: not-L not-C

which supports the valid conclusion in 22.

We predict that modus tollens from a hypothetical conditional will be easier than modus tollens from a factual conditional because the initial representation of the hypothetical conditional is more explicit than the one from the factual conditional. As a result, modus tollens can be made directly without any need to flesh-out the set of models.

We make a similar set of predictions for two further inferences, the denial of the antecedent from the minor premise,

24. Linda is not in Dublin.

and the affirmation of the consequent, from the minor premise,

25. Cathy is in Galway.

These inferences are fallacies on a conditional interpretation and reasoners make them only when they fail to flesh out their models fully. We again predict that the inference that depends on a negative instance -- the denial of the antecedent -- will be made more readily from the hypothetical than the factual conditional. In this case the inference is fallacious, and so our expectation is that the hypothetical conditional will support more fallacies than the factual conditional (see Byrne and Tasso, 1994, for details).

An Experimental Comparison of Inferences

Our aim was to compare the frequency of four sorts of inferences from hypothetical and factual conditionals. We predicted that the inferences that depended on an awareness of the negative instance -- modus tollens and denial of the antecedent -- would be made more often from the hypothetical conditional than from the factual one, because of the postulated differences in the explicitness of their representations.

Method

In the experiment we asked 80 people, untrained in logic, to make a single inference from a conditional. The design of the experiment was a fully between-subjects one. We gave half of the subjects the factual conditional and we gave the other half the hypothetical conditional (see Byrne and Tasso, 1994). Each subject carried out one of the four inferences outlined earlier. The content of the conditionals was based on people-in-places, expressed in the present tense, as illustrated in the examples. The components were negated explicitly where necessary (e.g., Linda is not in Dublin).

The subjects were undergraduate students in Trinity College, Dublin, who participated in the experiment voluntarily. They were tested in several medium-sized groups and they were randomly assigned to one of the eight conditions (10 subjects in each condition). They were given the premises printed on a sheet of paper, and their task was to write down what conclusion, if any, followed from them.

Results

The data corroborated our predictions, as Table 1 shows. As we expected, the subjects made more modus tollens inferences (80%) from the hypothetical conditional than from the factual one (40%), and they made more denial of the antecedent fallacies (80%) from the hypothetical conditional than from the factual one (40%), and both of these differences are reliable [Meddis quick-test (Meddis, 1984) $z = 1.77$, $n = 20$, $p < 0.05$, for each].

As Table 1 also shows, there were no reliable differences in the frequency of modus ponens inferences from the hypothetical conditional (90%) compared to the factual one (100%), nor in affirmation of the consequent fallacies from the hypothetical conditional (50%) compared to the factual one (30%, Meddis quick-test $z = 1$, and $z = 0.89$, respectively, $n = 20$, p = non-significant for each). These results corroborate our proposals about the representation of information in hypothetical conditionals. People represent both the negative case and the affirmative case for a hypothetical conditional, whereas they represent just the affirmative case for a factual conditional.

Table 1: Percentages of inferences made in each condition of the experiment

	MP	MT	DA	AC
Factual	100	40	40	30
Hypothetical	90	80	80	50

Key: MP, modus ponens; MT, modus tollens; DA, deny antecedent; AC, affirm consequent.

Discussion

The experimental evidence supports our suggestion that reasoners construct an initial representation of hypothetical conditionals that is more explicit than the initial representation of factual conditionals. For a hypothetical conditional they keep in mind not only the affirmative instances mentioned in the conditional (e.g., Linda in Dublin, Cathy in Galway) but also the implied negative instances. As a result, they make more inferences that depend on the representation of these negative instances (the modus tollens and denial of the antecedent inferences). Notice that our account does not propose that reasoners construct a logically more prudent representation of hypothetical conditionals; the more explicit representation enables them to make the valid modus tollens inference more readily, but it also renders them more vulnerable to logical fallacies such as the denial of the antecedent inference.

Our view is that hypothetical conditionals, including counterfactuals, can be encompassed within a general theory of conditionals. Hence these psychological results have implications for the philosophical debate on the

proper treatment of counterfactuals. The extension of our general theory of conditional reasoning to encompass counterfactual inference also has implications for artificial intelligence reasoning programs: inferences about factual matters and about hypothetical possibilities can be modelled by the same underlying mechanism.

We suggest that the answer to how people understand and reason with hypothetical conditionals requires a combination of ideas uncovered in cognitive psychological research on factual conditionals, philosophical and linguistic research on counterfactual conditionals, and social psychological research on the mutability of scenarios. The experiment we have reported here is one in a series that examines the inferences reasoners make from hypothetical conditionals, their evaluation of situations as verifying or falsifying them, and their spontaneous production of counterfactuals. The experiments indicate that reasoners have a coherent everyday logic for reasoning about what might be, and what might have been.

Acknowledgements

This research was supported by an Italian grant for postgraduate work to Alessandra Tasso. We thank Mark Keane, Phil Johnson-Laird, Ronan Culhane, Alberto Mazzocco, Vittorio Girotto, David Over, and Susana Segura Vera.

References

Barwise, J. (1986). Conditionals and conditional information. In Traugott, E.C., Ter Meulen, A., Snitzer Reilly, J., and Ferguson, C.A. (Eds.) *On Conditionals.* Cambridge: Cambridge University Press.

Braine, M. D. S. & O' Brien, D. P. (1991). A theory of IF: a lexical entry, reasoning program, and pragmatic principles. *Psychological Review,* 98, 182 - 203.

Byrne, R.M.J. (1989a). Suppressing valid inferences with conditionals. *Cognition,* 31, 61-83.

Byrne, R.M.J. (1989b). Everyday reasoning with conditional sequences. *Quarterly Journal of Experimental Psychology,* 41A, 141-166.

Byrne, R.M.J. and Johnson-Laird, P.N. (1989). Spatial reasoning. *Journal of Memory and Language,* 28, 564-575

Byrne, R.M.J. and Johnson-Laird, P.N. (1992). The spontaneous use of propositional connectives. *Quarterly Journal of Experimental Psychology,* 44A, 89-110.

Byrne, R.M.J. and Tasso, A. (1994). *Cognitive processes in counterfactual inferences: reasoning with hypothetical conditionals.* Mimeo, Trinity College, Dublin.

Chisholm, R. (1946). The contrary-to-fact conditional. *Mind,* 55, 289-307.

Dudman, V. H. (1988).Indicative and subjunctive conditionals. *Analysis,* 48, 113 - 122.

Evans, J.St.B.T., Newstead, S. and Byrne, R.M.J. (1993). *Human Reasoning: The Psychology of Deduction.* Hillsdale: Erlbaum.

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence,* 30, 35 - 79.

Goodman, N. (1973). *Fact, Forecast, and Fiction.* 3rd Edition. New York: Bobbs-Merrill

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.), *Syntax and Semantics, Vol.3, Speech Acts.* New York: Seminar Press.

Kvart, I. (1986). *A Theory of Counterfactuals.* Indianapolis: Hackett.

Jackson, F. (1991). *Conditionals.* Oxford: Oxford University Press.

Johnson-Laird, P.N. (1986). Conditionals and mental models. In Traugott, E.C., Ter Meulen, A., Snitzer Reilly, J., and Ferguson, C.A. (Eds.) *On Conditionals.* Cambridge: Cambridge University Press.

Johnson-Laird, P.N. and Byrne, R.M.J. (1989). *Only* reasoning. *Journal of Memory and Language,* 28, 313-330.

Johnson-Laird, P.N. and Byrne, R.M.J. (1991). *Deduction.* Hove and Hillsdale: Erlbaum.

Johnson-Laird, P.N., Byrne, R.M.J., and Tabossi, P. (1989). Reasoning by model: the case of multiple quantification. *Psychological Review,* 96, 658-673.

Johnson-Laird, P.N., Byrne, R.M.J., and Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review,* 99, 418-439.

Kahneman, D. and Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review,* 93, 136-153.

Kit-Fong Au, T. (1983). Chinese and English counterfactuals: the Shapir - Whorf hypothesis revisited. *Cognition,* 15, 155-188.

Lewis, D. (1973). *Counterfactuals.* Oxford, Blackwell.

Meddis, R. (1984). *Statistics Using Ranks.* Oxford: Basil Blackwell.

Miyamoto, J. M. & Dibble, E. (1986). Counterfactual conditionals and the conjunction fallacy.*Proceedings of the Eight Annual Conference of the Cognitive Science Society.* Hillsdale: Erlbaum.

Pollock, J. L. (1986). *Subjunctive reasoning.* Dordrecht Reidel.

Ramsey, (1931). *The foundations of mathematics and other logical essays.* London: Kegan Paul.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory.* Oxford: Basil Blackwell.