Understanding counterfactuals

David Over[1], Nicole Cruz[2,3], & Mike Oaksford[2]

[1]Psychology Department, Durham University
[2]Department of Psychological Sciences,
Birkbeck, University of London
[3]Laboratoire CHArt (Université Paris 8 & EPHE)

Acknowledgements

Thanks for comments on what follows to Reuben Stern, and to the participants of London Reasoning Workshop, 2017, and the conference in honour of Dorothy Edgington, London, Jan. 2018.

“If we had no faults, we would not take so much pleasure in noticing those of others.” (La Rochefoucauld, 1665).

A classical example (Adams, 1970)

If Oswald did not kill Kennedy, then someone else did.

If Oswald had not killed Kennedy, then someone else would have.

The indicative conditional is generally held to be true, and the counterfactual false.

What exactly is a counterfactual?

If A had been, then B would have been.

Does a counterfactual like the above logically imply that A and B are false?

Or does the assertion of it by a speaker presuppose or suggest that A and B are false?

The psychology of reasoning has not clearly answered these questions.

Modus Ponens (MP) for counterfactuals

Theorists almost universally agree that MP is valid for counterfactuals, but suppose we try to test people’s understanding of this logical rule in the traditional way by asking them to make these assumptions:

If A had been, then B would have been.
A is the case.

It is unclear whether people could make sense of these assumptions.

Counterfactuals and psychology

Counterfactuals have been studied more by social psychologists than by psychologists of reasoning.

Old academic, “I’m so relieved that I didn’t publish that bad idea in the early 1980s. If I had, I would be very embarrassed now.”

MP for counterfactuals

Old academic, “I’m so relieved that I didn’t publish that bad idea in the early 1980s. If I had, I would be very embarrassed now.”

Brilliant young PhD student: “You did publish that
bad idea in a review in 1981.”

MP and the classical example

If Oswald did not kill Kennedy, then someone else did.

If Oswald had not killed Kennedy, then someone else would have.

The indicative conditional is generally held to be true, and the counterfactual false. Suppose we learn that an expert review of the evidence has found that Oswald did not kill Kennedy?

New paradigm psychology of reasoning:

Belief change was a neglected topic in old paradigm, traditional psychology of reasoning, which focused on inferences from arbitrary assumptions.

But belief change, revision, updating over time, and dynamic reasoning, with utility judgments, are central topics in the new paradigm / Bayesian study of reasoning (Oaksford & Chater, 2007, 2013).

The Equation and the Ramsey test

The Equation (Edgington, 1995)

$$P(\text{if } A \text{ then } B) = P(B|A)$$

The Ramsey test (Ramsey, 1929)

To assess P(if A then B), suppose A, make changes to preserve consistency, and infer a degree of confidence in B under this supposition.

The logic of counterfactuals: Another neglected topic

Almost nothing was said in traditional psychology of reasoning about this logic.

There is one major study of MP, and none at all of centering, for counterfactuals.

Two fundamental inferences:
Do people consider them valid?

MP

If A had been, then B would have been.

A

Therefore B

Centering (one premise)

A & B

Thus, if A had been, then B would have been.

Two fundamental inferences: Coherence intervals

For MP, let P(if A then B) = x, and P(A) = y

P(B) should fall in the interval [xy, (1 – y) + xy]

For Centering (one premise), let P(A & B) = x

P(if A then B) should be ≥ x, since P(if A then B) = P(B|A) and P(A & B) = x = P(A)P(B|A).

Coherence intervals: More details

Suppose you flip a fair coin, I call out Heads or Tails while it is in the air, and I am correct at significantly above chance level.

Suppose P(A & B) = .5 for a range of cases, and the participants judge P(if A then B) $\geq$ .5 at an above chance level.

What are the best explanations of these findings?

Assertion: Yet another neglected topic

Assertion is a speech act, an action like any other, and can be given a Bayesian analysis.

An indicative conditional *if A then B* is asserted when $P(B|A)$ is high enough in context. This is not what the conditional means.

A counterfactual is asserted when, additionally, *A* is considered to be false. This is not what it means.

Void conditional assertions

A "void" indicative:

"If I am in the café, then I am drinking a coffee."

A "void" counterfactual:

"If I were speaking to you right now, it would be about void conditionals."

The de Finetti and Jeffrey tables are relevant here.

Conversations and belief revision

Consider the example from Byrne (1989):

If Mary has an essay to write, she is in the library.
If the library is open, Mary is in the library.
Mary has an essay to write.

The second premise is not needed for "suppression" and belief change. It can be deleted, and suppression will still occur as long as the premises are given in a dialogue (Stevenson & Over, 1995, 2001).

The Thompson & Byrne (2002) technique - See also Lewis (1973, 1.7)

John says that: "If Jack had gone to Calgary, then Barbara would have gone to Edmonton."
Mary replies that: "I know that Jack went to Calgary."

The conclusion of MP is highly endorsed using this technique. Note the inference is dynamic and results in belief revision.

A counterfactual cannot be modelled as meaning that its antecedent is false, nor can this falsity be part of the very definition of a counterfactual. Nor can the Ramsey test be that difficult.

The Thompson & Byrne (2002) example - Belief revision

John says, “If A had been, then B would have been.”

Mary replies: A

When we infer B with confidence, we revise our beliefs that A and B are false. Our assessment of P(if A then B), and the linked P(B|A), is invariant: it does not change.

However, we predict that P(if A then B), and P(B|A), will sometimes change for other materials, with <u>suppression</u> or <u>enhancement</u> of the counterfactual premise.

The endorsement of MP for counterfactuals: Cancelling the presupposition

The second speaker casts doubt on the reliability of the first speaker, but this does not make the major premise uncertain.

Can the pragmatic implication that the antecedent is false be cancelled more easily, in general, than a high conditional probability of the consequent given the antecedent? It seems that people will readily cancel the presupposition that the antecedent of a counterfactual is false, that the actual state of affairs makes the antecedent false.

The classic example: Major premise enhancement

John says: “If Oswald had not killed Kennedy then someone else would have.”
Latest scientific finding: Oswald did not kill Kennedy.

Most of us are not conspiracy theorists and would reject the major premise above as a counterfactual.

But given the minor premise, we take the counterfactual to be equivalent to the indicative. The probability of the former is “enhanced”, and so we are confident in the conclusion.

Summary of research questions

Does a counterfactual logically imply that its antecedent and consequent are false for ordinary people?

Are MP and centering valid for counterfactuals as assessed by ordinary people?

How do ordinary people revise their beliefs when they learn that the antecedent of a counterfactual is true?

A study of counterfactual MP: Major premise enhancement

Everyone in the room passed the exam.

John says: "If Jack had been in the room, he would have passed the exam."
Mary replies: "I know that Jack was in the room."

The counterfactual becomes equivalent to an indicative, and the conclusion of MP should be highly endorsed. Note that this example is a problem for Pearl's analysis.

Major premise enhancement: Analysis

Everyone in the room passed the exam.
John says: "If Jack had been in the room, he would have passed the exam."
Mary replies: "I know that Jack was in the room."

We infer *B* from *A* and *if A had been B would have been*. We had low confidence in the counterfactual, but we have high confidence in *A & B*. The counterfactual has become equivalent to the indicative *if A then B*, and by centering we now have to make P(B|A) high.

A study of counterfactual MP: Major premise suppression

John says: “If Jack had worked hard, he would have passed the exam.”
Mary replies: “I know that Jack did work hard.”

The counterfactual becomes equivalent to an indicative, but the major premise should be suppressed, and the conclusion of MP should not be highly endorsed.

Major premise suppression: Analysis

John says: "If Jack had worked hard, he would have passed the exam."
Mary replies: "I know that Jack did work hard."

We infer *B* from *A* and *if A had been B would have been*, but even so, have low confidence in *B*. We had high confidence in the counterfactual, but now we have high confidence in *A & not-B*. Centering makes our confidence in P(not-B|A) high, and so we must have P(B|A) low.

Counterfactual MP: Major premise invariance

John says: "If Jack had got 100%, he would have passed the exam."

Mary replies: "I know that Jack did get 100%."

Our confidence in the major premise is invariant, and the conclusion of MP should be highly endorsed.

Major premise invariance: Analysis

John says: “If Jack had got 100%, he would have passed the exam.”
Mary replies: “I know that Jack did get 100%.”

We infer *B* from   *A*   and *if A had been B would have been*.
We had high confidence in the counterfactual, and we have high confidence in *A & B*, and by centering this makes our confidence in P(B|A) high.

The buttons and the light
Adams (1975) and Edgington (2014)

If button A or button B is pressed, the light will come on. If neither is pressed, or both are, the light will say off.

The light is on, and we think we have good reason to believe A has been pressed. We then doubt the counterfactual:

“If B had been pressed then the light would have come on.”

But now we learn that B was pressed. Our confidence in the counterfactual is enhanced by centering.

Abductive counterfactuals
Edgington (2014)

A doctor says: "The patient has arsenic poisoning because, if he had taken arsenic, these would be his symptoms."

The lab report comes back: The patient has taken arsenic.

There is no point in MP here, for the patient obviously has the symptoms, but centering becomes important. Given the lab report and the symptoms, the doctor will become more confident in the abductive counterfactual.

Yet another classic example (Lewis, 1979)

We believe that Jim and Jack quarrelled yesterday, and Jack is angry about it. There is a case for being confident that: "If Jim asked Jack for help today, Jack would not help him."

There is also a case for being confident that: "If Jim were to ask Jack for help today, there would have been no quarrel, and so Jack would help him."

Suppose we now learn that Jim is asking Jack for help. The important question is <u>how</u> we learn it. Is it by intervention or observation? (Sloman & Lagnado, 2005)

Our coming experiments

We are developing experiments in which we manipulate background knowledge, causal structure, and assertions of major and minor MP premises by different speakers.

The aim is to study how the assertions of counterfactuals, and MP and centering inferences, affect belief revision in dynamic reasoning, with failures of invariance.

Conclusions

Traditional psychology of reasoning neglected the study of belief change, assertion, and counterfactual logic.

The probabilistic and Bayesian approaches to the field have started to overcome these limitations.

Believing any conditional would appear to have little point without the possibility of MP, but MP for counterfactuals is predicted to cause enhancement and suppression in some cases and so lead to failures of invariance. Centering should also be studied for counterfactuals.