Counterpossibles in Semantics and Metaphysics

Timothy Williamson

If whales were fish, their behaviour would differ from what it actually is. If whales were fish, their behaviour would be just as it actually is. Those sound like genuine alternatives. Yet, since whales are by nature mammals, they presumably *could not* have been fish; that would be contrary to their nature. Thus both conditionals are counterpossibles, counterfactual conditionals with impossible antecedents. Semantic orthodoxy makes all counterpossibles true. So the two conditionals were true, and not mutually exclusive after all. Is orthodoxy about counterpossibles correct? The problem is not just how best to tidy up an unimportant little corner of the logic and semantics of counterfactuals. It has significant theoretical and methodological ramifications in several directions. This paper defends orthodoxy against recent objections, and explains recalcitrantly unorthodox appearances by our pre-reflective reliance on a fallible heuristic in assessing conditionals.

1. *What is at stake*

A counterfactual is a conditional sentence like the so-called subjunctive 'If this were so, that would be so' and unlike the indicative 'If this is so, that is so'. 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, so 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.[1] Fortunately, the main issues about counterpossibles are not too sensitive to the precise shade of modality at issue.

Some subjunctive conditionals arguably have an epistemic reading (Edgington 2008, Vetter 2016). This paper is mainly concerned with non-epistemic readings of subjunctive conditionals, though some of its conclusions may generalize to epistemic readings too. It is sometimes hard to tell whether one is reading a conditional epistemically or not; the reader may wish to bear that issue in mind.

One view is that counterpossibles suffer some sort of presupposition failure (von Fintel 1998). If the presupposition were semantic, they might even lack a truth-value. However, that is too drastic. In explaining how one can tell that whales are not fish, someone might say 'If whales were fish, they would behave quite differently'. That utterance is felicitous even if it is common ground in the conversation that whales could not really have been fish. Of course, once it is common ground that the antecedent of a counterfactual is impossible, the conditional will often lose any conversational relevance it might previously have had, and so be infelicitous to utter. Thus there may be a defeasible pragmatic presupposition that the antecedent is possible, but it would not enter the semantics.

In discussing the semantics of the subjunctive conditional, I will treat it as a sentential operator, making a complex sentence $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ out of simpler sentences α and β. That may well be an over-simplification. On Kratzer's influential account (2012), 'if' restricts other sentential operators rather than being one in its own right. The arguments of this paper can be transposed to that more sophisticated setting.

It is convenient, though not crucial, to put the problem of counterpossibles in terms of possible worlds. We read 'possible' as *metaphysically possible*, the relevant standard for present purposes. The orthodox 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 α is true at *no* possible worlds?

The classic semantic accounts of counterfactuals, going back to Stalnaker (1968) and Lewis (1973), use the framework of compositional intensional semantics, and more specifically of possible worlds semantics (for simplicity in what follows, we hold fixed parameters such as the context and the assignment of values to variables, and leave them tacit). In this setting, we can explain why it is so natural to make all counterpossibles true. It is no optional peculiarity of Stalnaker's semantics, or Lewis's.

We equate the *intension* $|\alpha|$ of a sentence α with the set of possible worlds at which α is true. By hypothesis, the intension of a counterfactual is a function of the intensions of its antecedent and consequent:

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

We can 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. That is just the restricting effect of the antecedent. On that assumption, (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 would hold if its antecedent held. But if the antecedent is true at no possible worlds, then the consequent is true at no possible worlds at which the antecedent is true; (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 does not yet decide between making them all true and making them all false. However, we 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.[2]

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):[3]

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

If α *could not* hold without β, α *would not* hold without β. Then, 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).

Elsewhere, I have used (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 (Williamson 2007, pp. 156-77). This is one of several ways in which 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 (Nolan 1997, Brogaard and Salerno 2013, Kment 2014). 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).[4] Alternatively, one might seek a different semantic framework for counterfactuals that is less congenial to orthodoxy about counterpossibles. The issues must be examined in more depth.

The focus of resistance to orthodoxy about counterpossibles is usually on alleged counterexamples. Here is one (based on Nolan 1997). 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 that (13) is false and (14) true. The sick children in the mountains of South America were in no position to know about Hobbes's secret geometrical reasoning thousands of miles away; even if they had known, they had far more urgent things to care about. But, as we now know, the circle cannot be squared. The shared antecedent of (13) and (14) is metaphysically impossible. So orthodoxy makes (13) and (14) alike true. Thus, critics allege, (13) is a counterexample to orthodoxy: a false counterpossible. Such examples can be multiplied without limit.

The temptation to deny (13) and many similar counterpossibles is strong. But such inclinations are not always veridical. For the time being, we may treat them as defeasible evidence against orthodoxy about counterpossibles. Later, we will assess the prospects for an error theory about those inclinations.

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 (Field 1980). 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) $(M \wedge A) \mathrel{\square\!\!\rightarrow} C$

Here *M* is the platonist mathematical theory, *A* says that concrete reality is just as it actually is, and *C* says something 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.[5] 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 truth of (15) is insufficient for the truth of *C*. Fictionalists who implement their strategy with by counterfactuals and regard the rival metaphysical theory as a useful but impossible fiction have therefore been compelled to deny orthodoxy about counterpossibles (for instance, Dorr 2008).

We can articulate the relation between the move from (15) to *C* and orthodoxy about counterpossibles. The natural route from (15) to *C* is this. Suppose that *C* is (actually) false. Since *A* says that concrete reality is just as it actually is and *C* says something purely about concrete reality, ¬*C* does too. Hence ¬*C* is in effect part of what *A* says. Thus the opposite counterfactual holds:

(16) $(M \wedge A) \mathrel{\square\!\!\rightarrow} \neg C$

If (16) entails the negation of (15) we can therefore derive ¬(15) from ¬*C*, and so *C* from (15) by contraposition. Conversely, we can derive (15) from *C* just as we derived (16) from ¬*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 hold that opposite counterfactuals such as (15) and (16) are both true *only* if they are counterpossibles.[6] For they accept the following two principles. First, counterfactuals distribute over conjunction in the consequent:

(17) $(\alpha \mathrel{\Box\!\!\rightarrow} (\beta \wedge \gamma)) \equiv ((\alpha \mathrel{\Box\!\!\rightarrow} \beta) \wedge (\alpha \mathrel{\Box\!\!\rightarrow} \gamma))$

This holds on any standard semantics for counterfactuals. Second, no metaphysical possibility counterfactually implies a metaphysical impossibility:[7]

(18) $(\alpha \mathrel{\Box\!\!\rightarrow} \beta) \supset (\Diamond\alpha \supset \Diamond\beta)$

If something is impossible which would obtain if something else were to obtain, then the latter is impossible too. From (17) and (18) we can easily derive that the conjunction of opposite counterfactuals implies the impossibility of their antecedent:[8]

(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).

The apparently minor issue of counterpossibles is thus interrelated with various more central questions in both semantics and metaphysics.[9] However, the question should be separated from some other issues with which it is surprisingly often confused. That is the business of the next section.

2. *Misconceptions about orthodoxy*

In a recent critique of orthodoxy, Berit Brogaard and Joe Salerno characterize their target thus (2013, p. 642):

> 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' (2013, pp. 642-643). 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).[10] 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, recall that on the standard view of counterfactuals they have, as usual for complex sentences, a compositional semantics. Their meanings are built up out of the meanings of their constituents. For authors such as Stalnaker (1968) and Lewis (1973), the meaning of the counterfactual sentence $\alpha \mathrel{\square\!\!\rightarrow} \beta$ is built up out of the meanings of the sentences α and β combined with the meaning of the counterfactual operator $\mathrel{\square\!\!\rightarrow}$. 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 $\alpha \mathrel{\square\!\!\rightarrow} \beta$ and $\alpha \mathrel{\square\!\!\rightarrow} \gamma$, whatever the meaning of α.[11] 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 them 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 contributes 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. 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.

Of course, the informativeness of a sentence does not supervene purely on its intension in a standard semantic framework. For '2 + 2 =4' and a sentence expressing Fermat's last theorem have the same intension, the set of all possible worlds, even though the latter is more informative than the former. Informativeness is sensitive to *how* the given intension is expressed. But that point does not depend on one's view of counterpossibles.

Here is another way to verify the informativeness of counterpossibles, on the standard view. Suppose that someone is initially uncertain whether Plato is identical with Socrates. Then someone whom he trusts utters (21). Since he knows that its consequent is impossible (2 + 2 could not have been 5), he deduces (using (18)) that the antecedent is impossible too, and thereby comes to know that Plato could not have been identical with Socrates. He has gained useful information from (21).

What of 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'? 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 implies that *neither* the antecedent *nor* the consequent of (23) makes any contribution to the truth-value of (23). That is absurd because, without its specific antecedent and consequent, all that is left of (23) is the bare form of a counterfactual sentence 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 thus betrays a grave 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, including many (though not all) of Brogaard and Salerno's.[12] 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.

Of course, in terms purely of possible worlds, considering a counterpossible involves supposing that a member of the empty set of worlds obtains, which looks like a waste of time. But questions of uninformativeness and triviality are cognitive and computational, sensitive not only to sets of worlds but also to the linguistic guises under which they are presented. They may conceal or reveal the impossibility of the antecedent. Whether the hearer consciously accepts the vacuous truth of counterpossibles also makes a difference to their informativeness. The mere fact that a counterfactual conditional has an impossible antecedent tells us very little about its cognitive or computational status. For that, we need to know the linguistic guise of the antecedent.

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 (Grice 1989, pp. 26-28). 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, as Lewis knew (1973, p. 25).

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.[13] 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, (27)-(29) are true, and since there are no unicorns, (30)-(32) are true:

(27) Every golden mountain is a mountain.

(28) Every golden mountain is in Africa.

(29) Every golden mountain is a valley.

(30) Every unicorn is a mammal.

(31) Every unicorn is gentle.

(32) Every unicorn is a fish.

It would be absurd to claim that, on this standard account of the universal quantifier, the predicates in (27)-(32) ('is a mountain', 'is in Africa', 'is a valley', 'is a mammal', 'is gentle', 'is a fish') make no contribution to the truth-value, meaning or our understanding of (27)-(32) respectively. 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. For the same reason, it would be absurd to claim that, on the standard account, (27)-(32) are cognitively trivial or semantically uninformative. After all, one can understand (27)-

(29) without knowing that there are no golden mountains, and one can understand (30)-(32) without knowing that there are no unicorns.[14]

Just as we cannot simplify the proof of Fermat's Last Theorem by invoking (24), so we cannot simplify its proof by invoking (33):

(33) Every counterexample to Fermat's Last Theorem is a number both identical to 2 + 2 and identical to 5.

Although (33) is true, we cannot rely on it in a proof without having already proved it, any more than we can do so in the case of (24). Again, knowing the truth of (34) because one knows that no Englishman squared the circle does not make it conversationally appropriate to assert (34) in a discussion of seventeenth century English politics:

(34) Every Englishman who squared the circle became Lord Chancellor.

As the preceding examples suggest, our reactions to counterpossibles are often similar to our reactions to similarly constructed universal quantifications. For instance, Brogaard and Salerno write of 'the intuitive falsehood' of (35) (2013, p. 645):

(35) If my shirt had been red and non-red all over, then it would have been green.

A first, unreflective reaction to (35) is indeed sceptical: if it had been red and non-red all over, why should it have been green rather than any other colour? Compare (35) with (36):

(36) Every shirt that is red and non-red all over is green.

A first, unreflective reaction to (36) is equally sceptical: if a shirt is red and non-red all over, why should it be green rather than any other colour? In the case of universal quantification, we have learnt to override such immediate reactions. On reflection, (36) is true because it has no counterexample: no shirt is red and non-red all over and yet not green, because no shirt is red and non-red all over. Perhaps we should learn to override our immediate reactions to counterpossibles like (35) in a similar way. The last section of the paper discusses these issues in more depth.

3. Counterlogicals

David Lewis offered a brief independent argument for the truth of counterlogicals, counterfactuals with logically inconsistent antecedents. In principle, a moderate opponent of orthodoxy might claim that, although there are false counterpossibles, there no false counterlogicals. In practice, however, opponents of orthodoxy are typically less moderate: they claim that some counterlogicals are false. For if one is moved by examples of apparently false counterpossibles, one will probably also be moved by examples of apparently false counterlogicals, such as (37):

(37) If not everything were self-identical, everything would be self-identical.

Similarly, Brogaard and Salerno classify (38) as false (2013, p. 643):

(38) If intuitionistic logic were the correct logic, then the law of excluded middle would still be unrestrictedly valid.

Although (38) is a slightly trickier case, because its antecedent is inconsistent with metalogic rather than the logic of a non-metalinguistic object-language, for present purposes the difference does not matter.

The dialectical setting is this. Opponents of orthodoxy about counterpossibles typically grant classical logical principles for standard logical constants such as the truth-functors, quantifiers, identity, and even modal operators (not including the counterfactual conditional itself). They also grant standard structural principles about logical consequence, such as the cut rule and monotonicity. In that sense, classical logic is not at stake, though if one prefers a non-classical logic, one can have a similar debate about counterpossibles in that setting too. For present purposes, we may simply assume that both sides in the debate over counterpossibles accept classical logic.

Lewis compresses his argument into a single sentence (1973, p. 24):

> [I]t seems that a counterfactual in which the antecedent logically implies the consequent ought always to be true; and one sort of impossible antecedent, a self-contradictory one, logically implies any consequent.

Here is one way of unpacking Lewis's argument. For convenience, we represent his 'logically implies' by the logical truth of a material implication rather than the logical validity of an argument from premises to conclusion; nothing of present importance hangs on the choice. Logical truth is expressed by '⊢'. Suppose that α is 'self-contradictory':

(39) $\vdash \neg\alpha$

Then, by classical logic, α logically implies any consequent γ:

(40) $\vdash \alpha \supset \gamma$

Lewis endorses the plausible principle he calls 'Deduction within Conditionals' (1973, p. 132). It says that what is counterfactually implied by a given antecedent is closed under deduction:[15]

(DC) If $\vdash (\wedge_{i \in I} \beta_i) \supset \gamma$ then $\vdash (\wedge_{i \in I} (\alpha \mathrel{\Box\!\!\rightarrow} \beta_i)) \supset (\alpha \mathrel{\Box\!\!\rightarrow} \gamma)$

(DC) fits well with our practice of making deductions in developing a supposition to reach a counterfactual conclusion; it also yields the logical truth of (17) (the counterfactual conditional commutes with conjunction in the consequent). The present argument requires only the special case of (DC) with just one premise:

(DC1) If $\vdash \beta \supset \gamma$ then $\vdash (\alpha \mathrel{\Box\!\!\rightarrow} \beta) \supset (\alpha \mathrel{\Box\!\!\rightarrow} \gamma)$

If we substitute α for β in (DC1) and combine the result with (40), we derive (41):

(41) $\vdash (\alpha \mathrel{\Box\!\!\rightarrow} \alpha) \supset (\alpha \mathrel{\Box\!\!\rightarrow} \gamma)$

The reflexivity principle that everything counterfactually implies itself is unproblematic, and an axiom of Lewis's preferred logic of counterfactuals (1973, p. 132):

(R) $\vdash \alpha \mathrel{\Box\!\!\rightarrow} \alpha$

From (41) and (R) we derive (42):

(42) $\vdash \alpha \mathrel{\Box\!\!\rightarrow} \gamma$

Thus, by plausible principles of Lewis's logic of counterfactuals, we can derive (42), the theorem that a hypothesis counterfactually implies anything we like, from (39), the assumption that the hypothesis is logically inconsistent.

Lewis's argument and his conclusion require one minor qualification. For suppose that the object-language contains a rigidifying 'actually' operator A, where the sentence $A\alpha$ is true at an arbitrary world in a model if and only if α is true at the designated actual world of the model. Suppose further that logical truth is truth at the actual world of every model.[16] Substitute $\alpha \wedge \neg A\alpha$ for α. Then (43) holds because α and $A\alpha$ have the same truth-value at the designated actual world of any model:

(43) $\vdash \neg(\alpha \wedge \neg A\alpha)$

But we do not in general want (44):

(44) $\vdash (\alpha \wedge \neg A\alpha) \mathrel{\Box\!\!\rightarrow} (\alpha \wedge \neg\alpha)$

For example, if α means 'It is raining', then in effect (44) makes it a logical truth that if it had rained but not in this world, a contradiction would have obtained. The supposed logical truth is false if it could have rained but is not actually doing so. Since (R) is harmless, the trouble is with (DC), and more specifically with (DC1). They do not apply to languages with fancy modal operators such as 'actually' (if logical truth is truth at the actual world of every model). For such circumstances yield contingent logical truths, logically guaranteed to express truths but not logically guaranteed to express necessary truths. The standard Rule of Necessitation in modal logic fails under the same circumstances:

(RN) If $\vdash \alpha$ then $\vdash \Box\alpha$

For example, although (43) holds, (45) fails, because it could have rained in counterfactual but not in actual circumstances:

(45) $\vdash \Box\neg(\alpha \wedge \neg A\alpha)$

For these familiar reasons, (DC) and (DC1) must be restricted in the same way as (RN), for languages with fancy modal operators. In more ordinary cases, contingent logical truths are not involved, the restrictions are satisfied and this problem for Lewis's argument does not arise.[17]

The problem above for Lewis arises when logical implication is insufficient for strict implication. Ironically, in criticizing his argument, Brogaard and Salerno conflate logical

implication with strict implication. They miss not only that problem but also also the moderate option on which there are false counterpossibles but no false counterlogicals, which arises because logical implication is unnecessary for strict implication. They seem to be under the misapprehension that the opponent of orthodoxy about counterpossibles is bound to reject Lewis's argument about counterlogicals (Brogaard and Salerno 2013, pp. 648-9).

Those problems aside, Brogaard and Salerno's response to Lewis's argument is clearly to reject (DC) and (DC1). However, they reject them *only* for counterpossibles. For they say: 'All the typical rules governing counterfactuals are valid, when the antecedent is possible' (2013, p. 657). They treat it as a virtue of their account that it preserves a strong logic of counterfactuals with possible antecedents. However, from an abductive perspective, that puts increased pressure on their reasons for rejecting the standard rules in the first place (at least for simple languages without fancy modal operators such as 'actually'). Those reasons had better be robust enough to justify the sacrifice of a strong and simple theory such as Stalnaker or Lewis's logic of counterfactuals, in favour of one patched up with messy adjustments and repairs. In particular, given that counterfactuals have a compositional semantics, we should be suspicious of the idea that their behaviour is radically different on the rare occasions when the antecedent is impossible. For how are we supposed to have come to be using so oddly contoured a conditional?

4. *Objectivity and opacity*

A useful litmus test for the nature of a semantic construction is this: given co-referential inputs, does it produce co-referential outputs? If yes, it is *transparent*. If no, it is *opaque*. Of course, a proper account would work hard to clarify the relevant conception of reference, but for present purposes we can just use that schematic explanation.

To set up an example, consider 'Hesperus' and 'Phosphorus' as co-referential proper names, and 'Hesperus will explode tonight' and 'Phosphorus will explode tonight' as co-referential sentences. A new expression 'Prob' is introduced. It takes sentences as inputs and outputs complex singular terms for 'probabilities' in some sense. Compare two cases.

Case (i): We are authoritatively told that (46) is sometimes true:

(46) Prob(Hesperus will explode tonight) < Prob(Phosphorus will explode tonight)

Case (ii): We are authoritatively told that (47) is always true:

(47) Prob(Hesperus will explode tonight) = Prob(Phosphorus will explode tonight)

Case (i) implies the opacity of 'Prob'. It goes well with a subjective or epistemic reading on which Prob(α) is the credence or evidential probability of α. An agent may in some relevant sense be less confident that Hesperus will explode tonight than that Phosphorus will

explode tonight, or have less evidence for the former than for the latter. But case (i) goes poorly with an objective reading of 'Prob', on which Prob(α) is the physical chance of α. How can it be physically less probable that Hesperus will explode than that Phosphorus will explode tonight, when those are the very same state of affairs? By contrast, case (ii) is consistent with transparency. It goes poorly with a subjective or epistemic reading of 'Prob', for why should agents always be exactly as confident in 'Hesperus will explode tonight' as in 'Phosphorus will explode tonight', or always have exactly as much evidence for the former as for the latter? But case (ii) goes well with an objective reading of 'Prob', where the identity of the states of affairs forces the identity of their chances.

A natural explanation of the difference is that objective probabilities concern only the states of affairs themselves, whereas subjective and epistemic probabilities are also sensitive to differences in how agents represent those states of affairs — in modes of presentations, or guises, or something like that. The explanation may well need to be refined in all sorts of ways, but it looks to be at least roughly on the right track. At any rate, constructions that concern epistemic or subjective matters invite an opaque reading, while constructions that concern objective, non-epistemic and non-subjective matters do not invite an opaque reading. For simplicity, we will treat the opaque readings in the former case as the intended ones.

Normally, counterfactuals are on the objective side of this contrast. For instance, (48) and (49) should have the same truth-value:

(48) If Hesperus were to explode tonight, so would the Moon.

(49) If Phosphorus were to explode tonight, so would the Moon.

For the antecedents of (48) and (49) suppose the very same state of affairs to obtain. Normally, counterfactuals on a non-epistemic reading are transparent. They are about the very objects, properties, relations, and states of affairs their antecedents and consequents are about, not the ways in which agents represent those objects, properties, relations, and states of affairs.

If non-epistemic counterpossibles are uniform in this respect with other non-epistemic counterfactuals, they too are transparent. Although the objective impossibility of the antecedent might pragmatically trigger an epistemic reading, it is quite implausible that on an unequivocal non-epistemic reading the natural language counterfactual conditional should have been programmed all along to make the radical semantic switch to non-transparent behaviour just for the unusual case of an impossible antecedent. Transparency puts significant limits to the programme of using (presumed) counterpossibles to simulate one metaphysical theory from within another, fictionalist-style. For instance, suppose that on the true metaphysical view T, a mental event *m* is identical with a physical event *p*. On a rival view T*, *m* is distinct from *p*; T* is false, indeed impossible (by the necessity of identity). Proponents of T use counterpossibles to characterize how things are from the point of view of T*; for instance:

(50) T* $\Box\!\!\rightarrow m \neq p$

Of course:

(51) T* $\Box\!\!\rightarrow m = m$

But since $m = p$, both in fact and from the point of view of T, transparency yields (52) from (51):

(52) T* $\Box\!\!\rightarrow m = p$

Thus the impossible but internally coherent metaphysical theory T* still has mutually inconsistent counterfactual implications ($m = p$ and $m \neq p$), even though proponents of T* may be strict adherents of classical logic.

Unsurprisingly, Brogaard and Salerno deny that counterfactuals are transparent.[18] The question is whether substitution of co-referential terms makes a difference in truth-value only for counterpossibles, or for other counterfactuals too. Unfortunately, it is hard to extract a stable answer from their discussion.[19] The dilemma is this. First, suppose that substitution of co-referential terms makes a difference in truth-value only for counterpossibles. Then counterfactuals behave in radically different ways depending on the modal status of their antecedent: transparently, like a non-epistemic operator, if it is possible, opaquely, like an epistemic operator, if it is impossible. That suggests an implausibly hybrid semantics. A more uniform treatment is much to be preferred. Alternatively, suppose that substitution of co-referential terms makes a difference in truth-value for counterfactuals with possible antecedents too. Then we should expect some evidence of that, especially in the form of convincing examples. Brogaard and Salerno provide none. Indeed, we may expect them to prefer the first horn of the dilemma, given their already quoted remark that 'All the typical rules governing counterfactuals are valid, when the antecedent is possible', made in the context of the reasoning, which they endorse, from (53) and (54) to (55) (2013, p. 657):

(53) If the rocket had continued on that course, it would have hit Hesperus.

(54) Hesperus = Phosphorus.

(55) If the rocket had continued on that course, it would have hit Phosphorus.

Brogaard and Salerno's general approach to the semantics of counterfactuals is to keep the structure of Lewis's overall account but to add lots of impossible worlds, considered simply as sets of sentences, not required to be deductively closed, in order to handle counterpossibles. Of course, some sets of sentences do contain $m = m$ without containing $m = p$, even though the latter is in fact true; the singleton $\{m = m\}$ is an example. They flesh out their semantic theory by adding an account of relative closeness for impossible worlds (2013, p. 655):

> For any two impossible worlds w_1 and w_2, w_1 is closer to the base world than w_2 iff

(a) w_1 does not contain a greater number of sentences formally inconsistent with the relevant background facts (held fixed in the context) than w_2 does.

And if w_1 and w_2 contain the same number of sentences formally inconsistent with the relevant background facts (held fixed in the context):

(b) w_1 preserves a greater number of a priori* implications between sentences than w_1 does.

They explain a priori* implication thus: 'For a speaker *s* in a context *c*, P *a priori* implies* Q iff for *s* in *c*, Q is a relevant a priori consequence of P' (ibid.).

We need not examine Brogaard and Salerno's semantics in detail. It is hard to believe that one will get a useful measure of closeness by counting numbers of sentences, especially when w_1 and w_2 both contain a countable infinity of sentences. We are also not told how to evaluate open formulas at impossible worlds, as required in evaluating quantified counterpossibles. For instance, is the open sentence '*x* is bright' true or false at a world *w* containing the sentence 'Hesperus is bright' but not the sentence 'Phosphorus is bright', under the assignment of Hesperus, which is to say Phosphorus, as the value of the variable '*x*'?

For present purposes, a more interesting feature of Brogaard and Salerno's account is the appearance of the distinctively epistemic notion of a priori* implication in the account of relevant closeness for impossible worlds. They leave in place Lewis's account of the relative closeness of possible worlds, which depends on objective, non-epistemic features of the worlds (given a fixed context). Thus their overall relative closeness relation is patched together from epistemic and non-epistemic pieces. It is hard to avoid the impression that the account is being gerrymandered just to accommodate the marginal case of counterpossibles (see also the definition of validity in Berto, French, Priest, and Ripley 2016). Such a hybrid approach resembles an account of conditional probabilities on which they are purely objective when the conditioning event has a positive unconditional chance, but go epistemic or subjective when the conditioning event has zero unconditional chance: not an attractive option.

In addition to the general implausibility of such hybrid theories, there is a more specific problem. As opponents of orthodoxy like to emphasize, many counterpossibles do not simply crash when uttered. Speakers and hearers handle them in ways very similar to the ways in which they handle non-counterpossible counterfactuals. But if the semantics of counterfactuals is to be done in terms of the relative closeness of worlds, which turns out to work in radically different ways depending on whether the worlds are possible or impossible, then *shouldn't* speakers and hearers handle counterpossibles quite differently from how they handle other counterfactuals? Lewis's relative closeness for possible worlds is not remotely like Brogaard and Salerno's relative closeness for impossible worlds, so how come the same style of cognitive processing works for both? A more uniform account would be more plausible.

Closely related to the opacity of counterfactuals on the impossible worlds approach is the lack of a non-trivial relation of synonymy appropriately related to the compositional semantics. For suppose that the semantics is compositional with respect to some kind of meaning, m-meaning, in the sense that the m-meaning of a complex expression supervenes on the m-meanings of its atomic constituents and the way they are put together. The corresponding relation of synonymy is just sameness of m-meaning. If such m-synonymy is non-trivial, then it should sometimes hold between distinct atomic expressions. To take a standard example, suppose that 'furze' and 'gorse' are m-synonymous. Since the semantics is compositional for m-meaning, substituting one of those words for the other in a complex expression preserves its m-meaning. Thus (56) and (57) are m-synonymous (all occurrences of 'this' and 'it' being coreferential):

(56) If this were furze but not gorse, it would be furze but not gorse.

(57) If this were furze but not gorse, it would be furze but not furze.

But the semantics with impossible worlds will typically evaluate (56) and (57) differently: (56) will trivially be evaluated as true, whereas (57) will usually be evaluated as false because supposing that furze and gorse are distinct natural kinds requires nothing as extreme as supposing a contradiction, even though the former supposition is impossible too. Since (56) and (57) differ in truth-value, they differ in m-meaning, for the truth-values of sentences (in a context) should supervene on their compositional meanings, here m-meanings. This contradicts the previous result that (56) and (57) are m-synonymous. Since parallel considerations apply to any other supposed case of non-trivial m-synonymy, we must deny the initial assumption that m-synonymy is non-trivial. Another way of thinking about the issue is that the use of all sets of sentences as impossible worlds in the semantics for the counterfactual conditional in effect turns it into a quotational context, which is equally unattractive. The upshot is that the impossible worlds approach provides a compositional semantics for the counterfactual conditional only in a rather feeble sense.

5. *Counterfactual reasoning by reductio ad absurdum*

The prize 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. Here are examples adapted from Lewis (1973, p. 25):

(58) If p were the largest prime, $p! + 1$ would be prime.

(59) If p were the largest prime, $p! + 1$ would be composite.

They summarize a slight variation on Euclid's proof that there is no largest prime: (58) holds because if p were the largest prime, $p!$ would be divisible by all primes (since it is divisible by all natural numbers from 1 to p), so $p! + 1$ would be divisible by none; (59) 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 (DC) to conjoin the consequents of (58) and (59):

(60) If *p* were the largest prime, *p*! + 1 would be both prime and composite.

Since the consequent of (60) 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, for 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, especially when the hearer had already been told that there is no largest prime, and wanted to know why. At the very least, on a good semantic theory, the counterpossibles (58)-(60) should come out *true*, for they are soundly based on valid mathematical reasoning.

It is far from clear that (58)-(60) come out true on Brogaard and Salerno's theory. The point is not merely that, on their account, some impossible worlds include the sentence '*p* is the largest prime' while excluding the sentence '*p*! + 1 is prime' or '*p*! + 1 is composite' (or both), for that does not show that they are the closest worlds including '*p* is the largest prime' to the actual world. It is instructive to note Brogaard and Salerno's reaction to a much more elementary argument by *reductio ad absurdum* (Williamson 2007, p. 172):

> [I]f 5 + 7 were 13 then 5 + 6 would be 12 (and so by another eleven steps) 0 would be 1, so if the number of right answers I gave were 0, the number of right answers I gave would be 1.

They analyse that compressed argument as using *reductio ad absurdum* but argue that, as it stands, it is not cogent, because it does nothing to show that the nearest impossible world in which the antecedent is true has the required properties (2013 pp. 649-50). Exactly the same objection applies to the classic proof that there is no largest prime, cast in counterfactual terms. After all, what mathematician has ever supplemented a standard proof by *reductio ad absurdum* with considerations about the relative closeness to the actual world of various impossible worlds, considered as sets of sentences that need not be deductively closed? The cost of Brogaard and Salerno's rejection of my toy proof is a commitment to rejecting ordinary mathematical proofs by *reductio ad absurdum* cast in counterfactual terms. Nor is there any reason to expect special contextual factors somehow to rescue the hard mathematical proofs but not my easy one, or to postulate a different meaning for the counterfactual conditional in mathematics from its meaning in philosophy.

An initial reaction might be that Brogaard and Salerno must have played their cards badly, and that the trouble they have got themselves into reveals nothing about the strength of their original hand. But further reflection suggests that the problem for opponents of orthodoxy goes much deeper than that, and does not depend on the details of their view. For 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 α □→ ω. 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 in Nolan 1997.

One fallback is to concede orthodoxy for counterlogicals, countermathematicals, and the like, but still reject it for some countermetaphysicals. But not only is that a long retreat, it risks undermining the motivation for the residual resistance to orthodoxy in the first place, since the motivating features of the examples are already present in the case of counterlogicals. If those features are somehow illusory, who is to say that similar illusions are not also at work with countermetaphysicals? Thus the fallback is unstable.

Another response is to say that formulating mathematical arguments in counterfactual terms was always loose speech. One must then explain why those formulations *seemed* so compelling to the most rigorous and precise reasoners in our community — mathematicians and logicians — in effect, why they mistook loose speech for strict speech. That would be some sort of error theory. Such a response also risks instability, for if anti-orthodox theorists need an error theory, what is to stop orthodox theorists from applying their own error theory to the other side's considerations, and escaping the messy complications of anti-orthodoxy?

Yet another defensive move is to invoke context-dependence, arguing that (60) is true in the context of standard mathematical proof (contrary to Brogaard and Salerno's reaction) but false in some other contexts. This move might concede a reading of counterfactuals on which they can be used to define metaphysical modalities, while also providing another reading of them on which they can be used for fictionalist purposes in metaphysics. Of course, the appeal to context-dependence is a generic way of rejecting the validity of just about any deductive argument. To carry weight, it needs more specific support. It is not controversial that counterfactuals exhibit some degree of context-dependence in which factors are held fixed; if Julius Caesar had been in command during the Korean War, would he have used catapults or nuclear weapons? But that makes it all the more striking how hard it is to hear the mathematical argument for (60) as unsound. If context determined whether standard logic and mathematics were to be held fixed, (60) might be expected to trigger a shift to a context in which they are not, because it brings out so clearly that the antecedent is untenable without such a shift. Yet the argument for (60) remains compelling. Although one might be able to browbeat ordinary hearers into questioning the mathematical arguments, that is very different from the smooth, unreflective transitions characteristic of ordinary context-dependence. Without adequate motivation, the appeal to context-dependence is just an all-purpose objection to any valid argument. That context can shift the relative ranking of given worlds does not mean that it can shift which worlds are being ranked at all.

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?

6. *An error theory of apparently false counterpossibles*

Processing a non-obvious counterpossible typically *feels* very like processing a non-counterpossible counterfactual. Consider (13), 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, in this case, 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).

Schematically, what we seem to do when we assess the counterfactual $\alpha \mathrel{\square\!\!\rightarrow} \beta$ is this. First we counterfactually suppose α. Then, if within the scope of the counterfactual supposition we accept β, outside the scope of that supposition we accept $\alpha \mathrel{\square\!\!\rightarrow} \beta$. Similarly, if within the scope of the counterfactual supposition we reject β, outside the scope of that supposition we reject $\alpha \mathrel{\square\!\!\rightarrow} \beta$. More generally, whatever assessment we make of β within the scope of the counterfactual supposition of α we make of $\alpha \mathrel{\square\!\!\rightarrow} \beta$ outside the scope of that supposition; call that the *suppositional procedure*. 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.[20]

From an orthodox perspective, over-hasty applications of the suppositional procedure to counterpossibles produce false judgments. At first sight, that looks like a golden opportunity for unorthodox theorists to motivate their semantics. If an unorthodox semantic theory vindicates our way of assessing counterfactuals, while orthodox theories do not, that is a strong point in favour of unorthodoxy.

On further reflection, however, any prospect of fully vindicating the suppositional procedure goes dim. The first reason is just the use of counterfactuals in articulating mathematical proofs by reductio ad absurdum, discussed above. In a normal mathematical context, we counterfactually suppose that *p* is the largest prime, and in the scope of that supposition deduce the absurd conclusion that *p*! + 1 is both prime and composite. Still within the scope of the counterfactual supposition, we reject that as absurd. But we do not reject the corresponding counterfactual conditional ((60) above). Rather, we accept it in explaining our reasoning. Thus, on our best behaviour, we *contravene* the suppositional procedure.

A similar result follows given any sentence ω so absurd that we must reject it under any supposition whatever. For then we reject ω under the counterfactual supposition of ω, so the suppositional procedure tells us to reject the corresponding counterfactual conditional ω □→ ω. But on reflection we do not; we accept ω □→ ω because anything of the form α □→ α is a logical truth, even on the usual versions of impossible world semantics. On a variant of this theme, we might take ω to be 'Everything should actually be rejected'. In the scope of that counterfactual supposition, perhaps one should reject even ω itself, but outside the scope of that supposition one should not reject ω □→ ω.

The suppositional procedure also gets into trouble with probabilistic assessments. On its most natural application to such cases, the procedure involves equating one's credence in the conditional α □→ β with one's conditional credence in β on α:

(Cond) $\text{Prob}(\alpha \mathrel{\square\!\!\rightarrow} \beta) = \text{Prob}(\beta \mid \alpha)$

There is evidence that, in general, our assessment of the probability of a conditional is highly correlated with our assessment of the conditional probability of its consequent on its antecedent (Evans, Over, and Handley 2005). But (Cond) breaks down for counterlogical α. Under the standard equation of the conditional probability Prob(β | α) with the ratio Prob(α ∧ β)/Prob(α) of unconditional probabilities, the conditional probability is undefined when Prob(α) is zero, as the axioms of probability require it to be when α is a contradiction. If instead 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 in the sense that β holds at every point in the space at which α holds, so for vacuous α Prob(β | α) = Prob(¬β | α) = 1 for any β, violating 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.

In any case, (Cond) faces problems even when α is possible. The equation of probabilities of conditionals is more attractive for indicative conditionals than for subjunctive ones like α □→ β; combining both equations commits one to equating the probability of a subjunctive conditional with the probability of the indicative conditional with the same antecedent and consequent, thereby erasing a significant distinction. Moreover, a long line of impossibility results, initiated by David Lewis, shows that under

very general conditions there cannot be conditional propositions whose probabilities always equal the corresponding conditional probabilities (Lewis 1976, Hájek 2011).

Would-be vindicators of the suppositional procedure might respond to these problems by denying that the probability of β under the counterfactual supposition of α is the conditional probability of β on α. They might instead identify it with the probability of β under some other probability distribution $Prob_\alpha$ determined by α:

(Cond*) $\quad Prob(\alpha \mathrel{\square\!\!\rightarrow} \beta) = Prob_\alpha(\beta)$

Since the probability of α itself under the counterfactual supposition of α should be 1, $Prob_\alpha(\alpha) = 1$, and (Cond*) yields the same result since $Prob(\alpha \mathrel{\square\!\!\rightarrow} \alpha) = 1$. Thus not even (Cond*) solves the problem of counterlogicals, since $Prob_\alpha$ was supposed to be a probability distribution, which requires it to assign probability 0, not 1, to contradictions. Rejecting standard axioms of the probability calculus for the sake of one's favoured treatment of counterpossibles would be in the worst methodological taste, since the explanatory power of probability theory depends on its mathematical strength, which relies on those axioms.

For possible α, (Cond*) does not face impossibility results anything like as bad as those for (Cond). Lewis (1976) showed how to make sense of conditionals defined by equations like (Cond*) by defining $Prob_\alpha$ in terms of an operation of *imaging* on α. However, (Cond*) too has contentious results. For example, since $Prob_\alpha(\neg\beta) = 1 - Prob_\alpha(\beta)$, $Prob(\alpha \mathrel{\square\!\!\rightarrow} \neg\beta) = Prob_\alpha(\neg\beta) = 1 - Prob_\alpha(\beta) = 1 - Prob(\alpha \mathrel{\square\!\!\rightarrow} \beta) = Prob(\neg(\alpha \mathrel{\square\!\!\rightarrow} \beta)$.[21] But for possible α, by (Cond*):

$$Prob((\alpha \mathrel{\square\!\!\rightarrow} \beta) \wedge (\alpha \mathrel{\square\!\!\rightarrow} \neg\beta)) = Prob(\alpha \mathrel{\square\!\!\rightarrow} (\beta \wedge \neg\beta)) = Prob_\alpha(\beta \wedge \neg\beta) = 0$$

By the standard probability axioms, the probability of a two-way disjunction is the sum of the probabilities of the disjuncts minus the probability of their conjunction, so we have a probabilistic version of the contentious principle of conditional excluded middle for possible antecedents:

(PCEM) $\quad Prob((\alpha \mathrel{\square\!\!\rightarrow} \beta) \vee (\alpha \mathrel{\square\!\!\rightarrow} \neg\beta)) = 1$

Conditional excluded middle, the disjunction in (PCEM), is valid in Stalnaker's logic of conditionals, but invalid in Lewis's logic of counterfactuals.

Recently, natural language data have been used in defence of conditional excluded middle (Williams 2010). However, its epistemological consequences are highly problematic. Here is an example. Consider an epistemically reasonable probability distribution over worlds. Let *w* be a possible world where a fair coin is not tossed, *h* a possible world where it is tossed and comes up heads and *t* a possible world where it is tossed and comes up tails, *h* and *t* being as far as possible symmetrically related to *w*. Let *W* be true only in *w*, *H* true only in *h*, and *T* true only in *t*. Since *W* holds in only one world, for any α either *W* entails α in the probability space, in which case $W \mathrel{\square\!\!\rightarrow} \alpha$ has probability 1, or *W* is incompatible with α in that space, in which case $W \mathrel{\square\!\!\rightarrow} \alpha$ has probability 0. In particular, therefore, $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} H)$ has probability 1 or 0. Suppose that it has probability 1. Then, in effect, the distribution treats it as certain that *h* is selected over *t* as 'closer' to *w*: roughly, it

is certain that in *w* if the coin had been tossed it would have come up heads. But that is epistemically quite unreasonable. The only alternative is for $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} H)$ to have probability 0. But then $W \mathrel{\square\!\!\rightarrow} \neg((H \vee T) \mathrel{\square\!\!\rightarrow} H)$ has probability 1, so by conditional excluded middle $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} \neg H)$ has probability 1, and so too has $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} T)$. But that is equally epistemically unreasonable. Thus conditional excluded middle does not permit the required epistemic uncertainty between the worlds *h* and *t* with respect to *w*. Nor would it help to say that it is indeterminate (whatever that means) whether *h* or *t* is selected, and therefore indeterminate whether $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} H)$ has epistemic probability 1 or 0, since that is not an available epistemic state for us. Notably, the problem does *not* generalize to Lewis's semantics, since he can straightforwardly treat *h* and *t* as equally close to *w* and so evaluate both $(H \vee T) \mathrel{\square\!\!\rightarrow} H$ and $(H \vee T) \mathrel{\square\!\!\rightarrow} T$ as false at *w*, and so assign probability 0 to both $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} H)$ and $W \mathrel{\square\!\!\rightarrow} ((H \vee T) \mathrel{\square\!\!\rightarrow} T)$.[22]

What all these considerations suggest is that there is no reasonable way of fully vindicating the suppositional procedure. It breaks down for inconsistent antecedents, and may well do so for mundanely possible ones too. It does not follow that we do not use the suppositional procedure. It is plausibly our normal, unreflective way of evaluating counterfactual conditionals. But it is not 100% reliable. Like many of our methods of judgment, it is a useful but fallible heuristic. In particular, the linguistic data used to support the principle of conditional excluded middle may be the misleading outputs of such a heuristic.

For present purposes, the full power of the suppositional procedure is not needed. We need only consider a consequence of it that is independent of conditional excluded middle. Suppose that β is seen as inconsistent with γ. Then, normally, β is still seen as inconsistent with γ under the counterfactual supposition of α. Epistemically, the suppositional procedure treats combining β and γ under the counterfactual supposition of α as tantamount to combining $\alpha \mathrel{\square\!\!\rightarrow} \beta$ and $\alpha \mathrel{\square\!\!\rightarrow} \gamma$. Thus an inconsistency in the former is seen as tantamount to an inconsistency in the latter. That yields the following heuristic as a by-product of the suppositional procedure:

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

Thus, recognizing that β is inconsistent with γ, if we accept $\alpha \mathrel{\square\!\!\rightarrow} \beta$, we reject $\alpha \mathrel{\square\!\!\rightarrow} \gamma$. We can apply this rule when drawing out the implications of any counterfactual supposition α. In practice, when using the suppositional procedure, we need not explicitly accept $\alpha \mathrel{\square\!\!\rightarrow} \beta$ in order to reject $\alpha \mathrel{\square\!\!\rightarrow} \gamma$; it is enough to accept β within the counterfactual supposition of α, because the suppositional procedure treats those as equivalent.

For example, all this can be applied to the case of (13) and (14). '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 economically, having rejected 'Sick children in the mountains of South America at the time cared' within the counterfactual supposition of the antecedent, we can directly reject (13) by the suppositional procedure. The upshot is

the same. As already noted, the impossibility of the antecedent 'Hobbes (secretly) squared the circle' plays no role in our reasoning; logical relations under that supposition are treated no differently from usual.

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. The suppositional procedure endorses (HCC*) as a heuristic: if you accept $\alpha \mathrel{\Box\!\!\rightarrow} \beta$, you accept β under the counterfactual supposition of α, so by normal reasoning you reject $\neg\beta$ under the counterfactual supposition of α, so you reject $\alpha \mathrel{\Box\!\!\rightarrow} \beta$; likewise with β and $\neg\beta$ interchanged.

(HCC) and (HCC*) are equivalent under normal conditions. First, start with (HCC). Clearly, β is inconsistent with $\neg\beta$. 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*), and let it be given that β is inconsistent with γ. Thus γ entails $\neg\beta$. So, normally, from $\alpha \mathrel{\Box\!\!\rightarrow} \gamma$ you can derive $\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta$, by an informal analogue of Deduction within Conditionals.[23] But (HCC*) tells you not to accept both $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ and $\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta$. So, normally, you should not accept both $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ and $\alpha \mathrel{\Box\!\!\rightarrow} \gamma$. 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 (Evans, Handley, and Over 2005, pp. 1049-50). Moreover, if one evaluates the probabilities of counterfactuals according to (Cond*) (of which (Cond) is a special case), then $\text{Prob}(\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta) = 1 - \text{Prob}(\alpha \mathrel{\Box\!\!\rightarrow} \beta)$, so if one of $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ and $\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta$ is probable, the other is improbable, so not both can be accepted.

For the orthodox, (HCC) and (HCC*) are only heuristics because they 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 those cases, (HCC*) will never lead you astray. For if you accept one of $\alpha \mathrel{\Box\!\!\rightarrow} \beta$ and $\alpha \mathrel{\Box\!\!\rightarrow} \neg\beta$, and it is true (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.

Even on an unorthodox view of counterfactuals, the prospects for fully vindicating (HCC) and (HCC*) are bleak. The previous problems for the suppositional procedure with inconsistent antecedents apply just as much to them. For example, consider counterpossibles with explicit contradictions as antecedents:

(61) $(\alpha \wedge \neg\alpha) \mathrel{\Box\!\!\rightarrow} \alpha$

(62) $(\alpha \wedge \neg\alpha) \boxrightarrow \neg\alpha$

Both (61) and (62) look highly plausible; surely conjunctions counterfactually imply their conjuncts. But if both (61) and (62) are true, then they are consistent, even though they have the same antecedent and inconsistent consequents. Now unorthodox theorists may reject some instances of (61) and (62), 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. Dialetheists both assert that the Liar is true *and* assert that the Liar is not true. Presumably, therefore, both (61) and (62) 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.

The situation is this. The naive evaluation of some counterpossibles, such as (13), involves a move from the rejection of the consequent under the counterfactual supposition of the antecedent to the rejection of the whole conditional. The principles that rationalize such a move turn out to hold only for the most part; they are fallible heuristics. There are independent reasons to expect them to fail for at least some counterpossibles. Thus we should not rely on them uncritically.

How should we expect the heuristic evaluation of $\alpha \boxrightarrow \beta$ and $\alpha \boxrightarrow \neg\beta$ to go when α entails both β and ¬β? In effect, since both β and ¬β would eventually emerge as we developed the counterfactual supposition α for long enough, the heuristics make it a race between the contradictories as to which emerges first. If β emerges first, we accept $\alpha \boxrightarrow \beta$ and reject $\alpha \boxrightarrow \neg\beta$ before ¬β has time to emerge. If ¬β emerges first, we accept $\alpha \boxrightarrow \neg\beta$ and reject $\alpha \boxrightarrow \beta$ before β has time to emerge. Proponents of impossible worlds misinterpret this computational, inferential difference in terms of the relative closeness of impossible $\alpha \wedge \beta$ and impossible $\alpha \wedge \neg\beta$ 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.[24] 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.

We can test the three heuristics — the suppositional procedure, (HCC), and (HCC*) — by trying them out on three examples offered by Cian Dorr as 'some manifestly false counterfactuals whose antecedents seem to be metaphysically impossible' (2008, p. 37):[25]

(63) If I were a dolphin, I would have arms and legs.

(64) If it were necessary that there be donkeys, it would be impossible that there be cows.

(65) If there were unicorns, none of them would have horns.

Let us grant Dorr that in each case the antecedent is indeed metaphysically impossible.

First, consider (63). Let α = 'I am a dolphin', β = 'I have arms and legs', and γ = 'I do not have arms and legs'. In developing the antecedent supposition 'I am a dolphin', one holds fixed one's general knowledge of dolphins, including one's knowledge that they do not have arms and legs, and so derives 'I do not have arms and legs'. One thus accepts γ. Since β is manifestly inconsistent with γ, one rejects β. Any of the three heuristics then leads one to accept α □→ γ and reject α □→ β, which is (63). Thus the employment of any of the heuristics explains our rejection of (63).

Next, consider (64). Let α = 'It is necessary that there are donkeys', β = 'It is impossible that there are cows', and γ = 'It is possible that there are cows' (where the modal operators are read non-epistemically, as Dorr intends). In developing the antecedent supposition 'It is necessary that there are donkeys' with an eye to (64), one considers the modal status of 'There are cows'. In actuality, there are cows, so cows are possible. The supposed necessity of donkeys presents no good reason not to hold those facts fixed. Indeed, if one treats the modal status of cows and donkeys on a par, one would have to make cows necessary too, and therefore still possible. Hence one still favours 'It is possible that there are cows'. One thus accepts γ. Since β is manifestly inconsistent with γ, one rejects β. Any of the three heuristics then leads one to accept α □→ γ and reject α □→ β, which is (64). Thus the employment of any of the heuristics explains our rejection of (64).

Finally, consider (65). Let α = 'There are unicorns', β = 'No unicorn has horns', and γ = 'There are unicorns with horns'. In developing the antecedent supposition 'There are unicorns', one holds fixed with one's background conception of unicorns, including 'Unicorns have horns', and so derives 'There are unicorns with horns'. One thus accepts γ. Since β is manifestly inconsistent with γ, one rejects β. Any of the three heuristics then leads one to accept α □→ γ and reject α □→ β, which is (65). Thus the employment of any of the heuristics explains our rejection of (65).

Our impression that Dorr's examples are all false can be explained by our reliance on a fallible heuristic. On the suggested explanations, in no case does the impossibility of the antecedent figure in our assessment. We simply do not consider the modal status of the antecedent. That seems to fit the phenomenology of unreflectively assessing (63)-(65).

Any of the heuristics can also be applied to mathematical examples:

(66) If 289 were divisible by 3, 290 would also be divisible by 3.

A natural impression is that (66) is false. The explanation is along the usual lines. Let α = '289 is divisible by 3', β = '290 is divisible by 3', and γ = '290 is not divisible by 3'. In developing the antecedent supposition '289 is divisible by 3', one holds fixed one's background knowledge that the successor of a multiple of 3 is never a multiple of 3, and so derives '290 is not divisible by 3'. One thus accepts γ. Since β is manifestly inconsistent with γ, one rejects β. Any of the three heuristics then leads one to accept α □→ γ and reject α □→ β, which is (66). Thus the employment of any of the heuristics explains our rejection of (66).

But of course there are also equally cogent mathematical deductions of '290 is divisible by 3' from '289 is divisible by 3'; they are just a bit longer and less psychologically salient. Rejecting (66) would be unfaithful to the natural use of counterpossibles in formulating mathematical proofs.

Any of the three heuristics leads us to the false conclusion that (66) is false, and in very similar ways to the conclusions that (63)-(65) are false. To say the least, we should be wary of those conclusions too. Our impression that (63)-(65) are 'manifestly false' may well the product of our reliance on a highly fallible heuristic.[26]

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), (HCC*), and the suppositional procedure. For example, we accept both the counterpossibles (58) and (59) 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 obtained?' (Williamson 2007, pp. 162). Nevertheless, despite our ability to inhibit their operation, heuristics 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 (58) and (59) are mutually inconsistent after all.[27]

A useful analogy, already suggested in section 2, is with our naïve reactions to true universal quantifications with empty subject terms:

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

(68) Every unicorn is hornless.

A natural inclination is to judge (67) and (68) false. Even when one is told that there are no dolphins in Oxford and no unicorns, one still feels some resistance to accepting (67) and (68) respectively. That resistance is explicable by the hypothesis that we accept (69) and (70) on the basis of background information about dolphins and unicorns respectively, and are then inclined to reject (67) as inconsistent with (69) and (68) as inconsistent with (70):

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

(70) Every unicorn has a horn.

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.[28]

We can come to recognize the limitations of (HUQ) and (HUQ*) through natural reasoning. For instance, suppose that our rejection of (67) leads us to accept its negation:

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

From (71) we can validly reason to (72):

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

From (72) we can in turn validly reason to (73):

(73) There is a dolphin in Oxford.

But we know (73) to be false. That may lead us to realize that (67) 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.

A more general cognitive pattern underlying may explain these heuristics. For example, it is plausible that we use analogues of them for indicative as well as subjunctive conditionals, and for generic as well as universal quantifiers. We ignore the issue of the empty case. We continue using heuristics that do so 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 (67) 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 an accurate view of the truth-conditions of sentences of our native language formed using the most basic logical constants.[29] 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 (67) and (68) are false while (69) and (70) are 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, where some theories make sentences such as 'Muslims are terrorists' come out true.[30]

In the case of the universal quantifier, proper 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), (HCC*), or the suppositional procedure. How far should that view make us sceptical more generally about reliance on pre-theoretic assessments of counterfactuals in philosophy, semantics and elsewhere? Several points are worth noting.

First, 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 for counterfactuals. Blanket scepticism is not a sensible response.

Second, the problems posed for the heuristics by counterpossibles concern the rejection of $\Box\!\!\rightarrow$ conditionals, not their acceptance. 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 $\Box\!\!\rightarrow$ conditionals (Williamson 2007, pp. 179-207).

Third, nothing said here impugns the reliability of counterfactual judgments made on the basis of mathematical reasoning.

Fourth, the heuristics at issue are of such a general and pre-reflective nature that one might conjecture them to be universal amongst humans, more or less hard-wired into us independently of intelligence and education, with relatively little individual variation. But of course that does not mean that there will be no such variation in the inputs (acceptance of one counterfactual) or correspondingly in the outputs (rejection of another). Furthermore, whether the individual takes the output of the heuristic at face value or second-guesses it on the basis of theoretical reflection may well be highly sensitive to individual variation and educational background. It is an appropriate locus for the application of philosophical expertise.

Fourth, 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. The operation of one heuristic may also be pre-empted or inhibited by the operation of another.

Finally, 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.[31]

Notes

1 For more on the conception of metaphysical possibility as the most inclusive sort of objective possibility see Williamson 2016. One consequence discussed there is that, where ◊ expresses metaphysical possibility, the S4 principle ◊◊α ⊃ ◊α holds (contrary to Salmon 1989), for if ◊ expresses a sort of objective possibility, so does ◊◊. Correspondingly, □□α is no stronger than □α for this modality.

2 For Stalnaker (1968) there is a small wrinkle. For convenience, he postulates an impossible 'absurd world' λ where everything is true, and equates the truth-value of a counterpossible with the truth-value of its consequent at λ (the only world where its antecedent is true). However, the effect is the same, because the consequent is automatically true at λ. His semantics is still compositional with respect to intensions defined over possible worlds, because λ does not discriminate between sentences with the same intension.

3 Proof: Substitute β ∧ ¬β for α* and β for β* in (7). By elementary propositional modal reasoning (which can be carried out in the weakest normal system K), the result reduces to □¬α ⊃ □((α □→ β) ≡ ((β ∧ ¬β) □→ β)). Substitute β for α and ¬β for β in (8). The result is □((β ∧ ¬β) □→ β). Further elementary modal reasoning then yields (9).

4 If metaphysical possibility obeys the S4 axiom, as suggested in n. 1, then in the metaphysical sense impossible worlds are not even possibly possible, or possibly possibly possible, or …; their impossibility is of a more radical sort.

5 Field 1989 treats platonist mathematics as contingently false, but with respect to a logical rather than metaphysical modality. Such a logical modality arguably falls outside the range of objective modalities indicated above. For example, 'Socrates = Plato' is presumably possible in the alleged logical sense but is not metaphysically possible; the necessity of distinctness follows from the necessity of identity in S5, the best candidate for the propositional modal logic of metaphysical modality. See Williamson 2016 for more discussion.

6 For instance, it is axiom (a4) of Stalnaker (1968).

7 See Williamson 2007, p. 156. In the setting of a quasi-Lewisian approach to the semantics of counterfactuals, (18) corresponds to the Strangeness of Impossibility Condition Daniel Nolan is 'tempted to impose': 'any possible world

is more similar (nearer) to the actual world than any impossible world' (1997, p. 550; see pp. 566-7 for more discussion).

8 Proof: Substitute ¬β for γ in the right-to-left direction of (17) and β ∧ ¬β for β in (18) and use ¬◊(β ∧ ¬β).

9 But some applications of unorthodoxy about counterpossibles to metaphysics fail even on unorthodox terms. A case in point is Brogaard and Salerno's attempt to revive a modal analysis of essence (using counterfactuals) against well-known objections by Kit Fine (1994). As they explain Fine's argument, 'While Kripke's wooden table, Tabby, is necessarily a member of the set {Tabby}, it is not essential to Tabby that it be a member of that set' (Brogaard and Salerno 2013, p. 646). On their account of the ordinary use of 'essential', 'There being *F*s is essential to *x* iff if there were no *F*s then *x* wouldn't exist'; this is intended to allow that there being doctors may be essential to someone whose life they saved since if there were no doctors she wouldn't exist. Their account of the 'philosophical use' is the one relevant to Fine's challenge; on it 'There being *F*s is essential to *x* (or *x* is essentially *F*) iff (i) if there were no *F*s then *x* wouldn't exist, and (ii) it is metaphysically necessary that if *x* exists then *x* is *F*' (op. cit., p. 647). But this gets the Tabby example wrong on their own terms. For substitute 'Tabby' for '*x*' and 'member of a singleton of Tabby' for '*F*'. Then (ii) holds because it is metaphysically necessary that if Tabby exists then Tabby is a member of a singleton of Tabby. As for (i), their discussion assumes a contingentist view of existence, otherwise their account of the ordinary use would fail by their own lights in the doctors case. Hence we may assume that Tabby could have failed to exist. In those circumstances, there would have been no members of a singleton of Tabby, because there would have been no singleton of Tabby. Thus the relevant instance of (i) is a non-counterpossible, because its antecedent is possible; thus (i) behaves normally on their view. Moreover, it is metaphysically necessary that if Tabby exists then there are members of a singleton of Tabby, since Tabby itself is one. Hence (i) is true too. Thus, by Brogaard and Salerno's analysis, in the philosophical sense there being members of a singleton of Tabby is essential to Tabby (or Tabby is essentially a member of a singleton of Tabby), exactly the result they needed to avoid. In response to a different problem, they suggest requiring that 'F is essential to *a* only if: there might still be Fs even if *a* hadn't existed' (op. cit., p. 647 n.9), which handles the counterexample above at the unwise cost of making being Tabby inessential to Tabby.

10 In this paper I make the simplifying assumption that all counterfactuals are true or false in a given context.

11 Implementations of such a fine-grained conception of meaning have been familiar in the tradition of formal intensional semantics from the beginning; see Carnap 1947, pp. 56-9, and Lewis 1970, section V.

12 For example, Brogaard and Salerno's criticism of my 2007 in the passage beginning 'Interestingly enough' (2013, pp. 650-1) depends on neglect of the epistemic informativeness of counterpossibles.

13 For simplicity, I assume a fixed context and a set-sized domain.

14 For simplicity, I ignore the possibility that 'unicorn' suffers some form of reference failure worse than mere emptiness of extension.

15 Lewis implicitly requires the premise set I to be finite; without the Limit Assumption the infinitary version of the principle is invalid on his semantics (Lewis 1973, pp. 19-21). Informally, however, the infinitary version is almost as plausible as the finitary one; it is valid on Stalnaker's semantics. Lewis also requires the premise set to be nonempty, but the special case with an empty premise set is equally plausible, and valid on his semantics. It boils down to this: if $\vdash \gamma$ then $\vdash \alpha \mathrel{\Box\!\!\rightarrow} \gamma$ (the conjunction of an empty set of conjuncts is a tautology, since a conjunction is false only if at least one conjunct is false).

16 In the terminology of Davies and Humberstone 1980, this corresponds to real world validity rather than general validity.

17 See Williamson 2006 and 2007, pp. 295-6, for more discussion.

18 Brogaard and Salerno seem to think that the opacity of counterfactuals can be derived from their hyperintensionality: 'Hyperintensional operators do not permit substitutions of co-referring terms *salva veritate*' (2013, p. 650). That is false. For instance, let $\alpha \# \beta$ be true when α and β express the same Russellian structured proposition, and false otherwise. Then # is hyperintensional, because sentences with the same intension may express different Russellian propositions, but # is transparent, because (absent other opaque operators) substitutions in its scope of co-referring terms preserve the Russellian proposition expressed.

19 They write: 'Of course, we need to give a principled account of when counterfactuals create opaque contexts. They create opaque context [sic] when the antecedent or consequent which result [sic] from substituting one term for another does not follow a priori from the original. Since we are likely to use "Clark Kent" and "Superman" in such a way as to pick out the same individual, "Clark Kent has the same parents as I do" is an a priori implication of "Superman has the same parents as I do". So, substitution is legitimate' (2013, p. 650 n. 11). As stated, this seems to require 'Clark Kent is Superman' to be a priori. Perhaps

they mean a priori implications given the relevant identity as an auxiliary premise. In any case, none of this obviously helps with counterpossibles such as (50)-(52), where no peculiar constructions need appear in antecedent or consequent.

20 Evans, Handley, and Over 2005 provides a useful overview of the evidence for a suppositional account of the evaluation of conditionals, though they may sometimes be too quick in passing between psychological claims and semantic ones. See Evans and Over 2004, pp. 113-131.

21 Compare Boethius' claim that to negate the consequent is to negate the conditional (Kneale and Kneale 1962, p. 191).

22 For those uncomfortable with doubly embedded subjunctive conditionals, one can replace the outer one by an ordinary strict conditional, and do the whole argument with $\Box(W \supset ((H \vee T) \Box\!\!\rightarrow H)$ in place of $W \Box\!\!\rightarrow ((H \vee T) \Box\!\!\rightarrow H)$.

23 Even for the unorthodox, (DC) is *normally* reliable. Even for the orthodox, there are the abnormal cases discussed at the end of section 3, where (DC) must be restricted because the operative standard of validity is real world validity and the object-language contains fancy modal operators such as the rigidifying 'actually' operator. Such abnormal cases are rare in practice.

24 In some cases, we do attend to the impossibility of the antecedent. For example, we prefer 'If Hobbes had squared the circle, he would have done something geometrically impossible' to 'If Hobbes had squared the circle, he would have done nothing geometrically impossible'. The heuristic account has no difficulty with such examples: cognitively, the connection from antecedent to consequent is still easier with the former counterfactual than with the latter.

25 For (65), Dorr cites Kripke's influential argument for the impossibility of unicorns (1980, p. 156).

26 Our practical use of 'If I were you' counterpossibles and the associated imaginative exercises may also be explicable in terms of such heuristics, without appeal to any special semantics of the *de se* or the like.

27 An alternative hypothesis is that our behaviour manifests sensitivity to conversational implicatures rather than use of heuristics. For example, in asserting (13) one might be thought to counterfactually imply that sick children in the mountains of South America kept track of Hobbes's activities. If such a line worked, it would provide an alternative defence of orthodoxy about counterpossibles. However, attempts to cancel the alleged conversational implicatures work poorly: for example, saying 'I don't mean to imply that those

children were aware of Hobbes' hardly dissolves the cognitive dissonance caused by asserting 'If Hobbes had squared the circle, sick children in the mountains of South America at the time would have cared'. Although conversational implicatures may sometimes be in the mix, they are not the crucial ingredient.

28 An example where they go wrong: 'Every σ which is a non-σ is a σ' and 'Every σ which is a non-σ is a non-σ'.

29 See Peters and Westerståhl 2006, pp. 124-7, for a concise defence of the modern view of existential import. For a contrasting view see Strawson 1952, pp. 163-79.

30 Sterken 2015 makes this point.

31 This paper is based on my 2015 Beth Lecture in Amsterdam. Other versions of this material were presented at the University of Oxford, University of Edinburgh, the University of Connecticut, Yale University, Ohio State University, a conference on conditionals at the University of Belgrade, a conference on the impossible at the University of Turin, and a Mind, World and Action course in Dubrovnik. I thank the audiences and also Jennifer Nagel, Carola Barbero, Andrea Iacona, and Alberto Voltolini for many helpful questions, objections, and comments. Thanks also to Francesco Berto, Rohan French, Graham Priest, and David Ripley for showing me their critique of an earlier version of this paper (Berto, French, Priest, and Ripley 2016), which has prompted some important amplifications of the discussion of heuristics. An earlier version of some of the present material appeared as Williamson 201X.

References

Berto, Francesco, French, Rohan, Priest, Graham, and Ripley, Dave 2016: 'Williamson on counterpossibles'. Typescript.

Brogaard, Berit, and Salerno, Joe 2013: 'Remarks on counterpossibles', *Synthese*, 190, pp. 639-660.

Carnap, Rudolf 1947: *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: Chicago University Press.

Davies, Martin, and Humberstone, Lloyd 1980: 'Two notions of necessity', *Philosophical Studies*, 38, pp. 1-30.

Dorr, Cian 2008: 'There are no abstract objects', in Ted Sider, John Hawthorne, and Dean Zimmerman (eds.), *Contemporary Debates in Metaphysics*. Oxford: Blackwell.

Edgington, Dorothy 2008: 'Counterfactuals', *Proceedings of the Aristotelian Society*, 108, pp. 1-21.

Evans, Jonathan, Handley, Simon, and Over David 2005: 'Suppositions, extensionality, and conditionals: a critique of the mental model theory of Johnson-Laird and Byrne (2002)', *Psychological Review*, 112, pp. 1040-1052.

Evans, Jonathan, and Over, David 2004. *If*. Oxford: Oxford University Press.

Field, Hartry 1980: *Science without Numbers: A Defence of Nominalism*. Oxford: Blackwell.

Field, Hartry 1989: *Realism, Mathematics, and Modality*. Oxford: Blackwell.

Fine, Kit 1994: 'Essence and modality', *Philosophical Perspectives*, 8, pp. 1-16.

von Fintel, Kai 1998: 'The presupposition of subjunctive conditionals'. In U. Sauerland and O. Percus (eds.), *The Interpretive Tract* (MIT Working Papers in Linguistics 25), pp. 29-44.

Grice, Paul 1989: *Studies in the Ways of Words*. Cambridge, Mass.: Harvard University Press.

Hájek, Alan 2011: 'Triviality pursuit', *Topoi*, 30, pp. 3-15.

Kment, Boris 2014: *Modality and Explanatory Reasoning*. Oxford: Oxford University Press.

Kratzer, Angelika 2012: *Modals and Conditionals*. Oxford: Oxford University Press.

Kneale, William, and Kneale, Martha 1962: *The Development of Logic*. Oxford: Clarendon Press.

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

Lewis, David 1970: 'General semantics', *Synthese*, 22, pp. 18-67.

Lewis, David 1973: *Counterfactuals*. Oxford: Blackwell. Page references to 2nd ed., 1986.

Lewis, David 1976: 'Probabilities of conditionals and conditional probabilities', *Philosophical Review*, 85, pp. 297-315.

Nolan, Daniel 1997: 'Impossible worlds: a modest approach', *Notre Dame Journal for Formal Logic*, 38, pp. 535-72.

Peters, Stanley, and Westerståhl, Dag 2006: *Quantifiers in Language and Logic*. Oxford: Clarendon Press.

Salmon, Nathan 1989: 'The logic of what might have been', *Philosophical Review*, 98, pp. 3-34.

Stalnaker, Robert 1968: 'A theory of conditionals', *American Philosophical Quarterly Monographs* 2, pp. 98-112.

Sterken, Rachel 2015: 'Generics, content and cognitive bias', *Analytic Philosophy*, 56, pp. 75-93.

Strawson, Peter 1952: *Introduction to Logical Theory*. London: Methuen.

Vetter, Barbara 2016: 'Williamson modal epistemology, possibility based', *Canadian Journal of Philosophy*, 46, and in Mark McCullagh and Juhani Yli-Vakkuri (eds.), *Williamson on Modality*, London: Routledge.

Williams, J. Robert G. 2010: 'Defending conditional excluded middle', *Noûs*, 44, pp. 650-668.

Williamson, Timothy 2006: 'Indicative versus subjunctive conditionals, congruential versus non-hyperintensional contexts', in E. Sosa and E. Villanueva, eds., *Philosophical Issues*, *Volume 16: Philosophy of Language*, Oxford: Blackwell, pp. 310-333.

Williamson, Timothy 2007: *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson, Timothy 2016: 'Modal Science', *Canadian Journal of Philosophy*, 46, and in Mark McCullagh and Juhani Yli-Vakkuri (eds.), *Williamson on Modality*, London: Routledge.

Williamson, Timothy 201X: 'Counterpossibles', *Topoi*, forthcoming.