Counterfactual Fallacies *

Andrea Iacona†
ai@cc.univaq.it

ABSTRACT

A widely accepted claim about counterfactuals is that they differ from strict conditionals, that is, there is no adequate representation of them as sentences of the form $\Box\alpha \supset \beta$. To justify this claim, Stalnaker and Lewis have argued that some fallacious inferences would turn out valid if counterfactuals were so represented. However, their argument has a flaw, as it rests on a questionable assumption about the relation between surface grammar and logical form. Without that assumption, no consequence of the alleged kind is obtained, hence the claim may be rejected.

1.

A counterfactual is a conditional 'If it were the case that *p*, then it would be the case that *q*', where '*p*' is the antecedent and '*q*' is the consequent. For example, the following sentence is a counterfactual:

(1) If kangaroos had no tails, they would topple over

The obvious paraphrase of (1) is 'If it were the case that kangaroos have no tails, then it would be the case that they topple over'. A strict conditional is a sentence of the form $\Box\alpha \supset \beta$. In the familiar semantics of modal logic, $\Box\alpha \supset \beta$ is true in a world *w* if and only if $\alpha \supset \beta$ is true in every world accessible from *w*. If we call α-world a world in which α is true, this means that $\Box\alpha \supset \beta$ is true in *w* if and only if β is true in every accessible α-world. So it is tempting to say

* Many thanks to Andrea Borghini and José Diez for their comments on previous versions of this paper.
† University of L'Aquila, Italy.

that a counterfactual that has '*p*' as antecedent and '*q*' as consequent – a *p*/*q* counterfactual from now on – is a strict conditional that is true if and only if '*q*' is true in every world of some suitably restricted set in which '*p*' is true.[1]

However, Stalnaker and Lewis have argued that this temptation must be resisted. A strict conditional analysis of counterfactuals may appear tenable when one looks at this or that counterfactual, but it proves inadequate if one reflects on sets of counterfactuals and the logical relations they involve. At least three basic inference rules that hold for strict conditionals do not hold for counterfactuals, that is, there are at least three distinctive "counterfactual fallacies". The first is the fallacy of *strengthening the antecedent*. Consider the argument A1:

(2) If Otto had come, it would have been a lively party

---

(3) If Otto and Anna had come, it would have been a lively party

Imagine that Otto is a cheerful person, but that he just broke up with Anna after six months of endless rows. In such a situation (2) may be true even though (3) is false. In other words, (2) is consistent with

(4) If Otto and Anna had come, it would have been a dreary party

Therefore, A1 is invalid. But the following schema, S1, is valid:

$$\frac{\Box\alpha \supset \beta}{\Box (\alpha \wedge \gamma) \supset \beta}$$

For if β is true in all accessible α-worlds, *a fortiori* it will be true in all accessible α-worlds in which γ is true. This means that A1 cannot be represented as an instance of S1.[2]

The second is the fallacy of *transitivity*. Consider the argument A2:

[1] Mayo (1957) is among the early works in which it is suggested that counterfactuals amount to strict conditionals.

[2] Stalnaker (1991, p. 38); Lewis (1973, pp. 10–13 and 31). The sequence formed by (2) and (4) is called a Sobel sequence, from Lewis(1973, p. 10, fn).

(5) If Anna had gone to the party, Waldo would have gone
(6) If Otto had gone to the party, Anna would have gone

---

(7) If Otto had gone to the party, Waldo would have gone

Imagine that Waldo fancies Anna, although he never runs the risk of meeting his successful rival Otto. Imagine also that Otto was locked up at the time of the party, so that his going to the party is a remote possibility, but that Anna almost did go, as she hoped to meet him. In such a situation (5) and (6) may be true even though (7) is false. Therefore, A2 is invalid. However, the following schema, S2, is valid:

$\Box\beta \supset \gamma$
$\Box\alpha \supset \beta$

$\Box\alpha \supset \gamma$

For if all accessible α-worlds are β-worlds and all accessible β-worlds are γ-worlds, then all accessible α-worlds are γ-worlds. So A2 cannot be represented as an instance of S2.[3]

The third is the fallacy of contraposition. Consider the argument A3:

(8) If Otto had gone to the party, Anna would have gone

---

(9) If Anna had not gone, Otto would not have gone

Imagine that Otto wanted to go to the party but stayed away just to avoid Anna, while Anna would definitely have gone if Otto had been around. In such a situation (8) may be true even though (9) is false. Therefore, A3 is invalid. However, the following schema, S3, is valid:

[3] (Stalnaker, 1991, p. 38; Lewis, 1973, pp. 32–33). Note that S2 entails S1, as it is easily seen if α is replaced with $\alpha \wedge \beta$. So the failure of S1 alone suffices to discard S2.

$$\frac{\Box\alpha \supset \beta}{\Box\neg\beta \supset \neg\alpha}$$

For $\alpha \supset \beta$ and $\neg\beta \supset \neg\alpha$ have the same truth-value in every world. This means that A3 cannot be represented as an instance of S3 (Lewis, 1973, p. 35; Stalnaker, 1991, p. 39).

The Stalnaker-Lewis argument may be summarized as follows. If counterfactuals are strict conditionals, then A1–A3 instantiate S1–S3. But that is absurd. A1–A3 are invalid arguments, while S1–S3 are valid schemas. So counterfactuals are not strict conditionals. This paper is intended to provide a reason to doubt the Stalnaker-Lewis argument.

2.

The line of resistance that will be suggested differs from at least three objections that may be prompted by some contextualist accounts of counterfactuals as strict conditionals that have emerged recently. The assumption that the three objections share is that counterfactuals are highly context-sensitive strict conditionals, in that the accessibility relation associated with them varies as a function of their antecedent. On this assumption, counterfactuals with different antecedents are intuitively assessed relative to different contexts, because their antecedents select different sets of relevantly similar worlds.[4]

The first objection goes as follows. It is wrong to assume that A1–A3 are invalid arguments. In order to evaluate A1–A3, just as any other argument affected by context-sensitivity, the context must be held fixed. An argument is valid if and only if, for every context, if the premises are true relative to that context then the conclusion is true relative to that context. But A1–A3 are such that there is no context relative to which the premises are true and the

[4] The supposition that the counterfactuals in a Sobel sequence – hence in A1 – are strict conditionals that involve different contexts, initially dismissed in (Lewis, 1973, p. 13), is developed in (von Fintel, 2001) and in (Gillies, 2007). (Lowe, 1995, pp. 56–57), suggests that arguments such as A2 can be treated as cases of equivocation due to context-sensitivity.

conclusion false, hence they are valid. The invalidity of A1–A3 is only apparent, due to the context-shifts in their intended reading.[5]

This objection is not entirely convincing. Even if one grants that the counterfactuals in A1–A3 involve different contexts, and that no context makes the premises true and the conclusion false, one is not compelled to conclude that A1–A3 are valid. Certainly, the definition of validity as truth-preservation in any context entails that conclusion, so it clashes with our inclination to regard A1–A3 as invalid. But this clash does not show that our inclination is misplaced more than it shows that the definition is unable to handle such cases. In what follows it will be taken for granted that A1–A3 are invalid, just as they appear.

The second objection is opposite to the first, as it attacks the assumption that S1–S3 are valid schemas. A proponent of the view that counterfactuals are highly context-sensitive strict conditionals may grant that A1–A3 are invalid arguments and that A1–A3 instantiate S1–S3, but claim that S1–S3 are invalid precisely in virtue of that fact. For a schema is valid just in case all its instances are valid arguments.

This objection throws the baby out with the bathwater. To deny that S1–S3 are valid schemas is to deny the basic principles of modal logic. For the validity of S1–S3 follows from those principles. If S1–S3 are invalid, then the semantics of the language in which they are expressed is not the familiar semantics of modal logic, and $\Box$ does not have its familiar meaning. Even if one is willing to accept this consequence, which is not easy to swallow, the question remains of how one can maintain the claim that counterfactuals are strict conditionals in some sense that is relevant to the Stalnaker-Lewis argument. For that argument is intended to dismiss the claim that counterfactuals are strict conditionals in the familiar sense.

The third objection goes as follows. A1–A3 are invalid arguments, S1–S3 are valid schemas, but there is nothing absurd in the supposition that A1–A3 instantiate S1–S3. When $\Box$ occurs more than once in an argument and it is associated with different accessibility relations, the possibility that the premises of the argument are true and the conclusion false is not detectable

[5] A reasoning along these lines is offered in Brogaard & Salerno (2008), although it is not accompanied by a strict conditional analysis of counterfactuals. Cross (2011) questions the contextualist assumptions that underlie that reasoning.

from its logical form. In other words, the invalidity of A1–A3 is not amenable to formal explanation.

This objection is defeatist in at least one important respect. As long as formalization is understood in the usual way as a representation of logical form that displays fundamental logical properties such as validity, it is hard to make sense of the claim that A1–A3 are invalid arguments that instantiate S1–S3. To say so is to say something odd, namely, that although it is correct to represent the counterfactuals in A1–A3 as strict conditionals, such representation plays no role in a formal explanation of the logical properties of A1–A3. Nothing like this will be suggested here. Logical form does play a role in formal explanation, hence the logical properties of A1–A3 must be detectable from the logical form of the counterfactuals in them.

3.

So far there is nothing to object to the Stalnaker-Lewis argument. A1–A3 are invalid arguments, S1–S3 are valid schemas, and the supposition that A1–A3 instantiate S1–S3 leads to absurdity. The flaw of the argument lies elsewhere, namely, in the assumption that if counterfactuals are strict conditionals then A1–A3 instantiate S1–S3. Presumably, the rationale for this assumption is that the only way to represent a *p*/*q* counterfactual as a strict conditional is to suppose that its logical form is expressed by a formula $\Box\alpha \supset \beta$ where α stands for '*p*' and β stands for '*q*'. But that is not the only way, nor is the best. There is another way to represent a *p*/*q* counterfactual as a strict conditional, which is in accordance with the plausible hypothesis that the meaning of the counterfactual is that in any possible world in which *p*, and which resembles our world as much as the supposition that p permits it to, *q*. The view is that the logical form of a *p*/*q* counterfactual is $\Box\alpha \supset \beta$, where α does not stand for '*p*' but for the stronger condition that *p and for the rest things are like in our world as much as the supposition that p permits it to*. For example, in the case of (1) α expresses the condition that kangaroos have no tails and for the rest things are like in our world as much as kangaroos having no tails permits it to. That is, if γ stands for 'Kangaroos have no tails', α amounts to $\gamma \wedge \delta$, where δ expresses the similarity constraint required. The idea that underlies this view turns out clear if one reflects on the contrast between a *p*/*q* counterfactual and an overt strict conditional 'Necessarily, if *p* then *q*'. Consider (1) and the following sentence:

(10) Necessarily, if kangaroos have no tails, then they topple over

While the truth condition of (10) is that kangaroos topple over in any possible world in which they have no tails, the truth-condition of (1) is that kangaroos topple over in any possible world such that kangaroos have no tails and things are like our world as much as the supposition that kangaroos have no tails permits it to. Now consider a formal representation of (1) as $\Box\alpha \supset \beta$. If the same formula were assigned to (10), as required by the supposition that α stands for 'Kangaroos have no tails', there would be no way to distinguish (1) from (10) by looking at its formal representation. But this would go against something that is usually taken for granted about formalization, namely, that sentences with different truth-conditions are to be represented by means of distinct formulas, that is, formulas that can have different truth-values in the same interpretation. It is natural to expect that the difference in truth-conditions between (1) and (10) is formally represented, so that the corresponding formulas have different truth-values in some interpretation. Or at least, this is what Stalnaker, Lewis and many others would say. The simplest way to draw the distinction is to assign a different formula $\Box\alpha \supset \beta$ to (10), assuming that γ stands for 'Kangaroos have no tails' while α amounts to a stronger condition $\gamma \wedge \delta$. In substance, the idea is that the logical form of counterfactuals systematically diverges from their surface grammar, in that the antecedent of the formula that expresses their truth-condition does not correspond to their antecedent. In that sense counterfactuals differ from overt strict conditionals, whose antecedent is stated explicitly.[6]

On this view, A1–A3 do not instantiate S1–S3. Consider A1. If (2) is represented as $\Box\alpha \supset \beta$, then α does not stand for 'Otto has come' but for 'Otto has come and for the rest things are like in our world as much as Otto coming permits it to'. So (3) cannot be represented as $\Box (\alpha \wedge \gamma) \supset \beta$. Rather, it is to be represented as $\Box\gamma \supset \beta$, where γ expresses a condition that entails 'Otto and Anna has come' but is not reducible to a conjunction that includes α. For one thing is to require that a world is similar to ours as much as the truth of 'Otto has come' permits it to, quite another thing is to require that a world is

[6] In a longer paper, *Counterfactuals as Strict Conditionals*, I spell out the view that counterfactuals are strict conditionals whose antecedent is stated elliptically, and compare it with the account of counterfactuals suggested by Stalnaker and Lewis.

similar to ours as much as the truth of 'Otto and Anna has come' permits it to. Therefore, the schema instantiated by A1 is not S1 but the following, S4:

$$\frac{\Box\alpha \supset \beta}{\Box\gamma \supset \beta}$$

Consider A2. If (6) is represented as $\Box\alpha \supset \beta$, then (5) cannot be represented as $\Box\beta \supset \gamma$ but rather as $\Box\gamma \supset \delta$, where γ entails β. Therefore, the schema instantiated by A2 is not S2 but the following, S5:

$$\frac{\begin{array}{l}\Box\gamma \supset \delta \\ \Box\alpha \supset \beta\end{array}}{\Box\alpha \supset \delta}$$

Finally, consider A3. If (8) is represented as $\Box\alpha \supset \beta$, the antecedent of the formula that represents (9) cannot be $\neg\beta$ but a different formula γ that entails $\neg\beta$. Similarly, its consequent cannot be $\neg\alpha$ but a different formula δ that stands for 'Otto has not gone'. Therefore, the schema instantiated by A3 is not S3 but the following, S6:

$$\frac{\Box\alpha \supset \beta}{\Box\gamma \supset \delta}$$

Since S4–S6 are invalid schemas, the invalidity of A1–A3 is easily explained. A fallacy is a bad argument that may appear good at first sight, and counterfactual fallacies are no exception in this respect. A1–A3 may seem valid, in that the antecedents of the counterfactuals they contain make them look similar to other arguments that instantiate valid schemas. But in reality they are invalid, since they do not instantiate those schemas.

REFERENCES

Brogaard, B., & Salerno, J. (2008). Counterfactuals and context. *Analysis*, *68*, 39–46.

Cross,C.B. (2011). Comparative world similarity and what is held fixed in counterfactuals. *Analysis*, *71*, 91–96.

Gillies, A.S. (2007). Counterfactual scorekeeping. *Linguistics and Philosophy*, *30*, 329–360.

Lewis, D. (1973). *Counterfactuals*. London: Blackwell.

Lowe, E.J. (1995). The truth about counterfactuals. *Philosophical Quarterly*, *45*, 41–59.

Mayo, B. (1957). Conditional statements. *Philosophical Review*, *66*, 291–303.

Stalnaker, R. (1991). A theory of conditionals. In F. Jackson (Ed.), *Conditionals*. Oxford: Oxford University Press, 28–45.

von Fintel, K. (2001). Counterfactuals in a dynamic context. In M. Kenstowicz (Ed.), *Ken Hale: A Life in Language*. Cambridge, MA: The MIT Press, 123–152.