Counterpossibles

Timothy Williamson
University of Oxford, Oxford, U.K
timothy.williamson@philosophy.ox.ac.uk

Abstract

According to orthodoxy, all counterpossibles (counterfactual conditionals with impossible antecedents) are true, or at least not false. Nevertheless, some counterpossibles look false. The problem is not just how best to tidy up an unimportant corner of the logic and semantics of counterfactuals. It has significant theoretical and methodological ramifications in several directions. This paper defends the orthodox view against some recent objections, and explains the most recalcitrant appearances to the contrary by our pre-reflective reliance on a fallible heuristic in the assessment of counterfactuals.

1 What is at Stake

Typically, we use counterfactuals to talk about what would have happened if something had been different from how it actually was. Still, despite the etymology, a counterfactual may have a true antecedent; 'If she were depressed, that would explain her silence' does not imply that she is not depressed. But a counterpossible is a counterfactual whose antecedent is impossible, and therefore false.

What kind of impossibility is relevant? It is not epistemic. For consider this counterfactual:

(1). If thinking had never occurred, science would have flourished.

The antecedent of (1) is epistemically impossible, because it is incompatible with something we know: that we think. But that does not make the antecedent of (1) impossible in the relevant sense. Presumably, the universe could have been lifeless and thoughtless forever. In that case, science would *not* have flourished. Defenders of orthodoxy should agree that (1) is false. The special theoretical problem in evaluating 'If this had been so, that would have been so' arises when this *could not have been* so. The relevant sort of possibility is objective rather than epistemic or subjective. Moreover, what matters is the most inclusive sort of objective possibility, which we may call *metaphysical possibility*. For the special theoretical problem in evaluating a counterfactual does not arise when, although the antecedent could not *easily* have been so, it could still have been so.

It is convenient, though not crucial, to put the problem in terms of possible worlds. We take the worlds to be *possible* in the sense that it is metaphysically possible for any one of them to have obtained; metaphysical possibility as just explained is an appropriately inclusive standard for present purposes. The evaluation of the counterfactual $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ depends on the truth-value of β at relevant possible worlds at which α is true. But what happens if β is true at *no* possible worlds?

We may equate the *intension* $|\alpha|$ of a sentence α (in a context C) with the set of possible worlds at which α is true (in C). In a compositional intensional semantics of the usual type for counterfactuals, the intension of a counterfactual is a function of the intensions of its antecedent and consequent:

(2). $|\alpha \mathrel{\Box\!\!\rightarrow} \beta| = f(|\alpha|, |\beta|)$

Indeed, we can surely be more specific, for all that should matter about the consequent is at which possible worlds *where the antecedent is true* the consequent is also true, in other words, the intersection of the intension of the antecedent with the intension of the consequent. If so, (2) implies (3):

(3). $|\alpha \mathrel{\Box\!\!\rightarrow} \beta| = f(|\alpha|, |\alpha| \cap |\beta|)$

The truth-value of the consequent at worlds where the antecedent is false should be irrelevant to the truth-value of the conditional, for it concerns only what hold if its antecedent held. But (3) yields (4):

(4). If $|\alpha| = \{\}$ then $|\alpha \mathrel{\Box\!\!\rightarrow} \beta| = f(\{\}, \{\})$

Given (4), all counterpossibles have the same intension: they are indiscriminate. That is not yet to decide between making them all true and making them all false. However, we surely want any counterfactual (counterpossible or not) whose consequent merely repeats its antecedent to be a trivial necessary truth, true throughout the set of all possible worlds W:

(5). $|\alpha \mathrel{\Box\!\!\rightarrow} \alpha| = W$

Together, (4) (with $\beta = \alpha$) and (5) require $f(\{\}, \{\})$ to be W. Putting that back into (4), we get:

(6). If $|\alpha| = \{\}$ then $|\alpha \mathrel{\Box\!\!\rightarrow} \beta| = W$

In other words, all counterpossibles are necessary truths, and so truths. This is just the conclusion reached by Stalnaker, Lewis, and their successors in the mainstream of intensional semantics.

Similar arguments can be made in the modal object-language, without reference to worlds. For example, instead of (2) we can just require that counterfactuals with necessarily equivalent antecedents and necessarily equivalent consequents are themselves necessarily equivalent:

(7). $(\Box(\alpha \equiv \alpha^*) \,\&\, \Box(\beta \equiv \beta^*)) \supset \Box((\alpha \mathrel{\Box\!\!\rightarrow} \beta) \equiv (\alpha^* \mathrel{\Box\!\!\rightarrow} \beta^*))$

A plausible auxiliary assumption to complete the argument is simply that conjunctions necessarily counterfactually imply their conjuncts (if this and that were so, this would be so):

(8). $\Box((\alpha \wedge \beta) \mathrel{\Box\!\!\rightarrow} \alpha)$

Together, (7) and (8) imply (9), the analogue in the object-language of (6):

(9). $\Box\neg\alpha \supset \Box(\alpha \mathrel{\Box\!\!\rightarrow} \beta)$

A much simpler argument in the object-language for the truth of counterpossibles just uses a plausible and attractive assumption linking metaphysical modality to counterfactuals. It is that strict implication materially implies counterfactual implication:

(10). $\Box(\alpha \supset \beta) \supset (\alpha \mathrel{\Box\!\!\rightarrow} \beta)$

For, by elementary modal logic, an impossibility strictly implies anything:

(11). $\Box\neg\alpha \supset \Box(\alpha \supset \beta)$

By transitivity, (10) and (11) entail (12):

(12). $\Box\neg\alpha \supset (\alpha \mathrel{\Box\!\!\rightarrow} \beta)$

If one assumes the necessitation of (10), one can also derive the necessitation of (12).

One can use (10) in deriving equivalents of metaphysical modalities in counterfactual terms, as part of an argument for understanding our cognitive capacities for handling metaphysical modalities as a by-product of our cognitive capacities for handling counterfactual conditionals. In such ways, issues about counterpossibles have significant knock-on effects for more general philosophical questions.

Thus strong theoretical pressures push towards orthodoxy about counterpossibles. It is required by the standard simple and natural approach to the semantics of counterfactuals, and it contributes to a simple and natural picture of how counterfactuals and metaphysical modality fit together. Nevertheless, those pressures are not obviously irresistible. If one is willing to countenance impossible worlds in addition to possible worlds, one might be able to retain the world-based semantic framework for counterfactuals while rejecting orthodoxy about counterpossibles, as Nolan, Brogaard and Salerno, and Kment do. For if the semantic value of a counterfactual is sensitive to its behaviour at impossible worlds, that makes it easier to deny assumptions such as (2), (7), and (10). Alternatively, one might seek a different semantic framework for counterfactuals that is less congenial to orthodoxy about counterpossibles.

The focus of resistance to orthodoxy about counterpossibles is usually on alleged counterexamples. Here is one (from Nolan). What are the truth-values of (13) and (14)?

(13). If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared.

(14). If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would not have cared.

If one responds in a theoretically unreflective way, the natural snap answers are presumably that (13) is false and (14) true. The sick children in South America were in no position to know about Hobbes's secret reasoning thousands of miles away; even if they had known, they had more urgent things to care

about. But, as we know, the shared antecedent of (13) and (14) is metaphysically impossible. Therefore, according to orthodoxy, (13) and (14) alike are true. According to the critics, (13) is a counterexample to orthodoxy: a false counterpossible. Such examples can be multiplied.

The temptation to deny (13) and similar counterpossibles is strong. But such inclinations are not always veridical. For the time being, we may treat them as defeasible evidence against orthodoxy.

For some metaphysicians, rejecting orthodoxy also has more theoretical attractions. Here is an example. Nominalists crave the scientific advantages that platonists gain from quantifying over numbers and other abstract objects. How to emulate them? A common strategy, in this and similar cases, is *fictionalist*. One treats the envied rival metaphysical theory as a useful fiction. The proposal deserves to be taken seriously only if accompanied by a properly worked-out account of how reasoning on the basis of a fiction can nevertheless be a reliably truth-preserving way of getting from non-fictional premises to a non-fictional conclusion. For instance, if one reasons validly from true premises purely about concrete reality plus a false (by nominalist lights) auxiliary mathematical theory about abstract objects to a conclusion purely about concrete reality, the conclusion needs to be true too. But why should it be true? One way of implementing the fictionalist strategy is to use counterfactuals. The nominalist reasons in effect about *how things would be if the mathematical theory were to obtain and concrete reality were just as it actually is*. Thus the conclusion corresponds to this counterfactual:

(15). $(M \wedge A) \boxrightarrow C$

Here M is the platonist mathematical theory, A says that concrete reality is just as it actually is, and C says something specific purely about concrete reality. Thus, the truth of the counterfactual seems to guarantee the truth of its consequent, even though its antecedent is false (by nominalist lights), because the relevant counterfactual worlds are the same as the actual world with respect to concrete reality, which C is purely about. The trouble is that the nominalist may well regard platonism as not just *false* but *metaphysically impossible*: for instance, the structure of the hierarchy of pure sets (if any) seems to be a metaphysically non-contingent matter. For such a nominalist, M is impossible, so the counterfactual (15) is a counterpossible. But, given orthodoxy about counterpossibles, the impossibility of the antecedent guarantees the truth of the counterpossible, irrespective of its consequent, so the mere truth of (15) is insufficient for the truth of C. Fictionalists who implement their strategy by means of counterfactuals and regard the rival metaphysical theory as a useful but impossible fiction have therefore been compelled to deny orthodoxy about counterpossibles.

We can be a little more explicit about the relation between the move from (15) to C, on one hand, and orthodoxy about counterpossibles, on the other. The natural route from (15) to C is this. Suppose that C is (actually) false. By hypothesis, A says that concrete reality is just as it actually is, and C says something specific purely about concrete reality, hence $\neg C$ does too. Therefore, we can treat $\neg C$ as part of what A in effect says. Thus the opposite counterfactual surely holds:

(16). $(M \wedge A) \boxrightarrow \neg C$

If (16) excludes (15) we can therefore derive $\neg$(15) from $\neg C$, and so C from (15) by contraposition. Conversely, we can derive (15) from C just as we derived (16) from $\neg C$ (without relying on the mutual exclusion of counterfactuals). But orthodoxy rejects the assumption that (15) and (16) exclude each other, for both are true if their shared antecedent is impossible.

Most orthodox theorists will hold that opposite counterfactuals such as (15) and (16) are compatible *only* if they are counterpossibles. For they are likely to accept the following two principles in the logic of counterfactuals. First, counterfactuals distribute over conjunction in the consequent:

(17). $(\alpha \boxrightarrow (\beta \wedge \gamma)) \equiv ((\alpha \boxrightarrow \beta) \wedge (\alpha \boxrightarrow \gamma))$

This holds on any standard semantics for counterfactuals. Second, no metaphysical possibility counterfactually implies a metaphysical impossibility:

(18). $(\alpha \boxrightarrow \beta) \supset (\Diamond\alpha \supset \Diamond\beta)$

If something is impossible which would obtain if something else were to obtain, then the other thing is impossible too. From (17) and (18) we can easily derive that the conjunction of opposite counterfactuals implies the impossibility of their antecedent:

(19). $((\alpha \mathrel{\Box\!\!\rightarrow} \beta) \wedge (\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta)) \supset \neg\Diamond\alpha$

Even opponents of orthodoxy about counterpossibles may grant (19), since their unorthodoxy may be confined to counterpossibles, and a counterexample to (19) would require a possible antecedent. What opponents of orthodoxy reject, and proponents accept, is the converse of (19).

2 Misconceptions about orthodoxy

In a recent critique of orthodoxy, Berit Brogaard and Joe Salerno characterize their target thus: "Counterpossibles are trivial on the standard account. By 'trivial', we mean *vacuously true and semantically uninformative*. Counterpossibles are *vacuously true* in that they are always true; an impossibility counterfactually implies anything you like. And relatedly, they are *uninformative* in the sense that the consequent of a counterpossible makes no contribution to the truth-value, meaning or our understanding of the whole." Of this conjunction, orthodoxy as characterized above corresponds only to the first conjunct, the claim of vacuous truth. Brogaard and Salerno handle even that conjunct somewhat oddly. For instance, they say of the counterpossible (14) above: 'The intuition is that [(14)] is true, but non-vacuously'. By their own definition, the non-vacuous truth of (14) consists only in its truth, which both sides acknowledge, and the falsity of at least one other counterpossible, such as (13). Thus the relevant 'intuition' is not directed at (14) at all, but at some other counterpossible. However, that is a minor point compared to their inclusion of the second conjunct, semantic uninformativeness. For we need to be quite clear that semantic uninformativeness is no part whatsoever of the standard account. Consequently, that counterpossibles are trivial in Brogaard and Salerno's sense is no part whatsoever of the standard account.

To see this, we must recall that the standard view of counterfactuals is that, as usual for complex sentences, they have a compositional semantics. Their meanings are built up out of the meanings of their constituents. On the standard view, as found in authors such as Stalnaker and Lewis, the meaning of the counterfactual $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ is built up out of the meanings of the sentences α and β combined with the meaning of the counterfactual operator $\Box\!\!\rightarrow$. Many linguists, notably Kratzer (2012), have a subtler view of the semantic structure of conditional sentences, but the points to be made here can be transposed to such alternative settings. On a fine-grained conception of meaning, any difference in meaning between the sentences β and γ makes a difference in meaning between the counterfactuals $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ and $\alpha \mathrel{\Box\!\!\rightarrow} \gamma$, whatever the meaning of α. That applies just as much when α is impossible as when α is possible. For instance, the counterpossibles (13) and (14) differ by a 'not' in the consequent, which *ipso facto* makes a difference in meaning between (13) and (14). Thus it is just false that, on the standard view, 'the consequent of a counterpossible makes no contribution to the […] meaning […] of the whole'.

Equally objectionable is Brogaard and Salerno's claim that, on the standard view, 'the consequent of a counterpossible makes no contribution to […] our understanding of the whole'. For instance, consider these two counterfactuals:

(20). If Plato had been identical with Socrates, Plato would have been snub-nosed.

(21). If Plato had been identical with Socrates, 2 + 2 would have been 5.

We may assume that, by the necessity of distinctness, since Plato and Socrates are distinct, it is metaphysically impossible for Plato and Socrates to have been identical. Thus both (20) and (21) are counterpossibles. Nevertheless, we understand them by understanding their constituents and how they are put together. For instance, a failure to understand the constituent 'snub-nosed' prevents one from fully understanding (20), but does not prevent one from understanding (21). Thus, on the standard view, our understanding of the consequent of a counterpossible does contribute to our understanding of the whole counterpossible. Of course, if one happens to *know* that it is metaphysically impossible for Plato to have been identical with Socrates, and one accepts orthodoxy about counterpossibles, then one can work out that (20) is true even if one does not understand 'snub-nosed', but that is just an instance of the general point that one can know that a sentence states a truth without knowing what it states. For example, a trustworthy and trusted native speaker of Mandarin might utter a sentence of Mandarin and

tell me that it states a truth without telling me what truth it states. In any case, someone can understand (20) and (21) without knowing that it is impossible for Socrates to have been identical with Plato. Having spent too much time reading dodgy webpages, he might suspect that Plato *was* identical with Socrates. Alternatively, he might know that Plato was distinct from Socrates, but doubt the necessity of distinctness on faulty metaphysical grounds. In general, one can understand a counterpossible without knowing it to be a counterpossible, and one's understanding of it is relevantly like one's understanding of other counterfactuals. All these points arise naturally within the framework of a compositional approach to semantics, such as standard accounts assume.

What of the claim that, on the standard account, 'the consequent of a counterpossible makes no contribution to the truth-value […] of the whole'? At first sight, it looks more defensible, since truth-value is a more coarse-grained feature than either meaning or understanding. However, their claim about truth-values is unwarranted too. The only basis for making it is that, according to standard views, all counterfactuals with impossible antecedents have the same truth-value, because all are true, irrespective of their consequent. But, equally, according to standard views, all counterfactuals with *necessary consequents* have the same truth-value, because all are true, irrespective of their antecedent:

(22). $\Box\beta \supset (\alpha \mathrel{\Box\!\!\rightarrow} \beta)$

Both principles, (12) and (22), are corollaries of the quite general entailment (10) from any strict implication to the corresponding counterfactual; one can also derive the semantic analogue of (22) from (3) and (5). Thus, if the standard account implies that the consequent of a counterfactual with an impossible antecedent makes no contribution to the truth-value of the whole, by parity the standard account also implies that the antecedent of a counterfactual with a necessary consequent makes no contribution to the truth-value of the whole. But that combination is absurd. For consider a counterfactual such as (23) with an impossible antecedent *and* a necessary consequent:

(23). If 6 were prime, 35 would be composite.

By Brogaard and Salerno's style of reasoning, the standard account would imply that *neither* the antecedent *nor* the consequent of (23) makes any contribution to the truth-value of (23). That is absurd because, without its antecedent and consequent, all that is left of (23) is the bare counterfactual construction alone, which by itself certainly does not determine a truth-value. Obviously, standard theories of counterfactuals such as Stalnaker's and Lewis's have no such ridiculous consequence. Thus even Brogaard and Salerno's claim that, on the standard account, the consequent of a counterpossible makes no contribution to the truth-value of the whole is unwarranted. Such examples also tell in a parallel way against their already rejected claims that, on the standard account, the consequent of a counterpossible makes no contribution to the meaning and our understanding of the whole.

It is thus a misunderstanding of orthodoxy to suppose that it makes counterpossibles semantically uninformative or cognitively trivial. It simply makes them true. The misunderstanding is the source of many objections to orthodoxy. One such style of objection is this. According to orthodoxy, (24) is true, because Fermat's Last Theorem is a necessary truth:

(24). If Fermat's Last Theorem were false, 2 + 2 would be 5.

The critic then points out, correctly, that Andrew Wiles could not have simplified his famous proof by merely invoking (24) and thence deducing Fermat's Last Theorem by *reductio ad absurdum*. This does indeed refute the claim that (24) is uninformative or trivial, for given the latter claim it is harmless to rely on (24) in a proof. But it is hopeless as an argument against the claim that (24) is true, for the mere truth of a claim does not permit one to rely on it in a *proof*. For that, the claim must have some epistemically appropriate property: it must be an axiom, or have been already proved, or follow from previous steps in a way clear to expert mathematicians, or something like that. Since (24) has no such epistemically appropriate property, it offers no simplification of Wiles's proof. Thus the objection fails. More generally, assertibility requires some epistemically appropriate status, such as being known by the asserter, for which truth is insufficient. That point applies just as much to counterpossibles as to sentences of any other kind, and fits well with orthodoxy. Failure to appreciate it presumably comes from the confused idea that orthodoxy makes counterpossibles uninformative or trivial.

A subtler misconception about orthodoxy concerns speakers who know the impossibility of the antecedent. Consider (25):

(25). If Hobbes had squared the circle, he would have become Lord Chancellor.

I know that the antecedent of (25) is impossible; given orthodoxy, I know that (25) is true. Epistemically, I am in a position to assert (25) on those grounds. But if I do so in a discussion of seventeenth century English politics, something is obviously amiss. However, that point does not tell against orthodoxy, for orthodoxy can easily explain what is amiss. Given orthodoxy, I was also in a position to assert the more informative and equally relevant (26) instead:

(26). Hobbes could not have squared the circle.

Of course, (25) mentions political matters while (26) does not, but my grounds for asserting (25) make the mention factitious and misleading. Therefore, I should have asserted (26) — or, better, just kept quiet — instead, on Gricean grounds of conversational cooperation. Since I did not, my hearers may assume that I asserted (25) because I knew of some politically significant connection between squaring the circle and the Lord Chancellorship, and so be misled. If my hearers correctly identify my grounds for asserting (25), they will recognize the irrelevance of my contribution. Orthodoxy has no trouble in dealing with such cases.

In order to keep one's grip on the implications of orthodoxy, a salutary comparison is between the vacuous truth of counterpossibles and the vacuous truth of empty universal generalizations. The impossibility of the antecedent corresponds to the emptiness of the subject term. For it is widely agreed that 'Every N Vs' is true if and only if the extension of N is a subset of the extension of V. Thus, as a special case, if the extension of N is empty, it is a subset of the extension of V, whatever V is, so 'Every N Vs' is true. Consequently, since there are no golden mountains, 'Every golden mountain is a valley'. It would be obviously absurd to claim that, on this standard account of the universal quantifier, the predicates make no contribution to the truth-value, meaning or our understanding of such sentences. For universal generalizations have the same overall compositional semantic structure whether the subject term is empty or not, just as counterfactual conditionals have the same overall compositional semantic structure whether the antecedent is impossible or not. Similarly, it would be absurd to claim that, on the standard account, a sentence like 'Every golden mountain is in Africa' is cognitively trivial or semantically uninformative. One can understand it without knowing that there are no golden mountains.

Our reactions to counterpossibles are often similar to our reactions to analogous vacuous universal quantifications. In the latter case, we have learnt to override our immediate reactions. Perhaps we should learn to override our immediate reactions to counterpossibles in a similar way.

3 Counterfactual reasoning by reductio ad absurdum

The prime specimens of useful reasoning from an impossible supposition are arguments by *reductio ad absurdum* in mathematics. When we state them in everyday terms, it is natural to use counterfactual conditionals. Lewis gives these examples:

(27). If there were a largest prime p, $p! + 1$ would be prime.

(28). If there were a largest prime p, $p! + 1$ would be composite.

They summarize the classic proof that there is no largest prime: (27) holds because if p were the largest prime, $p!$ would be divisible by all primes (since it is divisible by all natural numbers up to p), so $p! + 1$ would be divisible by none; (28) holds because $p! + 1$ is larger than p, and so would be composite if p were the largest prime. To complete the proof, one can use Lewis's principle of Deduction within Conditionals to conjoin the consequents of (27) and (28):

(29). If there were a largest prime p, $p! + 1$ would be both prime and composite.

Since the consequent of (29) is a contradiction, one can deny the antecedent, and conclude that there is no largest prime.

Of course, one does not strictly *need* to formulate the proof in terms of counterfactual conditionals. One could use material conditionals instead, because all standard mathematical reasoning can be

formalized in purely extensional terms. Nevertheless, it is surely legitimate, indeed natural and appropriate, to use counterfactual conditionals. They nicely convey the role of the antecedent in the reasoning. At the very least, on a good semantic theory, the counterpossibles (27)-(29) should come out *true*, for they are soundly based on valid mathematical reasoning.

Consider any non-obvious impossibility α that can be shown, by more or less elaborate deductive reasoning, to lead to an obvious impossibility ω. The general anti-orthodox strategy is to be charitable by evaluating counterfactuals with α as the antecedent at impossible worlds or situations not closed under such reasoning, precisely in order to falsify counterpossibles such as $\alpha \mathrel{\Box\!\!\rightarrow} \omega$. But those are exactly the counterpossibles one needs to assert in articulating the argument by *reductio ad absurdum* against α. Thus the point generalizes, for instance to the use of counterlogical worlds.

Mathematical arguments by *reductio ad absurdum* are amongst the best arguments for counterpossibles we have. They tell us that if something non-obviously impossible were the case, something obviously impossible would be the case. We should accept the conclusions of those mathematical proofs. They provide strong evidence for orthodoxy. But how can we explain away the strongest evidence *against* orthodoxy, all the seemingly clear examples of false counterpossibles?

4 An error theory of apparently false counterpossibles

Processing a non-obvious counterpossible typically *feels* very like processing a non-counterpossible counterfactual. Consider (13) above, a good example of a seemingly false counterpossible ('If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared'). What goes on when we process it? In my case, before I consciously apply any theoretical considerations, it is something like this. I imagine Hobbes doing geometry in the secrecy of his room. I ask myself whether sick children in the mountains of South America at the time would have cared. I answer in the negative, because there was no way for them to have known about Hobbes's doings at the time, and even if they had known, they would hardly have cared. In the first instance, I assent to (14), the opposite counterfactual to (13), with the same antecedent but the negation of the consequent ('If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would *not* have cared'). My immediate inclination is then to deny (13), as excluded by (14). So far, the impossibility of the antecedent has played *no role whatsoever*. That is not to deny that I imagine Hobbes (secretly) squaring the circle. In some minimal, vague, unspecific way I do imagine him squaring the circle, but I could imagine him carrying out some genuine geometrical construction in much the same way. Now, in my case, theory kicks in. I remind myself that squaring the circle is impossible, and that opposite counterfactuals may both be true when their shared antecedent is impossible. I therefore countermand my inclination to deny (13).

What this suggests is that, in our unreflective assessment of counterfactual conditionals, we use a simple heuristic along these lines:

(HCC). Given that β is inconsistent with γ, treat $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ as inconsistent with $\alpha \mathrel{\Box\!\!\rightarrow} \gamma$

For instance, 'Sick children in the mountains of South America at the time cared' is obviously inconsistent with 'Sick children in the mountains of South America at the time did not care' (on the relevant readings), so in accordance with (HCC) we treat (13) as inconsistent with (14). Thus, having verified (14), we treat ourselves as having falsified (13). More generally, when drawing out the implications of a counterfactual supposition α, as soon as we have accepted $\alpha \mathrel{\Box\!\!\rightarrow} \gamma$, we take ourselves to be in a position to reject $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ for any β inconsistent with γ.

For many purposes, we can consider a simpler heuristic in place of (HCC):

(HCC*). If you accept one of $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ and $\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta$, reject the other

(HCC*) has the advantage over (HCC) of not using 'inconsistent', a term which could do with some clarification.

(HCC) and (HCC*) are equivalent under a wide range of conditions. First, start with (HCC). Clearly, β is inconsistent with ¬β. Then (HCC) tells you to treat $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ as inconsistent with $\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta$. Thus, if

you accept one of them, you should reject the other. In other words, you should obey (HCC*). Conversely, start with (HCC*). Suppose that you are given that β is inconsistent with γ. Thus γ entails ¬β. So, normally, from α □→ γ you can derive α □→ ¬β, by an informal analogue of Deduction within Conditionals. But (HCC*) tells you not to accept both α □→ β and α □→ ¬β. So, normally, you should not accept both α □→ β and α □→ γ. In other words, you should obey (HCC). However, since it is sometimes artificial to introduce an explicit negation when two sentences are obviously inconsistent, (HCC) may be the more natural heuristic.

There is psychological evidence that people reason in accordance with (HCC*), treating pairs of conditionals with the same antecedent and contradictory consequents as inconsistent, whether the conditionals are indicative or subjunctive. More generally, there is extensive psychological evidence that we tend to evaluate conditionals by evaluating their consequents on the supposition of their antecedents, with only subtle differences in treatment between indicatives and consequents. Thus we tend to treat cases where the antecedent is false as irrelevant to the evaluation. Our assessment of the probability of a conditional is highly correlated with our assessment of the conditional probability of its consequent on its antecedent. The well-supported suppositional model of our evaluation of conditionals predicts that it will conform to both (HCC) and (HCC*). For if β is inconsistent with γ, then it is inconsistent to accept both under the supposition of α. In probabilistic terms, the inconsistency of β with γ implies this relation between their conditional probabilities on α and the conditional probability of their disjunction on α:

(30). $\text{Prob}(\beta|\alpha) + \text{Prob}(\gamma|\alpha) = \text{Prob}(\beta \vee \gamma|\alpha) \leq 1$

Thus if Prob(β|α) is high, Prob(γ|α) is low, and *vice versa*. Given the close correlation between our assessments of conditional probabilities and our assessments of the probabilities of conditionals, this means that if we do not assess both conditionals in (HCC) or (HCC*) as probable.

For the orthodox, (HCC) and (HCC*) are only heuristics because they will lead you to reject true counterpossibles when α is impossible. However, it is plausible that usually, when counterfactual conditionals arise in practice, their antecedents are possible. In that case, (HCC*) will never lead you astray. For if you accept a true one of α □→ β and α □→ ¬β (which is not the responsibility of (HCC*)), (HCC*) will tell you to reject the other one, which will be false by (19). Given the near-equivalence of (HCC) and (HCC*), (HCC) will share much of the qualified reliability of (HCC*). Thus, on an orthodox logic of counterfactuals, both (HCC) and (HCC*) are reasonable though fallible heuristics.

One might wonder whether an unorthodox view of counterfactuals could treat (HCC) and (HCC*) as more than a fallible heuristic, by treating *no* counterpossibles as exceptions; that might even be an advantage of unorthodoxy. But that idea is unlikely to work. For consider counterpossibles with explicit contradictions as antecedents:

(31). $(\alpha \wedge \neg\alpha) \boxrightarrow \alpha$ (32). $(\alpha \wedge \neg\alpha) \boxrightarrow \neg\alpha$

Both (31) and (32) look highly plausible; surely conjunctions counterfactually imply their conjuncts. But if both (31) and (32) hold, then they are consistent, even though they have the same antecedent and inconsistent consequents. Now unorthodox theorists may reject some instances of (31) and (32), for instance when α itself is 'No conjunction counterfactually implies its conjuncts' or the like. But they cannot plausibly reject one of them in *all* cases. For instance, let α be 'The Liar is true', so $\alpha \wedge \neg\alpha$ makes the dialetheist claim about the Liar paradox that the Liar is both true and not true. The dialetheist both asserts that the Liar is true *and* asserts that the Liar is not true. Presumably, therefore, both (31) and (32) should hold on this reading of α, even for the unorthodox. Thus even they should regard (HCC) and (HCC*) as fallible heuristics, not as marking exceptionless rules of the logic of counterfactuals.

Probabilistic considerations point in the same direction, even if we ignore the long series of results, initiated by David Lewis, which show that conditionals cannot in general be identified with propositions whose probability is the conditional probability of the antecedent on the consequent. Under the standard equation of the conditional probability Prob(β|α) with the ratio $\text{Prob}(\alpha \wedge \beta)/\text{Prob}(\alpha)$ of unconditional probabilities, the conditional probability is undefined when Prob(α) is 0, as it is when α is a contradiction. If we treat conditional probabilities as primitive, we can sometimes assign Prob(β|α) a value even when Prob(α) = 0, but we still cannot do so when α holds at *no* point in the probability space, on pain of

violating basic principles of conditional probability. For Prob(β|α) should be 1 whenever α entails β, so for vacuous α Prob(β|α) should be 1 for every β, which violates the principle that Prob(¬β|α) = 1 − Prob(β|α). Thus the structure of probability theory rules out vacuous conditional probabilities. Rejecting standard principles of conditional probability to allow for vacuous conditional probabilities would be a fool's bargain. Thus we should expect (HCC) and (HCC*) to have exceptions for at least some impossible antecedents.

How will the evaluation of α □→ β and α □→ ¬β go when α entails β and ¬β? Since both β and ¬β would eventually emerge as we developed the counterfactual supposition α for long enough, (HCC) and (HCC*) make it a race between the contradictories as to which emerges first. If β emerges first, we accept α □→ β and so reject α □→ ¬β before ¬β has time to emerge. If ¬β emerges first, we accept α □→ ¬β and so reject α □→ β before β has time to emerge. On this view, the proponent of impossible worlds misinterprets this computational difference in terms of the relative closeness of impossible α ∧ β and impossible α ∧ ¬β worlds. One advantage of the heuristics account is that it explains our inattention to the impossibility of the antecedent in our cognitive processing of many counterpossibles. By contrast, accounts such as Brogaard and Salerno's that postulate a special standard of relative closeness for impossible worlds, apparently quite different from that appropriate for possible worlds, fail to explain the lack of felt adjustment to such a special standard in our cognitive processing of counterpossibles.

Of course, we are not completely helpless victims of our heuristics. Through conscious theoretical reflection, we can sometimes inhibit their operation. Our mastery of reasoning by *reductio ad absurdum* in mathematics shows our ability to defeat (HCC) and (HCC*). For example, we accept both the counterpossibles (29) and (30) in the proof that there is no largest prime, even though they have the same antecedent and mutually inconsistent consequents. Even in less formal settings, it is not psychologically compulsory to call off the search for β amongst the counterfactual consequences of α once ¬β has turned up. If we are asked an open-ended question such as 'What would have been the consequences if α had been the case?', we can continue the search in a way that allows for mutually inconsistent counterfactual consequences to emerge. That is in effect what we do when asked 'Could α have been the case?'. Nevertheless, despite our ability to inhibit their operation, (HCC) and (HCC*) remain the default, to which we may always be liable to revert when off our guard. For instance, if one puts aside one's mathematical sophistication, it is not hard to feel that (27) and (28) are mutually inconsistent after all.

A useful analogy, noted above, is with our naïve reactions to vacuously true universal quantifications:

(33). Every dolphin in Oxford has arms and legs.

A natural inclination is to judge (33) false, even if one doubts that there are no dolphins in Oxford. That resistance is explicable by the hypothesis that we accept (34) on the basis of background information about dolphins, and are then inclined to reject (33) as inconsistent with (34):

(34). Every dolphin in Oxford lacks arms and legs.

That suggests heuristics for universal quantification analogous to (HCC) and (HCC*):

(HUQ). Given that φ is inconsistent with ψ, treat 'Every σ φs' as inconsistent with 'Every σ ψs'

(HUQ*). If you accept one of 'Every σ φs' and 'Every σ ¬φs', reject the other

On the standard semantics for the universal quantifier, (HUQ) and (HUQ*) go extensionally wrong when and only when σ is empty in extension.

We can come to recognize the limitations of (HUQ) and (HUQ*) through natural reasoning. For instance, suppose that our rejection of (33) leads us to accept its negation:

(35). Not every dolphin in Oxford has arms and legs.

From (35) we can validly reason to (36), and thence to (37):

(36). Some dolphin in Oxford lacks arms and legs.

(37). There is a dolphin in Oxford.

But we know (37) to be false. That may lead us to realize that (33) is not false, though its utterance may induce a false presupposition. (HUQ) and (HUQ*) are fallible heuristics, defeasible by theoretical reflection, but they are still our default.

There may be a more general cognitive pattern underlying these heuristics. For example, it is plausible that we use analogues of (HCC) and (HCC*) for indicative as well as subjunctive conditionals, and

analogues of (HUQ) and (HUQ*) for generic as well as universal quantifiers. Very roughly, we ignore the empty case. We continue using heuristics based on that principle, even when the empty case is obviously relevant, until we resort to conscious reflection. Indeed, we may tend to use suppositional reasoning in evaluating universal and generic generalizations as well as conditionals. For instance, when asked to evaluate (33) or its generic analogue ('Dolphins in Oxford have arms and legs'), we may suppose that something is a dolphin in Oxford, and ask ourselves whether it has arms and legs.

Our theoretical grasp of universal quantification is currently more secure than it is of counterfactuals conditionals. We are consequently more comfortable in overruling (HUQ) and (HUQ*) than in overruling (HCC) and (HCC*). But it was not always so. Centuries of confusion about the existential import or otherwise of the universal quantifier bear witness to the difficulty of achieving a clear view of the truth-conditions of sentences of our native language formed using the most basic logical constants. Those who take themselves to have provided clear examples of false counterpossibles may be in a similar position to traditional logicians who took themselves to have provided clear examples of false universal generalizations with empty subject terms. Indeed, the primitively compelling nature of heuristics such as (HUQ) and (HUQ*) may have been the main obstacle to achieving a clear view of the truth-conditions of universal generalizations.

Imagine a philosopher attempting to craft a semantics for the universal quantifier to vindicate the heuristically driven judgments that (33) is false while (34) is true. He may invest immense patience and ingenuity in his project, but it is not going to end well. We should be similarly wary of attempts to craft a semantics for the counterfactual conditional to vindicate the heuristically driven judgments that some counterpossibles are false while others are true. There is a danger in semantics of unintentionally laundering cognitive biases into veridical insights, a danger evident in the semantics of generics.

With the universal quantifier, clear understanding was finally achieved through systematic, highly general semantic and logical theorizing, rather than by a more data-driven approach. The same may well hold for the counterfactual conditional. At any rate, it is methodologically naïve to take the debate over counterpossibles to be settled by some supposed examples of clearly false counterpossibles. As we have seen, a simple and mostly reliable heuristic would lead us to judge them false even if they were true.

On the view developed here, our assessments of counterfactuals are often based on fallible heuristics such as (HCC) and (HCC*). How far should that view make us sceptical more generally about reliance on pre-theoretic assessments of counterfactuals in philosophy, semantics and elsewhere?

The heuristics are reliable over wide ranges of cases. Just as we can gain lots of perceptual knowledge by relying on perceptual heuristics that are reliable over wide ranges of cases but fail under special conditions, so we can gain lots of modal knowledge by relying on heuristics such as (HCC) and (HCC*). Blanket scepticism is not a sensible response. Moreover, many assessments of counterfactuals will not rely on (HCC) and (HCC*) at all. (HCC) and (HCC*) are fundamentally devices for moving from the acceptance of one unnegated counterfactual to the rejection of another. They are not devices for accepting unnegated counterfactuals in the first place. Even if rejecting the latter counterfactual depends on the heuristic, accepting the former need not. Arguably, the key judgments in thought experiments, for instance that in such-and-such a Gettier case the subject would not know, involve the acceptance of unnegated counterfactuals, for which (HCC) and (HCC*) are not needed. Moreover, nothing said here impugns the reliability of counterfactual judgments made on the basis of mathematical reasoning. Even without theoretical reflection, we can inhibit the operation of the heuristics, as we sometimes need to do in order to maintain consistency. For instance, as already noted, we can continue the imaginative search for counterfactual consequences of a subjunctive supposition in an open-minded way that allows contradictions to arise. Nevertheless, the theorist who overrides the heuristic in favour of more reflective considerations should expect to feel some residual unease, at least at first. No matter how cogent the reflective considerations, the heuristic is too stupid to understand them; instead, it just goes on blindly pressing to have its way. If our access to the logic and semantics of our own language is essentially mediated by fallible heuristics, true theories may always feel Procrustean to us.