American Philosophical Quarterly
Volume 60, Number 3, July 2023

SUFFICIENT CONDITIONS FOR COUNTERFACTUAL TRANSITIVITY AND ANTECEDENT STRENGTHENING

Tristan Grøtvedt Haze

ABSTRACT This paper is about two controversial inference-patterns involving counterfactual or subjunctive conditionals. Given a plausible assumption about the truth-conditions of counterfactuals, it is shown that one can't go wrong in applying hypothetical syllogism (i.e., transitivity) so long as the set of worlds relevant for the conclusion is a subset of the sets of worlds relevant for the premises. It is also shown that one can't go wrong in applying antecedent strengthening so long as the set of worlds relevant for the conclusion is a subset of that for the premise. These results are then adapted to Lewis's theory of counterfactuals.

KEYWORDS conditionals, counterfactuals, conditional logic, transitivity, hypothetical syllogism, strengthening the antecedent.

Introduction

Consider the following patterns of inference involving subjunctive or counterfactual conditionals:

Hypothetical syllogism (i.e., transitivity):

$A \mathrel{\square\!\!\rightarrow} B$

$B \mathrel{\square\!\!\rightarrow} C$

$\therefore A \mathrel{\square\!\!\rightarrow} C$

Antecedent strengthening:

$A \mathrel{\square\!\!\rightarrow} B$

$\therefore (A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$

These inference-patterns can seem intuitively compelling considered in the abstract or using certain near-to-hand instances. The following instances, for example, seem like good bits of reasoning:

If you were to eat these berries, you'd get sick.

If you were to get sick, we'd have to go home.

∴ If you were to eat these berries, we'd have to go home.

If you had come to my house, I would have served coffee.

∴ If you had come to my house and brought biscuits, I would have served coffee.

Other instances, however, seem like bad bits of reasoning:

If J. Edgar Hoover had been a communist, he would have been a traitor.

If J. Edgar Hoover had been born a Russian, he would have been a communist.

∴ If J. Edgar Hoover had been born a Russian, he would have been a traitor.[1]

If you had jumped out the window, you would have hurt yourself.

> ∴ If you had jumped out the window and magically began to fly, you would have hurt yourself.[2]

This raises the question: under what conditions is it OK to reason according to these patterns? More precisely, under what conditions can these patterns be guaranteed not to lead from truth to falsity? The present paper addresses this question.

There is a venerable tradition of dealing with the fact that the patterns in question have bad instances by proposing related inference-patterns which avoid counterexamples and which leading theories predict to be valid. Some of these alternatives may be regarded as adding premises to the controversial patterns in question, while others may be regarded as modifying the premises. In this tradition we find the "substitutes" for the two patterns offered by Lewis (1973, 433). Bennett (2003, 332) continues the tradition by proposing a pattern, Limited Antecedent Strengthening, which is valid on the Lewisian semantics.

This tradition has furnished valuable results, but it is hard to shake the feeling that the apparently good instances of the patterns really are good bits of reasoning—aren't merely similar to good bits of reasoning—and are so all by themselves, without extra premises. These may not be formally valid patterns, but they somehow still appear to be patterns which we do well to remember and reason in accord with. This paper pursues a viewpoint according to which, when we appear to do good reasoning according to one of the patterns, we really are reasoning according to one of the patterns (not some similar pattern), and what licenses us need not be located in some implicit premise.

The license may instead be located in our understanding of what is and is not relevant to the counterfactuals involved. This is not a new idea: in Section 3, I consider a version of this approach due to Brogaard and Salerno. Brogaard and Salerno are successful in articulating a sufficient condition for such patterns not to lead us astray, but their condition is needlessly strong. I want to show that there can be good instances of the patterns which do not fulfil the Brogaard and Salerno condition but which do fulfil the weaker sufficient condition that I propose.

Perhaps surprisingly, I am not taking a stand here on the issue of whether the inference-patterns in question should be called *valid*. Since these patterns seem to have counter-instances, I suspect that we shouldn't call them valid. I want to articulate conditions under which instances of the patterns (whether or not the patterns themselves are valid) can be guaranteed to preserve truth.

Finally, let me mention, in order to set aside, a different approach one might take in pursuing the idea that we sometimes really do reason, appropriately, according to the controversial patterns. One might use Stalnaker's notion of *reasonable inference*, or something like it. For Stalnaker:

> an inference from a sequence of assertions or suppositions (the premises) to an assertion or hypothetical assertion (the conclusion) is reasonable just in case, in every context in which the premises could appropriately be asserted or supposed, it is impossible for anyone to accept the premises without committing himself to the conclusion. (Stalnaker 1975, 71)[3]

One might argue that sometimes when we reason according to the controversial patterns, we make a reasonable inference in this sense. Instead of working with such notions as appropriate assertion or supposition, my approach has been to stick to truth-conditional semantics and to look for minimal sufficient conditions for truth-preservation. It is these that I report here. To connect them back to actual cases of inferring according to the patterns, my idea is that we often know implicitly that these conditions are fulfilled. That, I think, is part of what makes the minimal sufficient conditions interesting, but my focus in this paper is on the conditions themselves. These conditions might be of use in telling a

Stalnakerian story about when an inference according to the controversial patterns is reasonable, but looking into that matter is a task that falls outside my scope here.

1. A Plausible Assumption

Consider the following plausible assumption about the truth-conditions of counterfactuals:

> (Assumption) A counterfactual $A \boxdot\!\!\rightarrow C$ is true iff in all relevant worlds the corresponding material conditional $A \supset C$ is true.

Or equivalently:

> A conditional $A \boxdot\!\!\rightarrow C$ is true iff all relevant A worlds (i.e., all relevant worlds in which A is true) are C worlds.

(The first formulation is most convenient for the purpose of proving the results of this paper.)

I want to emphasize that (Assumption) is *not* presented as indicating the form that a formal semantics of counterfactuals should take. In particular, I want to stay as neutral as I can on the live issue of whether and how the contextual variability of counterfactuals should be reflected in their formal semantic treatment.[4] To assume (Assumption) is merely to suppose that it is something we can correctly say about counterfactuals. Before proceeding, it should be noted that insofar as (Assumption), suitably understood, is true of indicative conditionals, the present results about counterfactuals carry over to indicatives—but this is not my focus here.[5]

2. Two Concepts of Relevant Worlds

When discussing counterfactuals in tandem with something like (Assumption), there are two importantly different ways of understanding "relevant world": broad and narrow. On the broad conception, a world can count as "relevant" for a counterfactual regardless of whether its antecedent holds at that world. (Such worlds may count as relevant because they match the actual world with respect to certain background facts, because they are similar enough to the actual world in the right respects, or in some cases perhaps merely by being possible—details may vary from case to case and theory to theory.) On the narrow conception, part of what it is to be a "relevant" world is to be a world where the antecedent holds. When illustrating the formal results, the present paper works with the former, broad conception. Why this is crucial will become clearer when we give the illustrations in Section 4, but let me pause here to illustrate the broad conception of relevant worlds a bit further.

Suppose Sarah likes cats, but strongly prefers sourcing pets from shelters and the like over sourcing them from pet shops, and has in fact sworn off ever obtaining a cat from a pet shop. With this in mind, we might say something like "If Sarah were to get a cat, it would not be from a pet shop." It is natural to understand what we are doing here as follows: we are, for the purposes of evaluating this conditional, holding fixed the fact that Sarah has sworn off ever getting a cat from a pet shop, and thus worlds in which Sarah does not mind getting a cat from a pet shop are ruled out as not relevant in the broad sense. On this picture, there is a bunch of broadly relevant worlds, and the most salient thing about them is that Sarah has, in these worlds, sworn off ever getting a cat from a pet shop. There will usually be other facts held fixed in the background too, such as that Sarah will not be forced by some malevolent party to get a cat from a pet shop despite her best intentions, and that there are viable ways to get a cat besides going to a pet shop. Then, we and our audience are in a position to verify the counterfactual by as it were selecting, from among the broadly relevant worlds, the ones where Sarah gets a cat, and checking that all of them are worlds in which the cat does not come from a pet shop. However, suppose that in the very same conversation

or chain of reasoning we also say "If Sarah were to get a cat, then she would have an animal." Here I think it is natural to regard a wider sweep of worlds as broadly relevant, including worlds in which Sarah has *not* sworn off ever getting a cat from a pet shop. This may be further brought out by reflecting that this sentence may have been intended by its utterer, and may be understood by its audience, in such a way that its significance is the same as, or very similar to, the following variants: "If Sarah were to get a cat, then no matter what, she would have an animal," or "If Sarah were to get a cat, then she would certainly have an animal, no matter where the cat came from." The salient background fact here is that cats are by nature animals. And this utterance need not cancel or interfere with the felicity of the earlier conditional, which was founded on Sarah's aversion to getting cats at pet shops. It seems we can, in a given conversation or piece of reasoning, keep track of what is relevant to—what lies behind, so to speak—different counterfactuals that are in play.

3. The Same-Set Proposal

To recall, the question addressed by the present paper is: under what conditions can the patterns of hypothetical syllogism and strengthening the antecedent be guaranteed not to lead us from truth to falsity?

Brogaard and Salerno (2008) suggests an answer. They use a simplified version of Lewis's (1973) account as their theoretical framework. Rather than talking directly of relevant worlds—which, on Lewis's account as commonly glossed, would be those closest to the world of evaluation in relevant respects—Brogaard and Salerno conduct their discussion in terms of background facts. In explanation, they write that "whether, in *w* [an arbitrary world], A counterfactually implies B is a matter of whether B holds in the A worlds that share (with *w*) the relevant background facts" (40). Thus, an expansion of background facts corresponds to a contraction of relevant worlds, and vice versa.

Brogaard and Salerno propose that "the set of contextually determined background facts must remain fixed when evaluating an argument involving subjunctives for validity" (Brogaard and Salerno 2008, 42). One set of background facts per argument. This offers a way of explaining the failure of certain instances of the patterns in question, such as those above about J. Edgar Hoover and jumping out of the window, while preserving the plausible idea that the patterns in question are often good ways of reasoning; what has gone wrong in the bad instances is that those instances involve counterfactuals whose sets of associated background facts differ from one another.

Transposed into the present framework and restricted to the patterns in question[6], Brogaard and Salerno's proposal amounts to this: for an instance of one of the patterns in question to be an instance of good reasoning, the set of worlds relevant for each counterfactual in the instance must be the same.[7] Call this *the same-set proposal*, and call the condition it imposes *the same-set condition*.

The basic idea here has been expressed by other authors. As Dohrn (ms.) notes, the same-set proposal was anticipated by Wright. Referring to the puzzlingness of uttering, in the same breath, "If Thatcher had been born and brought up a Russian, she would have been dedicated to the overthrow of the Western democracies" *and* "If Thatcher had been dedicated to the overthrow of Western democracies, she would have been a traitor to the UK," Wright asks:

> [W]hat is the explanation of this puzzlingness, if not that the convention is that when a number of counterfactuals are in play in a single context, some single range of relevant worlds [. . .] governs the assessment of them all? (Wright 1983, 138.)

As we can see, for Wright the idea applies to any situation where there are multiple

counterfactuals in play in a single context, not just when applying one of the inference-patterns in question. (Note also that Wright is working here with a broad conception of relevance (in the sense of Section 2), as opposed to a narrow one.) In von Fintel (2001), we find the idea applied to instances of strengthening the antecedent:

> It appears to me that speakers can reasonably offer arguments of the form of Strengthening the Antecedent. What should we say about such behaviour? Do they commit a fallacy? More likely, what we want to say is that they must be making tacit additional assumptions that make their inference valid. According to my account, the additional assumption that they are making is that the accessibility function [which is supposed to map the scenario at which a conditional is being evaluated to the set of scenarios relevant for its evaluation there] is such that it remains constant throughout the inference. (von Fintel 2001, 144. Parenthesis added)

As I have indicated, I have sympathy with this general approach of trying to say something precise in favor of these controversial patterns, and wish to carry it further. In the next section I prove some results which show, or at the very least strongly suggest, that the same-set condition for the legitimacy of an instance of one of the two patterns is too strict. It *is* a sufficient condition, but not the weakest one available; you will not go wrong if you adhere to the same-set condition, but you might miss out on some truths.

4. Subsethood is Enough

I want to show that (assuming (Assumption)):

(1) You can't go wrong in applying hypothetical syllogism so long as the set of worlds relevant for the conclusion is a subset of the sets of worlds relevant for the premises.

(2) You can't go wrong in applying strengthening the antecedent so long as the set of worlds relevant for the conclusion is a subset of that for the premise.

By "go wrong" I simply mean "have true premises and a false conclusion." Also, note that I say "subset" and not "proper subset"; fulfillment of the same-set condition by an instance of one of the inference-patterns in question is thus a special case of fulfillment of (the relevant one of) the above.

To show that (1) and (2) are true, I prove two theorems, illustrating each one with a piece of reasoning which, naturally interpreted, flouts the same-set condition but fulfills the relevant subset condition.

> I will use Rel(X) to denote the set of worlds relevant for a conditional X.
>
> Theorem 1. For any three counterfactuals $A \mathrel{\square\!\!\rightarrow} B$, $B \mathrel{\square\!\!\rightarrow} C$ and $A \mathrel{\square\!\!\rightarrow} C$ such that $Rel(A \mathrel{\square\!\!\rightarrow} C) \subseteq Rel(A \mathrel{\square\!\!\rightarrow} B)$ and $Rel(A \mathrel{\square\!\!\rightarrow} C) \subseteq Rel(B \mathrel{\square\!\!\rightarrow} C)$, $A \mathrel{\square\!\!\rightarrow} C$ will be true if $A \mathrel{\square\!\!\rightarrow} B$ and $B \mathrel{\square\!\!\rightarrow} C$ are true.
>
> *Proof*: Take any three counterfactuals $A \mathrel{\square\!\!\rightarrow} B$, $B \mathrel{\square\!\!\rightarrow} C$ and $A \mathrel{\square\!\!\rightarrow} C$ such that $Rel(A \mathrel{\square\!\!\rightarrow} C) \subseteq Rel(A \mathrel{\square\!\!\rightarrow} B)$ and $Rel(A \mathrel{\square\!\!\rightarrow} C) \subseteq Rel(B \mathrel{\square\!\!\rightarrow} C)$. Now suppose that $A \mathrel{\square\!\!\rightarrow} B$ and $B \mathrel{\square\!\!\rightarrow} C$ are true. Since all the worlds relevant for $A \mathrel{\square\!\!\rightarrow} C$ are also relevant for $A \mathrel{\square\!\!\rightarrow} B$ and $B \mathrel{\square\!\!\rightarrow} C$, and $A \mathrel{\square\!\!\rightarrow} B$ and $B \mathrel{\square\!\!\rightarrow} C$ are true, the material conditionals $A \supset B$ and $B \supset C$ will both be true at all the worlds relevant for $A \mathrel{\square\!\!\rightarrow} C$ (by (Assumption)). Since transitivity holds for material conditionals, $A \supset C$ will be true at all the worlds relevant for $A \mathrel{\square\!\!\rightarrow} C$, making $A \mathrel{\square\!\!\rightarrow} C$ true (by (Assumption)). Therefore, if $A \mathrel{\square\!\!\rightarrow} B$ and $B \mathrel{\square\!\!\rightarrow} C$ are true, $A \mathrel{\square\!\!\rightarrow} C$ will be true.

Illustration:

> If I had spoken to a cat then I would have spoken to an animal.
>
> If I had spoken to an animal then I would have been happy.
>
> ∴ If I had spoken to a cat then I would have been happy.

It is natural to think of the set of worlds relevant for the first premise as a superset of that for the conclusion. This could be further brought out by adding something like "no matter what" to the first premise.

Understanding the first premise in the way I have in mind, it seems natural to think that even worlds which differ substantially from actuality with respect to what tends to make me happy will be among the relevant ones; why would they be excluded? The point underlying the first premise, we might say, is that cats are by their very nature animals.[8] With the second premise on the other hand, it is natural to regard such worlds as non-relevant; the point underlying the second premise, we might say, is that speaking to animals happens to make me happy. That is a background fact that must hold at a world in order for that world to be relevant to the second premise.

Note here that for this illustration to make sense, we need the broad conception of relevant worlds distinguished in Section 2. Using the narrow conception (where for a world to be relevant to a conditional, the conditional's antecedent must hold at that world), it does not seem correct to say that the worlds relevant for the second premise are a subset of that for the first, for on the narrow conception all worlds relevant for the first premise are ones where I spoke to a cat, whereas there may be worlds relevant for the second premise in which I speak not to a cat, but to some other animal. But since "relevant worlds" in (Assumption) is intended the broad sense, this is not a problem.

Note also that if, following Kripke (1980), one regards the first premise as necessarily true, this does not make the above argument any less an instance of hypothetical syllogism, or any less illustrative of Theorem 1. I have encountered the objection that the first premise is necessarily true and therefore "redundant," with the result that there is something wrong with the above illustration. But this is confused for a number of reasons. Firstly, we often reason non-trivially with necessary truths as premises. Secondly, even if the above argument minus the first premise is already necessarily truth-preserving, it is not plausibly *a priori* that it is necessarily truth-preserving (Kripke's point being that it is not *a priori* that all cats are animals, even if it is necessarily the case that they are) and therefore the conclusion should not be held to follow deductively from the second premise alone, insofar as deduction is supposed to be an *a priori* affair. Thirdly, even if it were knowable *a priori* that if the second premise is true then the conclusion must be, there may be psychological or epistemic reasons to include the first premise anyway. The crucial point is just that we *can* use the premise in a good instance of hypothetical syllogism. Its being necessarily true in light of a familiar Kripkean doctrine does not affect this.

> Theorem 2. For any two counterfactuals $A \mathrel{\square\!\!\rightarrow} B$ and $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ such that $\mathrm{Rel}((A \,\&\, C) \mathrel{\square\!\!\rightarrow} B) \subseteq \mathrm{Rel}(A \mathrel{\square\!\!\rightarrow} B)$, $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ will be true if $A \mathrel{\square\!\!\rightarrow} B$ is true.
>
> *Proof*: Take any two counterfactuals $A \mathrel{\square\!\!\rightarrow} B$ and $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ such that $\mathrm{Rel}((A \,\&\, C) \mathrel{\square\!\!\rightarrow} B) \subseteq \mathrm{Rel}(A \mathrel{\square\!\!\rightarrow} B)$. Now suppose that $A \mathrel{\square\!\!\rightarrow} B$ is true. Since all the worlds relevant for $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ are also relevant for $A \mathrel{\square\!\!\rightarrow} B$, and $A \mathrel{\square\!\!\rightarrow} B$ is true, the material conditional $A \supset B$ will be true at all the worlds relevant for $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ (by (Assumption)). Since antecedent strengthening holds for material conditionals, $(A \,\&\, C) \supset B$ will be true at all the worlds relevant for $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$, making $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ true (by (Assumption)). Therefore, if $A \mathrel{\square\!\!\rightarrow} B$ is true, $(A \,\&\, C) \mathrel{\square\!\!\rightarrow} B$ will be true.

Illustration:

> (Assume for this illustration that the questioner and answerer below live in an area where there are many animals, and have just spent the day looking for cats. The questioner in this illustration does not know that all cats are animals, thinking mistakenly that some are robots.)
>
> Q: Do you think that, in view of how many animals there are around here, we would have found an animal today if we had stayed in this area and found a cat?

A: If we had found a cat today, then, no matter what, we would have found an animal today! All cats are animals! So yes, of course we would have found an animal today if we had stayed in this area and found a cat!

To put the reply in the form of an explicit argument:

If we had found a cat today, we would have found an animal today.

∴ If we had stayed in this area and found a cat today, we would have found an animal today.

Given how the conclusion was intended by the questioner, it is natural to regard its set of relevant worlds as being restricted to those in which the local area contains many animals. But again, for the premise it is natural to regard a larger sweep of worlds as relevant—and crucially, doing so does not affect the goodness of the argument.

5. Adapting the Results to the Lewisian Framework

A requirement close in spirit to the subset condition is available in the popular Lewis (1973) framework for counterfactuals. Given a counterfactual A $\boxdot\!\!\rightarrow$ C, call a sphere of worlds S *antecedent-permitting* for that counterfactual iff S contains A worlds (i.e. worlds at which the antecedent is true).

We may now formulate Lewisian analogues of (1) and (2) above:

(1L) You can't go wrong in applying hypothetical syllogism so long as every sphere S which is antecedent-permitting for one of the premises is antecedent-permitting for the conclusion.

(2L) You can't go wrong in applying strengthening the antecedent so long as every sphere S which is antecedent-permitting for the premise is antecedent-permitting for the conclusion.

We can see how (1L) and (2L) work by seeing how their conditions are flouted by the intuitively bad instances of the patterns that we began with:

If J. Edgar Hoover had been a communist, he would have been a traitor.

If J. Edgar Hoover had been born a Russian, he would have been a communist.

∴ If J. Edgar Hoover had been born a Russian, he would have been a traitor.

Consider a sphere S which contains worlds where Hoover is a communist, but none in which he is born a Russian. If there is such a sphere—and intuitively there will be if, for the evaluation of the first premise, country of origin counts more toward similarity than political affiliation—then the intuitive badness of the above inference may be explained in the Lewisian framework by pointing out that it violates (1L).

If you had jumped out the window, you would have hurt yourself.

∴ If you had jumped out the window and magically began to fly, you would have hurt yourself.

Consider a sphere S which contains worlds where you jump out the window, but none in which you magically begin to fly. If there is such a sphere—which intuitively there will be if people tend not to begin magically to fly—then the intuitive badness of the above inference may be explained in the Lewisian framework by pointing out that it violates (2L).

Finally, we may prove theorems in the Lewisian framework that are analogous to Theorems 1 and 2 above. We assume in the background a Lewisian system of spheres \$—a function mapping each world *w* to a nested set $\$_w$ of spheres (where spheres are construed as sets of worlds). Recall that, on Lewis's theory, a counterfactual A $\boxdot\!\!\rightarrow$ C is true at a world *w* iff A is false at all worlds, or there is an A & C world which is closer to *w* than any A & ~C world. In terms of spheres, the second part of this truth-condition amounts to the following: there is a sphere S $\in \$_w$ which contains an A world and which is such that all A worlds in S are C worlds.

Theorem 1L. For any world *w* and any three counterfactuals A $\boxdot\!\!\rightarrow$ B, B $\boxdot\!\!\rightarrow$ C and A $\boxdot\!\!\rightarrow$ C such that for all S $\in \$_w$, if there is (an A world

or) a B world in S then there is an A world in S, A $\Box\!\!\rightarrow$ C is true at *w* if A $\Box\!\!\rightarrow$ B and B $\Box\!\!\rightarrow$ C are true at *w*.

Proof: Take any three counterfactuals A $\Box\!\!\rightarrow$ B, B $\Box\!\!\rightarrow$ C and A $\Box\!\!\rightarrow$ C and a world *w* such that for all spheres S $\in$ $\$_w$, if there is (an A world or) a B world in S then there is an A world in S. Now suppose that A $\Box\!\!\rightarrow$ B and B $\Box\!\!\rightarrow$ C are true at *w*. Case 1. A is false at all worlds. In that case A $\Box\!\!\rightarrow$ C is (vacuously) true at *w*. Case 2. A is true at some world. Recall, we are supposing that A $\Box\!\!\rightarrow$ B and B $\Box\!\!\rightarrow$ C are true at *w*. That means that there is a sphere $S_{P1} \in \$_w$ such that there are A worlds in S_{P1} and all of them are B worlds and there is a sphere $S_{P2} \in \$_w$ such that there are B worlds in S_{P2} and all of them are C worlds. Now, $\$_w$ is nested, meaning that for all S and S' $\in$ $\$_w$, either S $\subseteq$ S' or S' $\subseteq$ S, so we consider two subcases.

Case 2a. $S_{P1} \subseteq S_{P2}$. We already have that there are A worlds in S_{P1} and all of them are B worlds as well as that there are B worlds in S_{P2} and all of them are C worlds. Now, since $S_{P1} \subseteq$ SP2, all A worlds in S_{P1} are also in S (and are B worlds), so we have that all A worlds in S_{P1}—and there are such—are C worlds. Since $S_{P1} \in \$_w$, A $\Box\!\!\rightarrow$ C is true at *w*.

Case 2b. $S_{P2} \subseteq S_{P1}$. We already have that there are B worlds in S_{P2} and all of them are C worlds. Now consider an arbitrary A world in S_{P2}—and there must be such, since we have that for all S $\in$ $\$_w$, if there is a B world in S then there is an A world in S. Since $S_{P2} \subseteq S_{P1}$ and all A worlds in S_{P1} are B worlds, that means our A world must be a C world, since it is in S_{P2}. But this was an arbitrary A world, and so all A worlds in S_{P2} are C worlds. Since $S_{P2} \in \$_w$, A $\Box\!\!\rightarrow$ C is true at *w*.

Theorem 2L. For any world *w* and any two counterfactuals A $\Box\!\!\rightarrow$ B and (A & C) $\Box\!\!\rightarrow$ B such that for all S $\in$ $\$_w$, if there is an A world in S then there is an A & C world in S, (A & C) $\Box\!\!\rightarrow$ B is true at *w* if A $\Box\!\!\rightarrow$ B is true at *w*.

Proof: Take any two counterfactuals A $\Box\!\!\rightarrow$ B and (A & C) $\Box\!\!\rightarrow$ B and a world *w* such that for all S $\in$ $\$_w$, if there is an A world in S then there is an A & C world in S. Now suppose that A $\Box\!\!\rightarrow$ B is true at *w*. Case 1. A & C is false at all worlds. In that case (A & C) $\Box\!\!\rightarrow$ B is (vacuously) true at *w*. Case 2. A & C is true at some world. Recall, we are supposing that A $\Box\!\!\rightarrow$ B is true at *w*. That means that there is a sphere $S_{P1} \in \$_w$ such that all A worlds in S_{P1} are B worlds. Now consider an arbitrary A & C world in S_{P1}—and there much be such, since we have that for all S $\in$ $\$_w$, if there is an A world in S then there is an A & C world in S. By the meaning of &, this is an A world, and therefore also a B world. But this was an arbitrary A & C world, and so there are A & C worlds in S_{P1} and all of them are B worlds. Since $S_{P1} \in \$_w$, (A & C) $\Box\!\!\rightarrow$ B is true at *w*.[9]

6. Conclusion

I have shown the informally stated (1) and (2) to be true given a plausible assumption about counterfactuals by proving two associated theorems, and have adapted these principles to the Lewision theory of counterfactuals, showing (1L) and (2L) to be true given Lewis's semantics by proving two associated theorems. The natural thing to conclude from all this, I think, is that instances of the inference-patterns in question which satisfy the relevant one of (1) or (2) (or the relevant one of (1L) or (2L)) are legitimate and non-fallacious.

Are the sufficient conditions I have given also necessary conditions for the legitimacy of instances of the inference-patterns in question? Clearly, there will be instances where the relevant condition is not fulfilled and yet the premises and conclusion are true, but whether it could ever be legitimate to *infer* the conclusion from the premises in such an instance is a further question. I am inclined to think that, insofar as we are talking about deductive inferences without suppressed premises, the answer is no, but I have not tried to argue for that here. And again, what we should say about the *validity* of these patterns turns on subtle questions which I have tried to steer clear of.

University of Melbourne

NOTES

1. Adapted from Stalnaker (1968, 106). It is interesting to note that this argument is more easily made to seem invalid, and the premises more readily understood the way they are intended, when the premises are put in this order.

2. I have not been able to determine the origin of this sort of example, but it is part of logical lore. I first encountered it in a course handout as an undergraduate.

3. This account is rendered mathematically precise at the end of Stalnaker's paper.

4. Lewis (1973) argues against semantic theories which treat counterfactuals as strict conditionals, that is, material conditionals with necessity operators applied to them. von Fintel (2001) and Gillies (2007) argue for retaining a strict-conditional analysis, but in a dynamic semantic framework. Iacona (2015) defends the strict conditional analysis of counterfactuals by instead proposing that counterfactuals have elliptically-stated antecedents. Moss (2012) defends Lewis's account from the objections of von Fintel and Gillies by means of pragmatic considerations. Karen Lewis (2018) attempts to steer between the von-Fintel-Gillies approach and Moss's, drawing on the strengths of each.

5. Theories of indicatives that are compatible with (Assumption) include those of Stalnaker (1968), Kratzer (1986), and Gillies (2009), but the details of what makes a possible world "relevant" in the relevant sense differ from theory to theory.

6. My (2016) argues that, as a general rule about arguments involving conditionals, Brogaard and Salerno's proposal is too strict: it seems as though two conditionals about unrelated subjects, involving different background facts, can with perfect propriety be put together into a single statement using conjunction introduction. (This may be a trivial piece of reasoning, but to be trivial is not to be illegitimate.) My contention there leaves open the possibility that Brogaard and Salerno are nevertheless correct that for an application of *one of the patterns in question* to be legitimate, all conditionals involved must be alike in background facts. The present paper challenges this.

7. I do not mean to imply that there are no options in philosophical space for agreeing with Brogaard and Salerno's proposal but disagreeing with the same-set proposal. Nevertheless, the same-set proposal seems like a natural transposition of their proposal into the present framework of (Assumption).

8. It is noteworthy that, in terms of Kratzer (1989), worlds where I spoke to a cat on some particular occasion are worlds in which the proposition that I spoke to an animal on that occasion *lumps* the proposition that I spoke to a cat on that occasion; given that I spoke to an animal on that occasion it is not a *separate fact* that I spoke to a cat. (In the article just cited Kratzer develops an interesting analysis of this notion of lumping in terms of situations, reckoned as parts of worlds.)

9. The above results are formulated in terms of a system of spheres $, a handy way of representing information about the comparative similarity of worlds. This information can be represented more directly as a *comparative similarity system*. Lewis (1973, 49) gives a recipe for deriving a system of spheres from a comparative similarity system. This recipe, however, is only defined for systems with no incomparabilities (distinct worlds *j* and *k* such that neither is more similar to some world *i* than the other, and nor are they equally similar to *i*). Other frameworks, notably those of Pollock (1976) and Kratzer (1977; 1979), permit incomparabilities. This raises the question of how the present results could be extended to such a framework. I will explain an imperfect way of doing it, and then explain why it is imperfect in the hope that someone can shed further light. Given an instance of one of the patterns and an incomparability-permitting comparative similarity system C, consider all the comparative similarity systems which agree with C on all comparisons but in which there are no incomparabilities. For each, use Lewis's recipe to derive a system of spheres $. See if the relevant sufficient condition (from 1L or 2L depending on the pattern in question) is met for each system of spheres. If so, the instance will be truth-preserving according to the original comparative similarity system C. This is guaranteed by the

fact, noted in Lewis (1973, 49) in connection with the aforementioned recipe, that "a counterfactual is true at a world according to the defined system of spheres $ if and only if it is true at that world according to the original comparative similarity system," together with the fact that a counterfactual is true on Pollock's or Kratzer's semantics if it is true on Lewis's or Stalnaker's semantics no matter how the missing comparisons are made. (See Lewis 1981, 226.) Why is this an imperfect solution? Because, in the case of infinitely many worlds, it's not the case that a counterfactual is true on Pollock's or Kratzer's semantics *only* if it is true on Lewis's or Stalnaker's semantics no matter how the missing comparisons are made, and when these diverge it is the Pollock-Kratzer prediction which seems correct: see Swanson (2014, 305–306) for an example. Swanson develops an alternative way for Lewis's theory to handle incomparabilities, where, instead of the simple supervaluational move of considering all the comparative similarity systems which agree with C on all comparisons but in which there are no incomparabilities, one employs *ordering supervaluationism*. The general idea here is that "a sentence is supertrue according to ordering supervaluationism iff there is some lower bound on interpretations such that the sentence is true according to every interpretation within that bound" (Swanson 2014, 298). However, given the structure of this (super)truth condition, where the main logical operator is an existential quantifier, I do not see how to adapt the above strategy so that Swanson's approach takes the place of the simple supervaluational move. (Thanks to an anonymous referee for this journal for information about this issue.)

REFERENCES

Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals* (Oxford: Oxford University Press).

Brogaard, Berit and Salerno, Joe. 2008. "Counterfactuals and context," *Analysis*, vol. 68, no. 297, pp. 39–46.

Dohrn, Daniel. m.s. "Counterfactuals, Accessibility, and Comparative Similarity," http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.591.7590&rep=rep1&type=pdf

Gillies, Anthony S. 2007. "Counterfactual scorekeeping," *Linguistics and Philosophy*, vol. 30, no. 3, pp. 329–360.

Gillies, Anthony S. 2009. "On truth-conditions for if (but not quite only if)," *Philosophical Review*, vol. 118, no. 3, pp. 325–349.

Haze, Tristan. 2016. "Against the Brogaard-Salerno Stricture," *The Reasoner*, vol. 10, no. 4, pp. 29–30.

Iacona, Andrea. 2015. "Counterfactuals as Strict Conditionals," *Disputatio*, vol. 7, no. 41, pp. 165–191.

Kratzer, Angelika, 1977. "What 'must' and 'can' must and can mean," *Linguistics and Philosophy*, vol. 1, no. 3, pp. 337–355.

Kratzer, Angelika. 1979. "Conditional necessity and possibility," in *Semantics From Different Points of View*, eds. Rainer Bäuerle, Urs Egli and Arnim von Stechow (Heidelberg: Springer Verlag), pp. 117–147.

Kratzer, Angelika. 1986. "Conditionals," *Chicago Linguistics Society*, vol. 22, no. 2, pp. 1–15.

Kratzer, Angelika. 1989. "An investigation of the lumps of thought," *Linguistics and Philosophy*, vol. 12, no. 5, pp. 607–653.

Kripke, Saul A. 1980. *Naming and Necessity* (Oxford: Blackwell).

Lewis, David K. 1973. *Counterfactuals* (Oxford: Blackwell).

Lewis, Karen S. 2018. "Counterfactual Discourse in Context," *Noûs*, vol. 52, no. 3, pp. 481–507.

Moss, Sarah. 2012. "On the Pragmatics of Counterfactuals," *Noûs*, vol. 46, no. 3, pp. 561–586.

Pollock, John L. 1976. "The 'possible worlds' analysis of counterfactuals," *Philosophical Studies*, vol. 29, no. 6, pp. 469–476.

Stalnaker, Robert C. 1968. "A Theory of Conditionals," in *Studies in Logical Theory* (*American Philosophical Quarterly Monographs 2*), ed. Nicholas Rescher (Oxford: Blackwell), pp. 98–112.

Stalnaker, Robert C. 1975. "Indicative conditionals," *Philosophia*, vol. 5, no. 3, pp. 269–286.
Swanson, Eric. 2014. "Ordering Supervaluationism, Counterpart Theory, and Ersatz Fundamentality." *Journal of Philosophy*, vol. 111, no. 6, pp. 289–310.
von Fintel, Kai. 2001. "Counterfactuals in a dynamic context," in *Ken Hale: A life in language*, ed. M. Kenstowicz (Cambridge: MIT Press), pp. 123–152.
Wright, Crispin. 1983. "Keeping Track of Nozick," *Analysis*, vol. 43, no. 3, pp. 134–140.