On the Substitution of Identicals in Counterfactual Reasoning*

Alexander W. Kocurek Forthcoming in *Noûs*

Abstract. It is widely held that counterfactuals, unlike attitude ascriptions, preserve the referential transparency of their constituents, i.e., that counterfactuals validate the substitution of identicals when their constituents do. The only putative counterexamples in the literature come from counterpossibles, i.e., counterfactuals with impossible antecedents. Advocates of counterpossibilism, i.e., the view that counterpossibles are not all vacuous, argue that counterpossibles can generate referential opacity. But in order to explain why most substitution inferences into counterfactuals seem valid, counterpossibilists also often maintain that counterfactuals with possible antecedents are transparency-preserving. I argue that if counterpossibles can generate opacity, then so can ordinary counterfactuals with possible antecedents. Utilizing an analogy between counterfactuals and attitude ascriptions, I provide a counterpossibilist-friendly explanation for the apparent validity of substitution inferences into counterfactuals. I conclude by suggesting that the debate over counterpossibles is closely tied to questions concerning the extent to which counterfactuals are more like attitude ascriptions and epistemic operators than previously recognized.

1 Introduction

A linguistic environment $\phi(x)$ is said to be ***transparent*** if any two coreferring names can be substituted for x in $\phi(x)$ *salva veritate*; otherwise, $\phi(x)$ is said to be ***opaque***. In other words, $\phi(x)$ is transparent if the following argument is valid for all names a and b:

$$\begin{aligned} &\phi(a) \\ &a = b \\ \therefore\ &\phi(b). \end{aligned}$$

This inference pattern is also known as the ***substitution of identicals***.

Frege [1892] famously pointed out (though not in these terms) that a transparent environment can be transformed into an opaque environment by embedding it under an attitude ascription. For example, let $P(x)$ stand for $\ulcorner x$ is a planet$\urcorner$. Intuitively, $P(x)$ is transparent: for instance, if Hesperus is a planet and Hesperus is Phosphorus, it logically

*Many thanks to Melissa Fusco, Wes Holliday, and three anonymous reviewers for their suggestions for improving this paper. I am grateful for all the feedback received from the audience members at the following venues where a version of this paper was presented: the 20th Oxford Graduate Philosophy Conference (2016), the Richard Wollheim Society at UC Berkeley (2016), the Berkeley-Stanford-Davis conference at UC Davis (2017), the Society for Exact Philosophy at the University of Calgary (2017), and the Central APA at Chicago (2018). I am grateful for comments received on this paper from Michael Longenecker (at the Central APA), Harrison Smith-Jaoudi (at the Berkeley-Stanford-Davis conference), and Timothy Williamson (at the Oxford Graduate Philosophy Conference).

follows that Phosphorus is a planet. So where h stands for 'Hesperus' and p stands for 'Phosphorus', the following argument is valid:

$$\begin{aligned} & P(h) \\ & h = p \\ \therefore\ & P(p). \end{aligned}$$

Now let B_a stand for ⌜Aisha believes that⌝. Intuitively, $\mathsf{B}_a\, P(x)$ is opaque: even if Aisha believes Hesperus is a planet, and even if Hesperus is Phosphorus, it does not logically follow that Aisha believes Phosphorus is a planet. So the following inference is not valid:

$$\begin{aligned} & \mathsf{B}_a\, P(h) \\ & h = p \\ \therefore\ & \mathsf{B}_a\, P(p). \end{aligned}$$

The inference is not valid even if we add that Aisha is a perfect logician, since Aisha still may not believe that Hesperus is Phosphorus.

The main question I want to address in this paper concerns the validity of the substitution of identicals into counterfactuals. Counterfactuals will generally invalidate the substitution of identicals when their constituents do. For instance, where '$\boxright$' stands for the counterfactual conditional, $\mathsf{B}_a\, P(h) \boxright \mathsf{B}_a\, P(x)$ is an opaque environment since $\mathsf{B}_a\, P(x)$ is. But what about the converse? Must a counterfactual validate the substitution of identicals if its constituents do? And if not always, when? In other words, under what conditions do counterfactuals preserve the transparency of their constituents?

At first, it seems plausible that counterfactuals *always* preserve the transparency of their constituents. After all, consider belief ascriptions. Why is it that belief ascriptions tend to generate opaque environments? The reason, quite simply, is that belief ascriptions are at least in part about the way in which we as agents perceive or represent reality—they are about an agent's underlying state of mind. So it is not surprising that the truth of a belief ascription would be sensitive to the representation used to pick out that individual. But counterfactuals are generally not about the way that mind-independent reality is perceived or represented (unless, of course, their constituents are). Rather, counterfactuals are typically concerned with mind-independent reality itself.

The only putative counterexamples in the literature come from counterfactuals with metaphysically impossible antecedents, also known as ***counterpossibles***. According to many of the extant semantic theories of counterfactuals, counterpossibles have trivial truth conditions.[1] On the face of it, though, some counterpossibles seem false. Consider, for instance, the following counterpossibles:

(*Salmon*) If my uncle Sam were a salmon, he would have wings.

[1]To illustrate, take the Lewis-Stalnaker semantics for counterfactuals, according to which $\phi \boxright \psi$ is true just in case (roughly) all of the closest possible ϕ-worlds are ψ-worlds [Stalnaker, 1968; Lewis, 1973]. If ϕ is impossible, then there are no possible ϕ-worlds, and so vacuously all of the closest possible ϕ-worlds are ψ-worlds. Hence, on this semantics, $\neg \Diamond \phi$ implies $\phi \boxright \psi$ for any ψ. A similar result holds for the semantic proposals found in Kratzer 1981, 1989, 2012; Galles and Pearl 1998; Lycan 2001.

(Con_{PA}) If PA were able to prove its own consistency, it would be consistent.

(*Intuitionism*) If intuitionistic logic were correct, logicians would be surprised.

None of these seem obviously true. (*Salmon*) seems contrary to biology; (Con_{PA}) seems in opposition to Gödel's second incompleteness theorem; and while (*Intuitionism*) may be true, it seems largely dependent on contingent sociological facts about logicians. So it is far from obvious that all counterpossibles are trivially true.

Now, let ***counterpossibilism*** be the view that some counterpossibles are true and some are false.[2] A well-known argument due to Williamson [2007] (which we review in § 3) shows that if counterpossibilism is true, then there are counterpossibles with transparent constituents that do not validate the substitution of identicals. Thus, the question of whether counterfactuals preserve the transparency of their constituents depends, at least in part, on whether counterpossibilism is true.

Let us suppose for a moment that counterpossibilism is true, and so a counterfactual with transparent constituents might fail to be transparent. Even so, counterfactuals *appear* to validate the substitution of identicals in a large number of cases. One might wonder, then, what explains the widespread appearance of validity of the substitution of identicals into counterfactuals. Since most of the counterexamples seem to involve counterpossibles, it is natural to explain the apparent validity of the substitution of identicals into counterfactuals by postulating that counterfactuals with possible antecedents always preserve the transparency of their constituents.[3] This principle is sometimes motivated by appealing to the so-called "strangeness of impossibility" constraint, which roughly states that nothing impossible would have obtained had something possible obtained (§ 3).[4]

In this paper, I will argue against this counterpossibilist explanation. That is, I will argue that if counterpossibles can fail to preserve the transparency of their constituents, then so can counterfactuals with possible antecedents. More specifically, I will argue that appealing to the possibility of the antecedent (i) cannot explain the apparent validity of instances of substitution involving counterpossibles (§ 4) and (ii) conflicts with the principle of the simplification of disjunctive antecedents (§ 5). If that is right, then explaining the apparent validity of the substitution of identicals into counterfactuals by restricting the counterexamples to counterpossibles is problematic.

Some might take this argument to constitute a refutation of counterpossibilism altogether, since this would seem to deprive the counterpossibilist any way of explaining the widespread appearance of the validity of the substitution of identicals into counterfactuals. However, I will propose an alternative counterpossibilist explanation for the apparent validity of the substitution of identicals into counterfactuals that does not rely on the modal status of the antecedent of the counterfactual in question and is compatible with the view that counterfactuals are generally opacity-generating (§ 6). This style of explanation is motivated by an analogy between the behaviors of counterfactuals and

[2]For a defense of counterpossibilism, also known as "non-vacuism", see Cohen 1987, 1990; Mares 1997; Nolan 1997; Goodman 2004; Vander Laan 2004; Krakauer 2012; Bjerring 2013; Brogaard and Salerno 2013; Kment 2014; Berto et al. 2017. For criticism, see Popper 1959; Stalnaker 1968; Lewis 1973; Bennett 2003; Williamson 2007, 2017.

[3]See, e.g., Brogaard and Salerno [2013]; Berto et al. [2017] for defenses of this approach.

[4]The name "strangeness of impossibility" comes from Nolan [1997, p. 550].

attitude ascriptions, and it suggests that the counterpossibilist ought to maintain that all counterfactuals have a partially epistemic flavor. Thus, I will conclude by arguing that the debate over counterpossibles is really at its core a debate about whether all counterfactuals are epistemic (§ 7).

2 Background

To start, let me introduce some notation and some background assumptions that are used throughout this paper. To clarify the discussion, I will use the following formal language to regiment a certain natural fragment of English:

$$\phi ::= P(a_1, \ldots, a_n) \mid (a_1 = a_2) \mid \neg\, \phi \mid (\phi \wedge \phi) \mid \Box\, \phi \mid (\phi \boxright \phi) \mid \mathsf{B}_a\, \phi \mid \mathsf{APK}\, \phi,$$

where $a_1, \ldots, a_n$ are names and P is any n-place predicate. We ignore quantifiers for simplicity. As usual, we treat the following symbols as defined from the more basic ones:

$$\begin{array}{rclcrcl} (a \neq b) & := & \neg(a = b) & & (\phi \equiv \psi) & := & (\phi \supset \psi) \wedge (\psi \supset \phi) \\ (\phi \vee \psi) & := & \neg(\neg\, \phi \wedge \neg\, \psi) & & \Diamond\, \phi & := & \neg\, \Box\, \neg\, \phi \\ (\phi \supset \psi) & := & \neg(\phi \wedge \neg\, \psi) & & \mathsf{APC}\, \phi & := & \neg\, \mathsf{APK}\, \neg\, \phi. \end{array}$$

Here is how one should read the operators:

$$\begin{array}{rcl} \Box\, \phi & = & \ulcorner \text{it is (metaphysically) necessary that } \phi \urcorner \\ \Diamond\, \phi & = & \ulcorner \text{it is (metaphysically) possible that } \phi \urcorner \\ \phi \boxright \psi & = & \ulcorner \text{if } \phi \text{ were the case, } \psi \text{ would be the case} \urcorner \\ \mathsf{B}_a\, \phi & = & \ulcorner a \text{ believes that } \phi \urcorner \\ \mathsf{APK}\, \phi & = & \ulcorner \text{it is } \textit{a priori} \text{ knowable that } \phi \urcorner \\ \mathsf{APC}\, \phi & = & \ulcorner \text{it is } \textit{a priori} \text{ possible (or conceivable) that } \phi \urcorner. \end{array}$$

We say a formula is ***atomic*** if it is either of the form $P(a_1, \ldots, a_n)$ or of the form $a_1 = a_2$. We say a formula is ***modal*** if it contains at least one of '$\Box$', '$\boxright$', 'B_a', and 'APK'. We say a formula is ***counterfactual*** if it contains '$\boxright$'. Finally, we say a formula is ***epistemic*** if it contains one of 'B_a' and 'APK'.

None of our discussion will depend on which semantics for counterfactuals one adopts. So where Γ is a set of formulas, I will write "$\Gamma \models \psi$" for "the argument from premises Γ to conclusion ψ is valid", whatever the correct notion of validity is. If $\Gamma = \{\phi_1, \ldots, \phi_n\}$, we may drop set brackets and write "$\phi_1, \ldots, \phi_n \models \psi$" in place of "$\{\phi_1, \ldots, \phi_n\} \models \psi$". Furthermore, we may write "$\models \psi$" in place of "$\varnothing \models \psi$".

If ϕ is a formula, let $\phi(x)$ be the result of replacing any number of names (possibly none) with the variable x. We will call $\phi(x)$ an ***environment***. The formula that results from replacing every occurrence of x in $\phi(x)$ with a name a will be denoted by "$\phi(a)$".[5] We will say an environment $\phi(x)$ is ***transparent*** if for any names a and b:

$$\phi(a), a = b \models \phi(b).$$

[5]Note that $\phi(b)$ need not be the result of replacing *every* occurrence of a in $\phi(a)$ with b. Rather, only those instances of 'a' represented by 'x' in $\phi(x)$ are replaced by 'b' in $\phi(b)$. So $\phi(x)$ might contain an occurrence of 'a' that is not replaced in $\phi(b)$.

Otherwise, we will say $\phi(x)$ is ***opaque***.[6]

We will make some fairly minimal assumptions about the behavior of $\models$ that both sides of the debate can agree on. First, most counterpossibilists are willing to grant the validity of classical reasoning, at least for the sake of discussion. So I will assume we may appeal to classical reasoning without harm throughout. Thus, we assume $\models$ obeys all of the classical structural rules (reflexivity, transitivity, weakening, etc.) and rules of inference (*modus ponens*, the deduction theorem, adjunction, etc.) and that every substitution instance of a propositional tautology is valid.

Second, we assume that the usual non-modal laws of identity all hold. That is, we assume that identity is reflexive and that all predicates (including identity) are transparent:

Self-Identity. $\models a = a$

Leibniz's Law. $\phi(a), a = b \models \phi(b)$ for any atomic $\phi(x)$.

These two principles collectively ensure that identity is an equivalence relation. They also ensure that we are not cheating by sneaking in some opaque environments into modal claims via opaque predicates.[7] By induction, it follows from these two principles that every non-modal environment is transparent, i.e.,

Transparency of Booleans. $\phi(a), a = b \models \phi(b)$ for any non-modal $\phi(x)$.

Thus, opacity can only arise in modal environments.

Third, we assume that metaphysical necessity is transparency-preserving. To make this precise, we adopt two principles of necessity that are widely accepted:

Necessity of Identity. $a = b \models \Box(a = b)$ and $a \neq b \models \Box(a \neq b)$.[8]

Necessitation. If $\phi_1, \ldots, \phi_n \models \psi$, then $\Box\,\phi_1, \ldots, \Box\,\phi_n \models \Box\,\psi$ (for $n \geqslant 0$).

These two principles have some useful consequences. First, by ***Necessitation***, we have:

Possibilization. If $\phi \models \psi$, then $\Diamond\,\phi \models \Diamond\,\psi$.

Second, and perhaps more importantly, if $\phi(x)$ is transparent, then $\Box\,\phi(x)$ is too. That is:

Transparency of Necessity. If $\phi(a), a = b \models \phi(b)$, then $\Box\,\phi, a = b \models \Box\,\phi(b)$.

It follows from ***Transparency of Booleans*** and ***Transparency of Necessity*** by induction that $\phi(a), a = b \models \phi(b)$ if ϕ is non-epistemic and non-counterfactual.

Using this notation, the principle of the substitution of identicals into counterfactuals can be stated as follows:

[6] More generally, where Γ is a set of formulas, we can define $\phi(x)$ to be ***transparent under*** Γ if $\Gamma, \phi(a), a = b \models \phi(b)$ for all names a and b. All of the substitution principles discussed in this paper could just as well be generalized to a premise set Γ in this way without loss, though we will not do so for simplicity.

[7] Potential examples of opaque predicates include $\ulcorner x$ is famous$\urcorner$ and $\ulcorner x$ worships $y\urcorner$.

[8] In Kocurek 2018, I argue against ***Necessity of Identity***. The view I defend also has the consequence that the substitution of identicals into counterfactuals with possible and transparent constituents fails. This paper can be seen as an attempt to argue for the plausibility of this consequence even if one wishes to maintain ***Necessity of Identity***.

Substitution. $\phi(a) \boxrightarrow \psi(a), a = b \vDash \phi(b) \boxrightarrow \psi(b)$ if $\phi(x)$ and $\psi(x)$ are transparent.

We can define ***counterpossibilism*** as the view that rejects the following two principles:

Vacuously True. $\neg \Diamond \phi \vDash \phi \boxrightarrow \psi$.

Vacuously False. $\neg \Diamond \phi \vDash \neg(\phi \boxrightarrow \psi)$.

That is, according to counterpossibilism, some counterpossibles are true while others are false.[9] We will call the rejection of counterpossibilism in this sense ***vacuism***.

It is worth noting that counterpossibilists sometimes argue that *no* non-trivial inference involving counterfactuals is valid.[10] Although this claim is often stated at an informal level, we can articulate it more precisely as follows:

No Counterfactual Logic. If $\phi_1, \ldots, \phi_n \vDash \psi$, then there is a non-counterfactual formula θ such that $\vDash \theta$ and $(\phi_1 \wedge \cdots \wedge \phi_n) \supset \psi$ is the result of uniformly substituting some occurrences of 0-ary atomics in θ with counterfactual formulas.

Informally, ***No Counterfactual Logic*** says that for the purposes of validity, counterfactuals can be treated on a par with propositional atomic formulas. Of course, if that is right, then the main thesis of this paper is trivial, since ***Substitution*** will fail for *all* counterfactuals.

For the purposes of this paper, I will assume that ***No Counterfactual Logic*** is false and that there are some interesting purely counterfactual inferences that are valid. While I think this assumption is justifiable, I do not want to spend much time developing justifications for it now. But I will note that the reason counterpossibilists accept ***No Counterfactual Logic*** typically has to do with counterpossibles whose antecedents are logically impossible, also known as ***counterlogicals***. It is highly controversial, even amongst counterpossibilists, whether counterlogicals are non-vacuous or coherent.[11] In any case, none of the counterpossibles dealt with in this paper are counterlogicals. So even if ***No Counterfactual Logic*** holds, we might interpret $\vDash$ to only codify principles of reasoning that hold in contexts where logical impossibilities are not entertained.

For the most part, we need not take a strong stance on exactly which counterfactual inferences are valid. Whenever I appeal to a counterfactual inference, I will say so explicitly. The only such inference I will unabashedly appeal to throughout is the following:

Triviality. $\vDash \phi \boxrightarrow \phi$.

[9] Strictly speaking, the failure of ***Vacuously True*** and ***Vacuously False*** only implies that some counterpossibles are *possibly* true and others are *possibly* false. We set aside the "counterpossibilist" view that counterpossibles are all *actually* true or all *actually* false. What is more, this definition of counterpossibilism leaves it open for counterpossibilists to accept the following vacuity principle:

Vacuous. $\neg \Diamond \phi \vDash \Box(\phi \boxrightarrow \psi) \vee \Box \neg(\phi \boxrightarrow \psi)$.

According to ***Vacuous***, every counterpossible is either necessary or impossible, but which it may depend on its constituents. No one as far as I am aware accepts ***Vacuous*** but rejects ***Vacuously True*** and ***Vacuously False***. So throughout, I will ignore the version of counterpossibilism that accepts ***Vacuous***.

[10] See, e.g., Cohen 1990, p. 131 and Nolan 1997, p. 554.

[11] For arguments that counterlogicals are vacuous or uninterpretable, see, e.g., Goodman 2004, p. 52 and Kment 2014, p. 73. For arguments that they are not vacuous, see, e.g., Mares 1997, pp. 517–518, Nolan 1997, p. 545, and Brogaard and Salerno 2013, p. 643.

Almost everyone on both sides of the counterpossibles debate accepts ***Triviality***.[12] And in any case, the instances of ***Triviality*** I will appeal to will generally be uncontroversial even for those who reject ***Triviality***.

3 The Explanatory Challenge

We will now briefly review (a) the argument that counterpossibilism implies the failure of ***Substitution*** and (b) the motivation for saving ***Substitution*** for counterfactuals with possible antecedents. We start with an argument due to Williamson [2007, p. 174] that counterpossibilism implies the failure of ***Substitution***.

Consider the following inference (call this the ***Main Argument***):

(M1) If Hesperus were not Phosphorus, Hesperus would not be Phosphorus.

(M2) Hesperus is Phosphorus.

(M3) ∴ If Hesperus were not Phosphorus, Phosphorus would not be Phosphorus.

Schematically, this can be represented in our formal language as follows:

$$\begin{aligned} & h \neq p \mathrel{\square\!\!\rightarrow} \underline{h} \neq p \\ & h = p \\ \therefore\ & h \neq p \mathrel{\square\!\!\rightarrow} \underline{p} \neq p. \end{aligned}$$

By ***Necessity of Identity***, (M1) and (M3) are counterpossibles. And by ***Triviality***, (M1) is trivially true. But if any counterpossible is ever false, then surely (M3) is false—or, at the very least, (M3) *could* be false.[13] If that is right, then the Main Argument is invalid. But the Main Argument is just an instance of ***Substitution***. So counterpossibilism implies the failure of ***Substitution***.

At first, this consequence may not seem so bad, since the Main Argument does not intuitively feel like valid reasoning. But even if we are willing to abandon ***Substitution***, we still need to explain why most ordinary instances of ***Substitution*** at least *appear* to be valid. To illustrate, consider the following example also from Williamson 2007 (call this the ***Rocket Argument***):

(R1) If the rocket had continued on its course, it would have hit Hesperus.

(R2) Hesperus is Phosphorus.

[12] It should be noted that ***Triviality*** is denied by vacuists who hold that all counterpossibles are vacuously *false*, though they will typically maintain a weakened version of ***Triviality***:

Possible Triviality. $\Diamond \phi \models \phi \mathrel{\square\!\!\rightarrow} \phi$.

It is also, of course, denied by counterpossibilists who maintain ***No Counterfactual Logic***. But even those counterpossibilists sympathetic to ***No Counterfactual Logic*** sometimes make an exception for ***Triviality*** [Nolan, 1997, pp. 554–555]. At no point do I appeal to ***Triviality*** to argue against either of these views.

[13] Many counterpossibilists would argue that (M3) is *necessarily* false. See, e.g., Krakauer 2012, pp. 76–78, Bjerring 2013, p. 335, Brogaard and Salerno 2013, pp. 652–653, and Kment 2014, p. 219 for a defense of this view. For potential problems for this view, see Nolan 1997, p. 550 and Vander Laan 2004, p. 271.

(R3) ∴ If the rocket had continued on its course, it would have hit Phosphorus.

The Rocket Argument seems to be intuitively valid, even if ***Substitution*** must be rejected. So if we cannot explain why arguments such as the Rocket Argument generally appear to be valid without appeal to ***Substitution***, then that would be sufficient grounds for rejecting counterpossibilism altogether. Call this ***the explanatory challenge***.

Let me pause to clarify what the explanatory challenge to counterpossibilism is. The problem is not that some (or many) of the instances of substitution into counterfactuals seem to be reliable forms of reasoning. After all, many fallacious forms of reasoning have reliable instances.[14] For instance, the inference from "This is a left shoe or a right shoe" and "This is a left shoe" to "This is not a right shoe" is invalid, but generally reliable. In such cases, however, we can readily explain why the inference is reliable. In this case, the inference is reliable because left shoes are generally not right shoes. The problem is that the counterpossibilist does not seem to have any good explanation for *why* the vast majority of these substitution inferences are reliable. Why are so many instances of substitution into counterfactuals with transparent constituents reliable if the inference is not universally valid?

As an analogy, consider the debate over modus ponens. Kolodny and MacFarlane [2010] argue that, in response to what is now called the miners paradox, we ought to reject the universal validity of modus ponens (for reasons I will not rehearse here). One immediate worry for their view, however, is that modus ponens is so essential for practical and theoretical reasoning that it is difficult to imagine how we could get along without it. Why is modus ponens such a reliable form of reasoning in such a diverse set of contexts? What else, besides the universal validity of modus ponens, could explain this?

Their strategy for addressing this problem is to argue that while modus ponens is not strictly valid, it only fails to be truth-preserving in a very restricted set of circumstances. In particular, their semantics predicts that modus ponens is truth-preserving (a) in categorical contexts (contexts where "one is drawing new conclusions from what one takes to be known facts" (p. 139)) and (b) in hypothetical contexts, so long as the consequent is information-invariant and the antecedent is either world-invariant or information-invariant (p. 142). Since most ordinary contexts where modus ponens is employed fall into one of these two categories, they are able to explain why modus ponens is generally reliable even if it is not unrestrictedly valid.

The explanatory challenge posed to the counterpossibilists is likewise to give an explanation for why substitution into counterfactuals (with transparent constituents) is generally reliable if it is not universally valid. This does not mean that they must explain the apparent validity of every apparently valid instance of ***Substitution*** in one fell swoop. Sometimes, an argument can seem valid due to particular features of the case that have nothing to do with its logical form (e.g., the shoe inference above seems valid simply because of certain background facts about shoes). As long as the counterpossibilist is able to give *some* (non-circular, non-*ad hoc*, etc.) explanation for why an instance of ***Substitution*** is safe when it is, they will have met the explanatory challenge posed here.

Since the only known counterexamples to ***Substitution*** involve counterpossibles, a natural suggestion for how to overcome the explanatory challenge would be to maintain

[14] Thanks to an anonymous reviewer for drawing this point to my attention.

that ***Substitution*** *only* fails when the counterfactuals involved are counterpossibles.[15] If that is right, then it seems as though we can explain why the substitution of identicals into counterfactuals appears to be a safe inference in ordinary cases by replacing ***Substitution*** with the following revised principle:

Possible Substitution. $\phi(a) \boxright \psi(a), a = b, \Diamond \phi(a) \models \phi(b) \boxright \psi(b)$ if $\phi(x)$ and $\psi(x)$ are transparent.

This principle can be motivated by a kind of conservatism with respect to the orthodox logic for counterfactuals: the standard counterfactuals inferences need not be rejected in ordinary circumstances, since in ordinary circumstances, the antecedents of the counterfactuals involved are metaphysically possible. One way to motivate ***Possible Substitution*** along these lines is to appeal to the following principle:[16]

Strangeness of Impossibility. $\Diamond \phi, \Box \psi \models \phi \boxright \psi$.

In words, nothing metaphysically impossible would have obtained had something metaphysically possible obtained. It turns out that ***Possible Substitution*** is actually entailed by ***Strangeness of Impossibility***, given certain plausible counterfactual principles (see § A). Thus, one might see ***Possible Substitution*** as a special case of a more general principle that is both plausible and accepted by a number of counterpossibilists.

Given ***Possible Substitution***, we can now explain the apparent validity of many instances of substitution into counterfactuals: those instances generally involve counterfactuals with possible antecedents. The Rocket Argument, for instance, seems valid because we implicitly assume the premise that the rocket could have continued on its course, which, if added, would render the argument explicitly valid.[17] And in most contexts, such an assumption is justified. But the Main Argument seems (and is) invalid because Hesperus is necessarily Phosphorus; and so (M1) is a counterpossible, in which case ***Possible Substitution*** does not apply. Thus, one might hope to account for the general reliability of ***Substitution*** by appealing to ***Possible Substitution***.[18]

[15] See Brogaard and Salerno 2013, p. 657 and Berto et al. 2017, p. 12 for this approach.

[16] This principle was introduced by Nolan [1997, p. 550]. The principle is usually stated in terms of the Lewis-Stalnaker semantics: no impossible world is as close to the actual world as any possible world. For defenses of this principle, see Mares 1997, pp. 521–522, Kment 2014 and Berto et al. 2017, p. 6. For criticisms, see Nolan 1997, p. 550, Vander Laan 2004, p. 271, and Bernstein 2016, p. 7.

[17] Berto et al. [2017, p. 12] argue that while the Rocket Argument is necessarily truth-preserving, it is not strictly valid, since it does not explicitly include this possibility premise. While considerations surrounding the semantic paradoxes might give us reason for denying that validity is necessary truth-preservation, such divergence does not concern us here, since we are assuming the validity of classical logic. Given the validity of the 5 axiom ($\Diamond \phi \supset \Box \Diamond \phi$), if the rocket could have continued on its course, then it necessarily could have, and so the Rocket Argument is already necessarily truth-preserving given ***Possible Substitution***. For our purposes, the issue is moot since I will argue in § 6 that the Rocket Argument is not even truth preserving.

[18] Williamson [2017, p. 16] has recently argued against this line of defense. His main worry is that the semantics for counterfactuals endorsed by a counterpossibilist who accepts ***Possible Substitution*** will likely be gerrymandered and disunified:

> [For these counterpossibilists,] 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.

Unfortunately, there are a variety of problems for this counterpossibilist explanation. While these problems do not show that ***Possible Substitution*** is false, I argue that they do motivate searching for an alternative explanation for the widespread apparent validity of substitution inferences into counterfactual environments.

4 Problem 1: Incompleteness

The first problem involves apparently valid instances of ***Substitution*** involving counterpossibles. Not all counterpossibles seem opaque, as the following argument illustrates (call this the ***Superman Argument***):[19]

(S1) If Superman and I had the same parents, then I would have had a brother.

(S2) Superman is Clark Kent.

(S3) ∴ If Clark Kent and I had the same parents, then I would have had a brother.

Let us suppose, for the sake of discussion, that it is impossible for Superman and I to have the same parents. (It is straightforward to modify the example if one rejects this.) Hence, (S1) and (S3) are counterpossibles and so ***Possible Substitution*** does not apply to this argument. Nevertheless, the argument seems valid. So even if we explain the apparent validity of arguments such as the Rocket Argument using ***Possible Substitution***, we will still need a further explanation for why many instances of ***Substitution*** involving counterpossibles such as the Superman Argument seem valid. Call this ***the incompleteness problem***.

To be clear, this worry does not call into question ***Possible Substitution*** as a principle of counterfactual resaoning. Though I will raise doubts about ***Possible Substitution*** later (§ 6), we may grant it holds for the sake of argument. The worry is that ***Possible Substitution*** does not fully address the explanatory challenge. While the counterpossibilist can explain the felt validity of instances of ***Substitution*** involving counterfactuals with possible antecedents, that explanation does not account for the Superman Argument, or other instances of substitution into counterpossibles that seem valid. If the counterpossibilist cannot supplement ***Possible Substitution*** with other plausible principles that explain why these instances of substitution into counterpossibles seem valid, then explanatory challenge still stands, and counterpossibilism is still in trouble.

In § 6, I will propose a solution to the explanatory challenge on behalf of the counterpossibilist that avoids the incompleteness problem (one that, as it so happens, will call into question the solution invoking ***Possible Substitution***). Here, as motivation for that proposal, I want to consider some natural alternative responses to this problem that appeal to *a priori* possibility, and demonstrate their failure.[20]

Counterpossibilists have in turn responded by proposing what seem to be fairly unified semantic theories for counterfactuals that both tolerate non-vacuous counterpossibles and validate ***Possible Substitution***. See Brogaard and Salerno 2013; Berto et al. 2017 for examples.

[19] This argument comes from Kallestrup [2009, fn. 11].

[20] Brogaard and Salerno [2013, fn. 11] respond to the incompleteness problem as follows:

> Hyperintensional contexts do not always resist substitution even if they sometimes do. Of

One response to the incompleteness problem is to maintain the epistemic analogue of ***Possible Substitution***, viz., that substitution is licensed in the scope of counterpossibles when their antecedents are *a priori* possible:[21]

Conceivable Substitution. $\phi(a) \boxdot\!\!\rightarrow \psi(a), a = b, \mathsf{APC}\,\phi(a), \mathsf{APC}\,\phi(b) \models \phi(b) \boxdot\!\!\rightarrow \psi(b)$ if $\phi(x)$ and $\psi(x)$ are transparent.

Since it is *a priori* possible for Superman and I to have the same parents, ***Conceivable Substitution*** would validate the Superman Argument. Unfortunately, it would also validate the Main Argument. After all, it is *a priori* possible that Hesperus is not Phosphorus. Thus, by ***Conceivable Substitution***, substitution would be licensed in (M1), and thus the Main Argument would be valid. Hence, if the counterpossibilist wants to reject the validity of the Main Argument, they cannot endorse ***Conceivable Substitution***.

Here is a more subtle way to explain the validity of the Superman Argument by appealing to the *a priori*. Generally, when one makes a counterfactual supposition, one tries to make as small of a change to the actual world as possible while accommodating the supposition. While it is notoriously difficult to make this idea precise,[22] it is natural to think that in making a counterfactual supposition, one should preserve as many *necessary* truths as is feasible. Of course, counterpossibilism states that one cannot preserve every necessary truth under counterfactual suppositions that are metaphysically impossible. Nevertheless, one should still aim to maximize the number of necessary truths that are preserved under such a supposition. For instance, on the counterfactual supposition that Superman and I had the same parents, one should maintain (say) that modal realism is false (assuming it is actually false), since Superman and I having the same parents does not contradict the falsity of modal realism. More generally, if there is no *a priori* reason to reject a necessary truth on a counterfactual supposition, one should hold that necessary truth fixed. This idea can be codified into a principle as follows:

Minimize Impossibility. $\Box\,\theta, \neg\,\mathsf{APK}(\phi \supset \neg\,\theta) \models \phi \boxdot\!\!\rightarrow \theta$.

Now, if $\psi(x)$ is transparent, then $a = b \models \Box(\psi(a) \supset \psi(b))$. Hence, by ***Transparency of Necessity*** and ***Minimize Impossibility***, we have the following:

$$a = b, \neg\,\mathsf{APK}(\phi \supset a \neq b) \models \phi \boxdot\!\!\rightarrow (\psi(a) \supset \psi(b)).$$

course, we need to give a principled account of when counterfactuals create opaque contexts. They create opaque context when the antecedent or consequent which result 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'.

But this response does not look promising: "Clark Kent and I have the same parents" is not an *a priori* implication of "Superman and I have the same parents" since it is not an *a priori* implication of either of these that Superman is Clark Kent. Moreover, this explanation would imply that the Main Argument is valid by symmetric considerations. So this cannot be what explains the apparent validity of the Superman Argument. I suspect the second proposal sketched in this section could be construed as a charitable reconstruction of their general idea.

[21] Thanks to an anonymous reviewer for encouraging me to say more about this proposal.

[22] See, e.g., Goodman 1947; Lewis 1973; Kratzer 2012 for discussion.

Thus, assuming some plausible counterfactual principles, one can show that ***Minimize Impossibility*** entails the following (see § A for the proof):

A Priori Substitution. $\phi(a) \boxright \psi(a), a = b, \neg \mathsf{APK}(\phi(a) \supset a \neq b), \neg \mathsf{APK}(\phi(b) \supset a \neq b) \models \phi(b) \boxright \psi(b)$ if $\phi(x)$ and $\psi(x)$ are transparent.

In other words: substitution is licensed into counterfactuals so long as their antecedents (and substituted instances) are not *a priori* incompatible with the identity premise. This principle would explain why the Superman Argument seems valid since neither Superman and I having the same parents nor Clark Kent and I having the same parents *a priori* imply that Superman is not Clark Kent. It is also compatible with the invalidity of the Main Argument, since the antecedent of (M1) is just the explicit negation of "Hesperus is Phosphorus".

In light of ***Necessity of Identity***, ***A Priori Substitution*** can also be viewed as an epistemic analogue of ***Possible Substitution***. One way to formulate ***Possible Substitution*** is as follows:

$$\phi(a) \boxright \psi(a), a = b, \neg \Box(\phi(a) \supset a \neq b), \neg \Box(\phi(b) \supset a \neq b) \models \phi(b) \boxright \psi(b).$$

Under the assumption that $a = b$, the additional premises in this inference pattern (viz., $\neg \Box(\phi(a) \supset a \neq b)$ and $\neg \Box(\phi(b) \supset a \neq b)$) are equivalent to the possibility premise (viz., $\Diamond \phi(a)$) in ***Possible Substitution***. Thus, ***A Priori Substitution*** can be viewed as the epistemic version of this alternative formulation of ***Possible Substitution***. The difference is that ***A Priori Substitution*** is not equivalent to ***Conceivable Substitution*** since identity facts are not generally knowable *a priori*.[23]

Unfortunately, while ***A Priori Substitution*** does not technically render the Main Argument valid, it validates a closely related argument. Consider the following slightly modified version of the Main Argument (call this the ***Venus Argument***):

(V1) If Hesperus were not Phosphorus, Hesperus would not be Phosphorus.

(V2) Hesperus is Venus.

(V3) Venus is Phosphorus.

(V4) ∴ If Hesperus were not Phosphorus, Phosphorus would not be Phosphorus.

This argument is regimented in our formal language as follows:

$$\begin{aligned} h \neq p \boxright \underline{h} \neq p \\ h = v \\ v = p \\ \therefore h \neq p \boxright \underline{p} \neq p. \end{aligned}$$

[23]Specifically, the proof that $\Diamond \phi(a)$ implies $\neg \Box(\phi(a) \supset a \neq b)$ relies essentially on ***Necessity of Identity***. Thus, without the *a priori* analogue of ***Necessity of Identity***, one cannot show that ***A Priori Substitution*** implies ***Conceivable Substitution***.

In other words, the Venus Argument is just the Main Argument except we split "Hesperus is Phosphorus" into two separate identity claims using 'Venus'. Arguably, if the Main Argument is invalid, the Venus Argument is not valid either. However, the Venus Argument is valid if ***A Priori Substitution*** holds.

First, notice that if no substitution occurs in the antecedent, then ***A Priori Substitution*** reduces to the following simplified principle:

A Priori Substitution in Consequent. $\phi \boxdot\!\!\rightarrow \psi(a), a = b, \neg \mathsf{APK}(\phi \supset a \neq b) \vDash \phi \boxdot\!\!\rightarrow \psi(b)$ if $\psi(x)$ is transparent.

Informally: substitution is licensed into the consequents of counterfactuals so long as their antecedents are not *a priori* incompatible with the identity claim. Now, the Venus Argument can be broken up into two parts, each of which is valid if ***A Priori Substitution in Consequent*** holds. The first part involves substitution via the first identity claim (call this part the ***Venus Argument Part 1***):

(V1) If Hesperus were not Phosphorus, Hesperus would not be Phosphorus.

(V2) Hesperus is Venus.

(V2.5) ∴ If Hesperus were not Phosphorus, Venus would not be Phosphorus.

Since the claim that Hesperus is not Phosphorus is not *a priori* incompatible with the claim that Hesperus is Venus (that is, $\neg \mathsf{APK}(h \neq p \supset h \neq v)$), (V2.5) follows from (V1) and (V2) by ***A Priori Substitution in Consequent***. Using (V2.5), we can now infer our conclusion (V4) via the following inference (call this part the ***Venus Argument Part 2***):

(V2.5) If Hesperus were not Phosphorus, Venus would not be Phosphorus.

(V3) Venus is Phosphorus.

(V4) ∴ If Hesperus were not Phosphorus, Phosphorus would not be Phosphorus.

Again, since the claim that Hesperus is not Phosphorus is not *a priori* incompatible with the claim that Venus is Phosphorus (that is, $\neg \mathsf{APK}(h \neq p \supset v \neq p)$), (V4) follows from (V2.5) and (V3) by ***A Priori Substitution in Consequent***. Stitching both parts together, we obtain the validity of the Venus Argument.

This shows that if the counterpossibilist does not want to accept the validity of the Venus Argument, they must reject ***A Priori Substitution in Consequent***, and hence ***A Priori Substitution***. In that case, they cannot appeal to ***A Priori Substitution*** to explain the felt validity of the Superman Argument. An alternative explanation must be sought for why the Venus Argument is invalid while the Superman Argument seems valid. The trouble is it is not at all clear what other explanation appealing to the *a priori* can do both jobs.

A natural thought to have about the Venus Argument is the following.[24] In effect, the Venus Argument is just an indirect way of implementing the substitution of 'Hesperus' with 'Phosphorus' in the Main Argument that utilizes two identity claims rather than

[24]Thanks to an anonymous reviewer for suggesting this possibility.

one. So perhaps the reason substitution is not licensed is that the antecedent of the counterfactual premises are incompatible with the *all* the identity claims used in these two substitutions *taken together*. That is, the Venus Argument is invalid because the antecedent of (V1) ($h \neq p$) is *a priori* incompatible not with either of the identity claims taken individually ($h = v$, $v = p$) but with their conjunction ($h = v \wedge v = p$).

The problem is that it is not clear how this observation can be used to invalidate either of the two parts of the Venus Argument taken separately. Suppose, just for the sake of argument, that the counterpossibilist objects to the Venus Argument Part 1. Consider that part of the argument by itself as its own argument. How does the fact that the antecedent of (V1) is *a priori* incompatible with the conjunction of (V2) and some other claim not stated in the argument explain why the Venus Argument Part 1 is invalid?

Perhaps we can explain this by postulating that substitution is licensed when the antecedent of the counterfactual premise is not *a priori* incompatible with the identity premise *together* with the relevant background facts. Since one relevant background fact is the fact that Venus is Phosphorus, the antecedent of the counterfactual premise (V1), the identity premise (V2), and the relevant background facts are *a priori* incompatible. The worry with this approach, however, is that it inadvertently fails to validate the Superman Argument. After all, one relevant background fact to consider with regards to the Superman Argument is the fact that I have such-and-such parents and Clark Kent does not have those parents. So even here, the antecedent of the counterfactual premise (S1), the identity premise (S2), and the relevant background facts are *a priori* incompatible. Thus, I do not see anyway of spelling out what counts as a "relevant background fact" that simultaneously invalidates the Venus Argument and explains the felt validity of the Superman Argument.

In sum, it is unclear how the counterpossibilist can appeal to the notion of *a priori* to give a complete and general explanation for why many instances of ***Substitution*** involving counterpossibles seem valid. The complications arising from this approach are far too great, and, in any case, there is a simpler and more unified solution to the explanatory challenge at their disposal (one that does away with appeals to ***Possible Substitution***).

Not only is there a worry that the possibility premise in ***Possible Substitution*** is an incomplete answer to the explanatory challenge, there is also a worry that it is an idle one. Remember, the claim is that the Rocket Argument seems valid because we are implicitly assuming a missing premise, viz., that the rocket could have continued on its course. This suggests that if we explicitly deny that premise in the Rocket Argument, the argument should no longer seem valid. But that does not seem to be the case. Suppose we modify the Rocket Argument in this way (call this the ***Impossible Rocket Argument***):

(IR1) If the rocket had continued on its course, it would have hit Hesperus.

(IR2) Hesperus is Phosphorus.

(IR3) It is impossible for the rocket to have continued on its course.

(IR4) ∴ If the rocket had continued on its course, it would have hit Phosphorus.

Does this argument feel any less valid than the original Rocket Argument? Presumably not. Both arguments seem equally valid—our intuitions do not seem sensitive to the possibility of the rocket continuing on its course. But if that is right, then how could we

possibly explain the felt validity of the Rocket Argument by appealing to the assumption that the rocket could have continued on its course? It does not seem that we can.

5 Problem 2: Simplification of Disjunctive Antecedents

We saw above that using ***Possible Substitution*** to explain the felt validity of the Rocket Argument does not explain the felt validity of arguments like the Superman Argument, which involve valid substitution into counterpossibles. This is a general problem for any attempt to simultaneously reject ***Substitution*** and explain the felt validity of arguments such as the Rocket Argument using ***Possible Substitution***.

The next worry is less general than the previous one. In particular, it only applies to those who accept the ***S***implification of ***D***isjunctive ***A***ntecedents:

SDA. $\models [(\phi \vee \psi) \boxrightarrow \theta] \equiv (\phi \boxrightarrow \theta) \wedge (\psi \boxrightarrow \theta)$.

Although ***SDA*** does not hold on all semantic theories for counterfactuals, many have found ***SDA*** to be a quite plausible counterfactual principle.[25] Moreover, there are reasons for being sympathetic to ***SDA*** if one is sympathetic to counterpossibilism. After all, counterpossibilism implies that counterfactuals are generally representation-sensitive in the sense that counterfactuals are sensitive to the differences between representations of an individual. ***SDA*** also codifies a kind of representational sensitivity, albeit one of a slightly different flavor: it tells us that counterfactuals are sensitive to differences in the logical representations of their antecedents.

Here is the rough idea. Consider the following counterfactual principle:

Replacement. If $\phi \dashv\models \phi'$, then $\phi \boxrightarrow \psi \dashv\models \phi' \boxrightarrow \psi$.

It is well-known that ***SDA*** and ***Replacement*** together entail the following principle:[26]

Antecedent Strengthening. $\phi \boxrightarrow \theta \models (\phi \wedge \psi) \boxrightarrow \theta$.

But many find ***Antecedent Strengthening*** counterintuitive.[27] For example, the following inference seems invalid:

(K1) If I were a karate master, I would have won the fight.

(K2) ∴ If I were a karate master and had broken my leg earlier, I would have won the fight.

Thus, if one wants to reject ***Antecedent Strengthening***, then either ***SDA*** or ***Replacement*** has to go. In light of this, if one wants to keep ***SDA***, one must admit that counterfactuals are sensitive to differences in the logical representation of their antecedents. So even though

[25] For a defense of ***SDA***, see, e.g., Fine 1975, 2012; Nute 1975, 1978; Ellis et al. 1977; Alonso-Ovalle 2006, 2008; Starr 2014; Santorio Forthcoming. For criticisms, see, e.g., Loewer 1976; Lewis 1977; Mckay and van Inwagen 1977; Warmbrōd 1981.

[26] Proof: $\phi \dashv\models \phi \vee (\phi \wedge \psi)$ by classical reasoning, so $\phi \boxrightarrow \theta \dashv\models (\phi \vee (\phi \wedge \psi)) \boxrightarrow \theta \models (\phi \wedge \psi) \boxrightarrow \theta$.

[27] See Lewis 1973. With that said, von Fintel [2001] argues that ***Antecedent Strengthening*** is valid when entailment is interpreted in terms of Strawson entailment. See Gillies 2007; Moss 2012; Lewis 2017 for discussion.

the counterpossibilist is committed to a different kind of representational sensitivity for counterfactuals, it is natural for a counterpossibilist to want to at least be open to the possibility of maintaining ***SDA***.

Unfortunately, those who accept ***Possible Substitution*** but reject ***Substitution*** cannot accept ***SDA***. To see why, consider the following argument (call this the ***Planet Argument***):

(P1) If Hesperus were not Phosphorus, Hesperus would be the second planet from the sun.

(P2) Hesperus is Phosphorus.

(P3) It is possible for Hesperus to be the second planet from the sun.

(PC) ∴ If Hesperus were not Phosphorus, Phosphorus would be the second planet from the sun.

Intuitively, if the Main Argument is invalid, then the Planet Argument is also invalid. For instance, it seems that the premises of the Planet Argument are consistent with the premise "If Hesperus were not Phosphorus, Hesperus and Phosphorus would not be equidistant from the sun"; and if we add this premise to the argument, the invalidity of the argument is even more apparent. However, ***Possible Substitution*** and ***SDA*** entail that the Planet Argument is valid.

The reason is simple: if you disjoin an impossible claim and a possible claim, the resulting disjunction is possible, which one can then employ in simplification inferences. More precisely, where $Sec(x)$ stands for $\ulcorner x$ is the second planet from the sun$\urcorner$, we have:

(P1) $h \neq p \boxright Sec(h)$ (premise)

(P2) $h = p$ (premise)

(P3) $\Diamond Sec(h)$ (premise)

(P4) $\Diamond(h \neq p \vee Sec(h))$ (***Possibilization***, (P3))

(P5) $Sec(h) \boxright Sec(h)$ (***Triviality***)

(P6) $(h \neq p \vee Sec(h)) \boxright Sec(h)$ (***SDA***, (P1), (P5))

(P7) $(h \neq p \vee Sec(h)) \boxright Sec(p)$ (***Possible Substitution***, (P2), (P4), (P6))

(PC) $h \neq p \boxright Sec(p)$ (***SDA***, (P7))

So if the Planet Argument is invalid, then we must give up either ***Possible Substitution*** or ***SDA***.

Rejecting ***SDA*** is not out of the question: ***SDA*** is as controversial as a counterfactual principle gets! So this argument is admittedly not, by any means, a knockdown objection to ***Possible Substitution***. But it does put the counterpossibilist who uses ***Possible Substitution*** to explain substitution failures in an uncomfortable position. For one thing, as we noted above, the counterpossibilist has reasons to be sympathetic to ***SDA*** given that it

embodies a kind of representational sensitivity. While they need not adopt ***SDA***, neutrality would be preferred if the counterpossibilist can achieve it.

Moreover, the conclusion (P7) already seems bad enough. It would already be highly problematic if the counterpossibilist were committed to the validity of the argument from (P1)–(P3) to (P7), since even that argument does not seem valid. But note that while ***SDA*** is invoked in deriving (P6), we only use the uncontroversial direction of ***SDA***, viz., the right-to-left direction. Even the opponents of ***SDA*** agree that this direction of ***SDA*** holds.[28] So we do not need the controversial direction of ***SDA*** to create problems for those who explain the apparent validity of ***Substitution*** with ***Possible Substitution***: the uncontroversial direction of ***SDA*** already causes problems.

6 A Counterfactual Substitution Principle

In the previous sections, I argued against using ***Possible Substitution*** to explain the apparent validity of the substitution of identicals into counterfactuals. But I do not think that this is the end of the road for counterpossibilism. In this section, I will sketch a more unified counterpossibilist explanation that can account for why most instances of the substitution of identicals in counterfactuals are reliable. On the explanation I propose, the apparent validity of arguments like the Rocket Argument have nothing to do with the possibility of their antecedents. In fact, as I explain below, counterfactuals with possible and transparent constituents may nevertheless fail to be transparent.

To see how the explanation works, it helps to compare counterfactuals with other opaque environments, such as those involving belief ascriptions. Let us first answer an easy question: why do belief ascriptions seem to generate opaque environments? For example, let us assume throughout that Aisha is a perfect logician, believing all the logical consequences of her beliefs.[29] Given this, why does it not follow from the premise that Aisha believes that Hesperus is a planet together with the premise that Hesperus is Phosphorus that Aisha also believes Phosphorus is a planet? That is, why is the following argument invalid?

$$\begin{aligned} &\mathsf{B}_a\, P(h) \\ &h = p \\ \therefore\ &\mathsf{B}_a\, P(p). \end{aligned}$$

Intuitively, the answer is this: while Hesperus is, in fact, Phosphorus, Aisha might fail to believe that. Even if Aisha is a perfect logician, she can still fail to believe that Hesperus is Phosphorus since that fact is only knowable *a posteriori*. But once we add the premise

[28] There are some who reject even this direction of ***SDA***, but only for reasons that do not affect the points made here. For instance, Briggs [2012, p. 156] defends a semantics on which $(\phi \vee \psi) \boxrightarrow \theta$ is equivalent to $(\phi \boxrightarrow \theta) \wedge (\psi \boxrightarrow \theta) \wedge (\phi \wedge \psi \boxrightarrow \theta)$. If that is right, then we cannot infer $(\phi \vee \psi) \boxrightarrow \theta$ from $\phi \boxrightarrow \theta$ and $\psi \boxrightarrow \theta$ alone: we must also check that $(\phi \wedge \psi) \boxrightarrow \theta$ holds. Fortunately, this does not affect our inference to (P6), since the premise we would need to add, viz., $(h \neq p \wedge Sec(h)) \boxrightarrow Sec(h)$, is trivially true. So our application of ***SDA*** in inferring (P6) is still licensed, even on this view.

[29] If needed, the reader is free to interpret the inferences in this section that involve belief ascriptions as carrying with them an enthymematic premise that Aisha believes the logical consequences of her beliefs.

that she believes Hesperus is Phosphorus, the argument becomes valid—regardless of whether Hesperus really is Phosphorus. In other words, while the argument above is fallacious, the argument below is not (assuming Aisha is a perfect logician):

$$\begin{aligned} &\mathsf{B}_a\, P(h) \\ &\mathsf{B}_a(h = p) \\ \therefore\ &\mathsf{B}_a\, P(p). \end{aligned}$$

Consider now opaque environments involving *a priori* knowability claims. Clearly, it is *a priori* knowable that if Hesperus is a planet, then Hesperus is a planet. And Hesperus is Phosphorus. But it does not follow that it is *a priori* knowable that if Hesperus is a planet, then Phosphorus is a planet. That is, the following argument is invalid:

$$\begin{aligned} &\mathsf{APK}(P(h) \supset P(h)) \\ &h = p \\ \therefore\ &\mathsf{APK}(P(h) \supset P(p)). \end{aligned}$$

The reason, again, is simply that it is not *a priori* knowable that Hesperus is Phosphorus. If we replaced the premise that Hesperus is Phosphorus with the (false) premise that it is *a priori* knowable that Hesperus is Phosphorus, then the resulting argument would be valid (though unsound). That is, while the previous argument is not valid, the following argument is valid:

$$\begin{aligned} &\mathsf{APK}(P(h) \supset P(h)) \\ &\mathsf{APK}(h = p) \\ \therefore\ &\mathsf{APK}(P(h) \supset P(p)). \end{aligned}$$

What I want to suggest is that a counterpossibilist should say the same thing about counterfactuals. Consider, for instance, the Rocket Argument, which can be formalized as something like the following (where $Cont(x)$ stands for $\ulcorner x$ continued on its course$\urcorner$, $Hit(x, y)$ stands for $\ulcorner x$ hit $y\urcorner$, and r is a name for the rocket):

$$\begin{aligned} &Cont(r) \mathrel{\square\!\!\rightarrow} Hit(r, h) \\ &h = p \\ \therefore\ &Cont(r) \mathrel{\square\!\!\rightarrow} Hit(r, p). \end{aligned}$$

While this argument might seem valid at first, it implicitly relies on a hidden unstated premise, viz., that Hesperus would still be Phosphorus had the rocket continued on its course. Thus, while the argument above is not strictly valid, the following argument is valid (again, regardless of whether Hesperus is, in fact, Phosphorus):

$$\begin{aligned} &Cont(r) \mathrel{\square\!\!\rightarrow} Hit(r, h) \\ &Cont(r) \mathrel{\square\!\!\rightarrow} h = p \\ \therefore\ &Cont(r) \mathrel{\square\!\!\rightarrow} Hit(r, p). \end{aligned}$$

This idea can be codified more precisely into a (simplified) counterfactual substitution principle (we will return to the more general version covering substitution in the antecedents of counterfactuals in a moment):

Counterfactual Substitution in Consequent. $\phi \boxrightarrow \psi(a), \phi \boxrightarrow a = b \vDash \phi \boxrightarrow \psi(b)$ if $\psi(x)$ is transparent.

It is very natural to implicitly assume this missing premise ($\phi \boxrightarrow a = b$) in most substitution inferences involving counterfactuals—so natural that it is hardly worth stating explicitly. If the antecedent of a counterfactual is not relevant to the truth of an identity claim that obtains, then there is no reason to assume that if the antecedent had obtained, the identity might have failed to obtain. In the Rocket Argument, for example, without any further information, the rocket's course does not seem to have anything to do with Hesperus being Phosphorus. Thus, it seems very plausible to assume that had the rocket continued on its course, Hesperus would still be Phosphorus. And if we added that premise to the argument explicitly, ***Counterfactual Substitution in Consequent*** would render the resulting argument valid.

Contrast this with the Main Argument. There, the very first premise is an explicit denial of the missing premise needed for ***Counterfactual Substitution in Consequent*** to apply, viz., if Hesperus had not been Phosphorus, Hesperus *would* be Phosphorus. In other words, to render the Main Argument valid using ***Counterfactual Substitution in Consequent***, one would have to add an apparently absurd premise of the form $\phi \boxrightarrow \neg \phi$. So if ***Counterfactual Substitution in Consequent*** were the principle underlying these judgments, we would not expect the Main Argument to sound plausible at all. Similar reasoning applies to the Venus Argument: the argument is not valid unless we can assume that if Hesperus were not Phosphorus, Hesperus and Phosphorus would still be Venus, which seems patently false. Thus, ***Counterfactual Substitution in Consequent*** does not predict the Venus Argument would be (or even seem) valid.

Counterfactual Substitution in Consequent does not imply either ***Possible Substitution*** or ***A Priori Substitution in Consequent*** without additional assumptions.[30] Nonetheless, this missing counterfactual premise ($\phi \boxrightarrow a = b$) is often (though not always) safe to assume when the antecedent of the counterfactual is possible—which would explain why ***Possible Substitution*** seemed so natural in the first place. And when the antecedent is impossible, it is still often (though not always) safe to assume when there is a lack of an *a priori* connection between the antecedent and the identity claim—which would explain the appeal of ***A Priori Substitution in Consequent***. So ***Counterfactual Substitution in Consequent*** seems to capture what was intuitive about these alternative substitution principles without their costs.

To see that ***Counterfactual Substitution in Consequent*** is really what is doing the work

[30] We could derive ***Possible Substitution*** from ***Counterfactual Substitution in Consequent*** if the following holds for all a and b:

$$a = b, \Diamond \phi \vDash \phi \boxrightarrow a = b.$$

Likewise, we could derive ***A Priori Substitution in Consequent*** from ***Counterfactual Substitution in Consequent*** if the following holds for all a and b:

$$a = b, \neg \mathsf{APK}(\phi \supset a \neq b) \vDash \phi \boxrightarrow a = b.$$

While these might seem reasonable, arguments such as the Revised Rocket Argument to follow suggest they are not universally valid. And since ***A Priori Substitution in Consequent*** validates the Venus Argument, I think the plausibility of ***Counterfactual Substitution in Consequent*** counts against the principle above.

in explaining the apparent validity of arguments like the Rocket Argument, it helps to see how things change when we explicitly deny this hidden premise. Thus, consider the following argument (call this the ***Revised Rocket Argument***):

(RR1) If the rocket had continued on its course, it would have hit Hesperus.

(RR2) Hesperus is Phosphorus.

(RR3) The rocket could have continued on its course.

(RR4) If the rocket had continued on its course, Hesperus would not be Phosphorus.

(RR5) ∴ If the rocket had continued on its course, it would have hit Phosphorus.

Suppose a rocket with a completely accurate map of our entire solar system is heading towards the object labeled 'Hesperus' on its map. On its way, it receives the following instructions from home base: "Check your database to see whether Hesperus is Phosphorus. If Hesperus is not Phosphorus, continue on your course. Otherwise, abort and come back home." (We should also say the rocket will go wherever it is instructed to go without complications—it has plenty of fuel, it is functioning properly, no asteroids are heading toward it, and so forth.) The rocket checks its database, sees that Hesperus is Phosphorus, and thus aborts mission and heads back to Earth.

In this scenario, it seems that all of the premises are true, but the conclusion is false. (RR1) seems true, since the rocket was already heading towards Hesperus. (RR2) is given as part of the case. (RR3) is obviously true; after all, it is metaphysically possible for the rocket to have received no instructions from headquarters or to have a glitch in the computer, in which case it would have continued on its course as before. (RR4) seems true, since the instructions told the rocket to continue on its course if and only if Hesperus was not Phosphorus. And yet (RR5) seems false: the rocket will continue on its course if and only if Hesperus is not Phosphorus, and in that case, its trajectory is headed towards Hesperus. So if the rocket had continued on its course, it would not have hit Phosphorus. Insofar as one has counterpossibilist intuitions, then, the Revised Rocket Argument argument will seem intuitively invalid.

One might object that (RR4) cannot possibly be true given the truth of (RR3). If it is really metaphysically possible for the rocket to continue on its course, then it simply cannot be true that Hesperus would not be Phosphorus had the rocket continued on its course. For if (RR4) were true, that would mean that something impossible would have obtained had something that is possible obtained. Instead, something else would have had to happen, such as the rocket going on the fritz or getting hit by an asteroid (even though we tried to stipulate away all such complications).

But notice that this kind of response assumes ***Strangeness of Impossibility***:

Strangeness of Impossibility. $\Diamond\phi, \Box\psi \models \phi \boxrightarrow \psi$.

For in saying that (RR4) must be false, one must rely on the fact that (a) the rocket could have continued on its course, and (b) necessarily, Hesperus is Phosphorus. Otherwise, there does not seem to be a principled reason for holding that the premises of the Revised Rocket Argument could not hold all at once in the scenario described above. And as noted

in § 3, under plausible counterfactual principles, ***Strangeness of Impossibility*** entails ***Possible Substitution***. So this response would be fine for someone sympathetic to ***Possible Substitution***. But the question is just whether our commitment to the validity of ***Possible Substitution*** is stronger than the intuition that the Revised Rocket Argument is invalid. We already saw in the previous sections that maintaining ***Possible Substitution*** solely for the sake of explaining failures of substitution is problematic. So the main independent motivation for holding on to ***Possible Substitution*** in light of this putative counterexample seems to be a prior commitment to a more general principle such as ***Strangeness of Impossibility***.

Now is a good time to point out, however, that there are alleged counterexamples to ***Strangeness of Impossibility***. Consider, for instance:[31]

(*Lewis*) If Lewis were right about modality, modal realism would be true.

Supposing modal realism is actually false, it is necessarily false. Moreover, this counterfactual seems true (whereas "If Lewis were right about modality, modal realism would be false" does not). But the antecedent of this counterfactual is perfectly possible; Lewis *could* have been right about modality—that is, it is metaphysically possible for Lewis to have had all the correct views on modality. There is more to say about such an example, of course, but it at least illustrates that failures of ***Strangeness of Impossibility*** are motivated independently of substitution inferences.[32]

With all that said, ***Counterfactual Substitution in Consequent*** is not quite the full story. If we want to explain the felt validity of arguments like the Superman Argument, which

[31]Other counterexamples can be found in Nolan 1997, p. 550 and fn. 21 and Vander Laan 2004, p. 271. It is worth noting that Vander Laan's counterexample involves a kind of *reductio* conditional that some might think should be viewed as an indicative conditional.

[32]One common response to this counterexample to ***Strangeness of Impossibility*** is that it involves some kind of scope ambiguity. On a "*de re*" reading, we hold fixed Lewis's actual views on modality and are assert *of them* that if they were right, modal realism would be true. On a "*de dicto*" reading, we are saying if the sentence "Lewis is right about modality" were true, modal realism would be true. The former reading, so the response goes, is true but a counterpossible, whereas the latter reading is not a counterpossible but is false.

It seems correct that there are multiple ways to interpret (*Lewis*), some according to which it is true and others according to which is false. However, it is worth bearing in mind that the claim that there is a scope ambiguity in (*Lewis*) is a syntactic claim. It would be more plausible if the antecedent had been phrased with 'Lewis's views', as in "If Lewis's views about modality were right,…". But the current phrasing, "If Lewis were right about modality,…" makes it difficult to maintain that there is a scope ambiguity. The difference in interpretation is more likely due to context-sensitivity than to an ambiguity.

As an analogy, suppose Aisha is taller than Bart but neither is particularly tall. Now consider:

(*Tall*) If Bart were tall, Aisha would be tall.

Like (*Lewis*), (*Tall*) seems to have two interpretations: one on which it is true (because Aisha is taller than Bart, so if Bart counts as tall, so does Aisha) and one on which it is false (because Bart's height doesn't affect Aisha's height). But the difference can hardly be attributed to a scope ambiguity with 'tall'. A more plausible explanation of the two different interpretations utilizes the fact that counterfactuals are context-sensitive: there are multiple ways to satisfy the antecedent by holding fixed different features of the actual world, and which way of satisfying the antecedent is more salient depends on features of the context. The fact that the antecedent of (*Lewis*) has both possible and impossible realizations does not undermine its status as a counterexample to ***Strangeness of Impossibility*** so much as elevate it: it is quite telling that the most natural interpretation of (*Lewis*) is one on which the antecedent describes an impossible scenario, even though an interpretation on which it describes a possible scenario is available.

involve substituting coreferring names in the antecedent of a counterfactual, we cannot appeal to ***Counterfactual Substitution in Consequent***, which only concerns substitution in the consequent of a counterfactual. And this time, we cannot appeal to an analogy with belief ascriptions or *a priori* knowability claims, since there is not really an analogue of substituting coreferring names "in the antecedent" for such environments.

Fortunately, we can still derive a more general substitution principle from ***Counterfactual Substitution in Consequent*** using another plausible counterfactual principle, which is a limited version of the replacement of counterfactually equivalent antecedents:[33]

Antecedent Replacement. $\phi(a) \boxrightarrow \psi, \phi(a) \boxrightarrow \phi(b), \phi(b) \boxrightarrow \phi(a) \vDash \phi(b) \boxrightarrow \psi$.

But notice that by applying ***Counterfactual Substitution in Consequent*** and ***Triviality***, $\phi(a) \boxrightarrow \phi(b)$ follows from:

$$\phi(a) \boxrightarrow a = b.$$

Likewise, using the same principles, $\phi(b) \boxrightarrow \phi(a)$ follows from:

$$\phi(b) \boxrightarrow a = b.$$

Combining these facts together, we get a more general substitution principle along the following lines:[34]

[33] One might try to justify ***Antecedent Replacement*** using the following more general principle:

Counterfactual Replacement. $\phi \boxrightarrow \psi, \phi \boxrightarrow \phi', \phi' \boxrightarrow \phi \vDash \phi' \boxrightarrow \psi$.

I do not want to take a stand on ***Counterfactual Replacement*** here. I will simply note that we cannot justify ***Antecedent Replacement*** with ***Counterfactual Replacement*** if we want to maintain ***SDA***. For ***Counterfactual Replacement*** together with ***SDA*** also entail ***Antecedent Strengthening***, so long as we assume the following:

Counterfactualization. If $\phi \vDash \psi$, then $\vDash \phi \boxrightarrow \psi$.

Since $\phi \vDash \phi \vee (\phi \wedge \psi)$, we have $\vDash \phi \boxrightarrow (\phi \vee (\phi \wedge \psi))$. And since $\phi \wedge \psi \vDash \phi$, we have $(\phi \wedge \psi) \boxrightarrow \phi$. So by ***Triviality*** and ***SDA***, it follows that $\vDash (\phi \vee (\phi \wedge \psi)) \boxrightarrow \phi$. Hence, by ***Counterfactual Replacement*** and ***SDA***, $\phi \boxrightarrow \theta \vDash (\phi \vee (\phi \wedge \psi)) \boxrightarrow \theta \vDash (\phi \wedge \psi) \boxrightarrow \theta$. So we cannot use ***Counterfactual Replacement*** to argue for ***Antecedent Replacement*** if we want to also maintain ***SDA***. Still, ***Antecedent Replacement*** seems *prima facie* plausible even for those who reject ***Counterfactual Replacement*** for these reasons.

[34] Note that ***Counterfactual Substitution*** requires *both* the antecedent *and* its substitution to counterfactually imply the identity claim in order to carry out the substitution. Most of the time, when $\phi(a) \boxrightarrow a = b$ holds, then so does $\phi(b) \boxrightarrow a = b$. So the inference is usually safe even with just one of these premises. Problems arise if we drop either of these premises from ***Counterfactual Substitution***, however. For example, if the following held for all a and b and all transparent $\phi(x)$ and $\psi(x)$:

$$\phi(a) \boxrightarrow \psi(a), \phi(a) \boxrightarrow a = b \vDash \phi(b) \boxrightarrow \psi(b),$$

then one instance of this would be:

$$a = b \boxrightarrow \theta, a = b \boxrightarrow a = b \vDash a = a \boxrightarrow \theta,$$

which, by ***Triviality***, reduces to:

$$a = b \boxrightarrow \theta \vDash a = a \boxrightarrow \theta$$

But this seems too strong. From the mere fact that if Alfonzo were Borka, he would go skiing, it does not seem to follow that if Alfonzo were Alfonzo, he would go skiing.

Counterfactual Substitution. $\phi(a) \boxdot\!\!\rightarrow \psi(a), \phi(a) \boxdot\!\!\rightarrow a = b, \phi(b) \boxdot\!\!\rightarrow a = b \vDash \phi(b) \boxdot\!\!\rightarrow \psi(b)$ if $\phi(x)$ and $\psi(x)$ are transparent.

In the case where no substitution occurs in the antecedent ϕ, ***Counterfactual Substitution*** reduces to ***Counterfactual Substitution in Consequent***. Again, the additional premises are very natural to assume in ordinary circumstances. In the Superman Argument, for example, it is very plausible that had Superman and I had the same parents, Superman would still have been Clark Kent, and likewise if Clark Kent and I had the same parents. Thus, the felt validity of such substitution inferences involving counterpossibles can be vindicated.

So on the one hand, appealing to ***Possible Substitution*** to meet the explanatory challenge seems to raise a number of undesirable complications. Not only does it seem explanatorily irrelevant, but it seems ill-equipped to handle the incompleteness problem from § 4 and it conflicts with ***SDA***. By contrast, ***Counterfactual Substitution*** seems to explain the felt validity of substitution inferences involving counterfactuals in a simple, unified, and elegant way. It faces no obvious incompleteness problem and it does not force the counterpossibilist to take a stand on ***SDA***. It vindicates the intuitions behind other substitution inferences such as ***Possible Substitution*** and ***A Priori Substitution*** without validating them universally. It does not require (but does not rule out) appealing to *a priori* connections when explaining the apparent validity of substitution inferences. It can be easily integrated into most hyperintensional frameworks for counterpossibles put forward in the literature. And, as far as I can see, it faces no devastating counterexamples. Therefore, I propose the counterpossibilist meet the explanatory challenge using ***Counterfactual Substitution*** and simply drop appeal to ***Possible Substitution*** altogether.

7 Epistemicizing Counterfactuals

In the previous sections, I argued that serious problems plague anyone who defends the substitution of identicals into counterfactuals with possible antecedents but rejects the substitution of identicals into counterpossibles. I argued that the more attractive alternative is to reject the substitution of identicals even into counterfactuals with possible antecedents and explain away the apparent validity by appeal to an unstated premise, viz., that the identity claim would still hold had the antecedent obtained.

To be clear, nothing I have said thus far settles whether counterpossibilism is true, and I cannot possibly hope to settle such a debate here. The goal of this paper was simply to point out the flaws in one counterpossibilist response to the problems surrounding substitution and to offer a better one. In this final section, I would like to briefly return to the debate over counterpossibles and to offer a new (albeit tentative) way of viewing that debate in light of these conclusions about the substitution of identicals.

A common worry that vacuists raise against counterpossibilism is that non-vacuous counterpossibles threaten our ordinary forms of counterfactual reasoning. If counterpossibilism is true, the thought goes, then many of the mundane forms of counterfactual reasoning that we thought were acceptable will no longer be valid. It is tempting for a counterpossibilist to try to reassure those who are compelled by this worry by arguing that counterpossibilism preserves all of the usual counterfactual inferences when the

antecedents of the counterfactuals involved are possible. This makes it seem as if counterpossibilism is just trying to account for some extreme cases of such reasoning and that our ordinary forms of counterfactual reasoning are still reliable.

But as we saw, conservatism is not an option for the counterpossibilist. What this debate over the substitution of identicals reveals is that the counterpossibilist cannot remain modest with respect to all ordinary forms of counterfactual reasoning. Their disagreement with orthodoxy goes much deeper than merely how to account for some special cases of counterfactual reasoning. Rather, the non-vacuity of counterpossibles reveals that the behavior of counterfactuals is quite analogous with the behavior of attitude ascriptions more generally in that both seem sensitive to the *way* an object is presented to an agent, not just *what* object is being presented. In other words, the analogy between counterfactuals and attitude ascriptions (such as belief reports and *a priori* knowability claims) indicates an implicit commitment by the counterpossibilist to a kind of epistemic view about counterfactuals on which all counterfactuals have an epistemic flavor.

It is well-known that counterfactuals can sometimes be read epistemically rather than circumstantially (one can replace "circumstantial" with "metaphysical" or "alethic" if desired).[35] Here is a widely-discussed example from Edgington [2008, pp. 16–17] that illustrates this point:

> There is a treasure hunt. The organizer tells me 'I'll give you a hint: it's either in the attic or the garden.' Trusting the speaker, I think 'If it's not in the attic it's in the garden.' We are competing in pairs: I go to the attic and tip off my partner to search the garden. I discover the treasure. 'Why did you tell me to go to the garden?' she asks. 'Because if it hadn't been in the attic it would have been in the garden: that's (what I inferred from) what I was told.'

In this context, the speaker's response to the question sounds perfectly correct. But notice that there is a sense in which it might be false. For suppose the person who was in charge of hiding the treasure did not even know that there was a garden. Then the treasure would not have been in the garden even if it were not in the attic.

The best way to interpret the speaker's assertion is epistemically rather than circumstantially: what the speaker is roughly trying to say is that if the treasure were not in the attic, then it would have followed from her evidence that it was in the garden. On this epistemic interpretation of the counterfactual, what the speaker asserted was correct. On the circumstantial interpretation, by contrast, what the speaker asserted might not be correct: if the treasure were not in the attic, it might have been hidden in another location that the person in charge of hiding the treasure knew about.

This suggests a new proposal about counterpossibles (proposed by Vetter [2016a]): counterpossibles are only ever non-vacuous on their *epistemic* reading. On their *circumstantial* reading, the usual semantics for counterfactuals applies, and thus, such counterfactuals are vacuous. And since charitable conversational partners will generally gravitate toward non-vacuous non-contradictory readings of a speaker's utterance, it makes sense

[35] The label "epistemic" here, which is standard terminology in the literature, might be misleading. It is not that this interpretation of counterfactuals is connected to one's knowledge *per se*. Rather, it is that this interpretation is connected to one's information state (or, more broadly, one's mental state).

that English speakers would generally gravitate towards the non-vacuous epistemic readings of these counterfactuals over their vacuous circumstantial counterparts. So even though vacuism is correct insofar as the non-epistemic readings of counterfactuals are all vacuous, the counterpossibilist is still tracking something genuine, viz., the semantic non-vacuity of the epistemic reading of counterpossibles.

While this position does not neatly fall into either the vacuist or counterpossibilist camp, it seems closer in spirit to vacuism than to counterpossibilism.[36] It would be entirely unsurprising if the epistemic readings of counterpossibles were not vacuous, since epistemic environments are generally hyperintensional. So if counterpossibilist grants that counterpossibles are vacuous on their circumstantial interpretation and only claims that counterpossibles non-vacuous on their epistemic reading, then their position is a trivial one. Moreover, the types of arguments vacuist give in favor of their position (e.g., that counterfactuals are generally not about representations) indicate that the vacuists do not have an epistemic reading of counterpossibles in mind, and though they disagree with the soundness of such arguments, counterpossibilists do not think that they are beside the point. This suggests that the original debate over counterpossibles can be restated as follows: counterpossibilists think that counterpossibles are generally non-vacuous on their circumstantial reading, whereas vacuists think that all counterpossibles are vacuous on that reading. Assuming this is a fair construal of the debate, the position that counterpossibles are non-vacuous only on their epistemic reading would be a thoroughly vacuist one. And while the vacuist could deny the existence of an epistemic reading of counterfactuals, allowing such a reading affords them the benefit of easily explaining away the counterpossibilist intuitions.

It is controversial whether there really are two separate readings of the counterfactual as opposed to just one.[37] Part of the problem is that if there are two distinct readings of the counterfactual, then there are not many ways for testing whether or not a particular counterfactual is to be read circumstantially or epistemically, since the circumstantial 'would' does not exhibit many of the normal features circumstantial modals exhibit.[38]

Vetter [2016a, pp. 17–18] has recently argued that one possible test for determining whether the intended reading of a counterfactual is epistemic or circumstantial is to see if the counterfactual obeys the substitution of identicals. If it does not, that suggests that we are dealing with an epistemic reading of the counterfactual rather than the circumstantial one. The reason is just the one we noted for thinking counterfactuals are generally transparent in § 1: circumstantial modals not generally about representational features of our language, but rather about mind-independent reality.

If I am right, however, that counterpossibilists ought to deny that counterfactuals with possible antecedents are transparency-preserving, then this test cannot be accepted by counterpossibilists unless they also accept that counterfactuals lack a circumstantial reading. For if counterfactuals have both readings, and if one maintains that substitution is licensed on the circumstantial reading, then the Main Argument is valid on that reading.

[36] Thanks to an anonymous reviewer for encouraging me to clarify this point.

[37] See, e.g., Veltman 2005, p. 174.

[38] Vetter [2016b, section 6] outlines some of the reasons why distinguishing between the circumstantial and epistemic readings of counterfactuals is difficult. For instance, like epistemic modals, and unlike circumstantial modals, counterfactuals can scope over tense and aspect.

In that case, one might start to question whether our intuitions about counterpossibles on their circumstantial readings are reliable. In addition, if substitution is licensed on the circumstantial reading of counterfactuals because the way objects are presented is irrelevant to their truth, then it is not clear why the replacement of necessary equivalents (a principle incompatible with counterpossibilism) should not also be licensed. For necessary equivalents are, in a sense, merely different ways of presenting one and the same state of affairs.

At first, this might seem to be a problem for counterpossibilism. On the one hand, this test seems to be a useful guide in general for distinguishing circumstantial and non-circumstantial modality,[39] and is supported by a quite plausible explanation for why substitution would fail in one case but not another. On the other hand, while it may be controversial whether there is an epistemic reading of counterfactuals, it is hardly controversial whether there is a circumstantial reading. This seems to put counterpossibilist in a bind: either reject this quite plausible independently well-motivated test for dinstuishing circumstantial and epistemic modals, or deny that counterfactuals have a circumstantial reading at all, neither of which seems very palatable. What I want to suggest in closing is that, actually, the conclusion that counterfactuals lack a circumstantial reading might not be so bad.

For one thing, it is far from clear that the differences between the so-called epistemic and circumstantial interpretations of counterfactuals arises out of two different *readings* of the counterfactual conditional. In fact, the hypothesis that there are two readings seems hard to square with the fact that one cannot apparently in the same breath consistently assert a counterfactual with one reading and deny the same counterfactual with a different reading. As a test, suppose you learn that the person who hid the treasure in Edgington's example knew nothing of a garden. Does what the speaker said still sound correct? To my ears, the answer is negative.

An alternative way to explain the difference between the so-called readings is through context-sensitivity. After all, it is already well-established that counterfactuals are sensitive to what possibilities are salient in context. To use a classic example from Lewis [1973, pp. 66-67], whether Caesar would have used catapults had he been in command in Korea depends on what contextually-salient facts about Caesar are being held fixed. Given this, it is unclear why one could not explain the felt differences between different "readings" of counterfactuals by appealing to context shifting.

What is more, the thesis that counterfactuals lack a circumstantial reading does not imply that counterfactuals are *about* representational features of objects. None of the counterfactuals discussed in this paper were about representations; they were all about the world and the objects in it. It is just that, according to counterpossibilism, the truth conditions of counterfactuals (like the truth conditions of attitude ascriptions) depend on the particular representation used to represent objects. In other words, while the condition expressed by a counterfactual need not be a condition on representational devices such as

[39]Here, I am excluding epistemic modals such as 'might' and first-person attitude reports, where substitution inferences seem uniformly valid. For example, "Superman might be powerful; Superman is Clark Kent; therefore, Clark Kent might be powerful" seems valid. Likewise, "I think Superman is powerful; Superman is Clark Kent; therefore, I think Clark Kent is powerful" may turn valid under certain conceptions of logical consequence.

names, the condition that is expressed by a counterfactual depends on the representational devices used in stating the counterfactual.

So while the counterpossibilist could maintain the two-readings view of counterfactuals and argue that counterpossibles are non-vacuous even on the circumstantial reading, I think this would be ill-advised. The argument that circumstantial modals validate the substitution of identicals is fairly compelling. One might even take it as *definitional* of circumstantial modality that it not be representation-sensitive. If counterfactuals are representation-sensitive even on their circumstantial reading, one has to ask where this sensitivity comes from. By contrast, the counterpossibilist who maintains that counterfactuals only have an epistemic reading do not face such difficult questions. The representation-sensitivity of counterfactuals arises in a natural way that is analogous to how it arises in attitude ascriptions.

Some might nevertheless hesitate to accept a view on which counterfactuals lack a circumstantial reading. If you find yourself in this camp, keep in mind that this claim follows from the claim that circumstantial modals validate the substitution of identicals and the claim that counterfactuals generally invalidate the substitution of identicals. The fact that counterfactuals invalidate the substitution of identicals according to counterpossibilism shows that counterfactuals have to behave more like epistemic operators than circumstantial ones in that they must be sensitive to modes of presentation and not just to reference. Indeed, there is some evidence that counterpossibilists themselves already conceive of counterfactuals in this way.[40] Thus, properly understood, I think it is reasonable to think of the debate over counterpossibles as being closely tied to questions concerning the extent to which counterfactuals are more like attitude ascriptions and epistemic operators than previously recognized.

A Appendix

In this appendix, we verify some of the claims that were made in §§ 3–4 regarding entailment relations between various principles of counterfactual reasoning.

First, in § 3, it is claimed that ***Strangeness of Impossibility*** entails ***Possible Substitution*** given certain plausible counterfactual principles. The plausible counterfactual principles we need for the proof are listed below:

Agglomeration. $\phi \boxrightarrow \psi, \phi \boxrightarrow \theta \vDash \phi \boxrightarrow (\psi \wedge \theta)$.

Possible Closure. If $\psi \vDash \theta$, then $\phi \boxrightarrow \psi, \Diamond \phi \vDash \phi \boxrightarrow \theta$.

[40] For instance, Mares [1997] develops a counterpossibilist semantics where truth is assessed relative to information states rather than to worlds. Brogaard and Salerno [2013] develop a counterpossibilist semantics for counterfactuals invoking the notion of an "*a priori* connection" between propositions. Vander Laan [2004, pp. 269–271] and Brogaard and Salerno [2013, p. 648] emphasize that while counterpossibles can be informative, counterfactuals with "unentertainable" suppositions cannot. Thus, counterfactuals are sensitive to the epistemic or doxastic notion of entertainability as opposed to the metaphysical notion of possibility. On the other hand, those that generally adopt ***Strangeness of Impossibility*** do not like to characterize their view as "epistemicizing" counterfactuals. Thus, Berto et al. [2017, p. 11] explicitly deny that counterfactuals are epistemic and also commit themselves to ***Strangeness of Impossibility*** (p. 6). See also Kment 2014.

Fact A.1. ***Strangeness of Impossibility***, ***Agglomeration***, and ***Possible Closure*** entail:

Strong Possible Closure. $\phi \boxrightarrow \psi, \Diamond \phi, \Box(\psi \supset \theta) \models \phi \boxrightarrow \theta$.

Proof: First, by ***Strangeness of Impossibility***:

$$\Diamond \phi, \Box(\psi \supset \theta) \models \phi \boxrightarrow (\psi \supset \theta)$$

So by ***Agglomeration***:

$$\phi \boxrightarrow \psi, \Diamond \phi, \Box(\psi \supset \theta) \models \phi \boxrightarrow (\psi \wedge \psi \supset \theta)$$

Applying ***Possible Closure*** to the conclusion, we obtain ***Strong Possible Closure***. ■

Strong Possible Closure also entails ***Possible Closure*** given the principles from § 2.

Fact A.2. ***Strangeness of Impossibility*** and ***Strong Possible Closure*** entail ***Possible Substitution***.

Proof: If $\psi(x)$ is transparent, then by ***Necessity of Identity*** and ***Necessitation***, we have:

$$a = b \models \Box(\psi(a) \supset \psi(b)).$$

With an application of ***Strong Possible Closure***, we obtain ***Possible Substitution***. ■

Next, it is claimed in § 4 that ***Minimize Impossibility*** entails ***A Priori Substitution*** given certain plausible counterfactual principles. The principles we need are ***Agglomeration*** from above, ***Antecedent Replacement*** from § 6, and the following principles:

A Priorization. If $\phi_1, \ldots, \phi_n \models \psi$, then $\mathsf{APK}\,\phi_1, \ldots, \mathsf{APK}\,\phi_n \models \mathsf{APK}\,\psi$.

Conceivable Closure. If $\psi \models \theta$, then $\phi \boxrightarrow \psi, \mathsf{APC}\,\phi \models \phi \boxrightarrow \theta$.

Fact A.3. ***Minimize Impossibility***, ***Agglomeration***, ***A Priorization***, and ***Conceivable Closure*** entail:

A Priori Substitution in Consequent. $\phi \boxrightarrow \psi(a), a = b, \neg\,\mathsf{APK}(\phi \supset a \neq b) \models \phi \boxrightarrow \psi(b)$ if $\psi(x)$ is transparent.

Proof: Since $\psi(x)$ is transparent, we have:

$$a = b \models \Box(\psi(a) \supset \psi(b)).$$

Moreover, by ***A Priorization***, we have:

$$\mathsf{APK}(\phi \supset \neg(\psi(a) \supset \psi(b))) \vDash \mathsf{APK}(\phi \supset a \neq b).$$

By applying ***Minimize Impossibility***, we obtain:

$$a = b, \neg\, \mathsf{APK}(\phi \supset a \neq b) \vDash \phi \boxrightarrow (\psi(a) \supset \psi(b)).$$

Thus, using ***Agglomeration***, it follows that:

$$\phi \boxrightarrow \psi(a), a = b, \neg\, \mathsf{APK}(\phi \supset a \neq b) \vDash \phi \boxrightarrow (\psi(a) \wedge (\psi(a) \supset \psi(b))).$$

Moreover, by ***A Priorization***:

$$\neg\, \mathsf{APK}(\phi \supset a \neq b) \vDash \mathsf{APC}\, \phi.$$

So by ***Conceivable Closure***, we obtain ***A Priori Substitution in Consequent***. ■

Fact A.4. ***A Priori Substitution in Consequent*** and ***Antecedent Replacement*** entail ***A Priori Substitution***.

Proof: By ***A Priori Substitution in Consequent*** and ***Triviality***, we have:

$$a = b, \neg\, \mathsf{APK}(\phi(a) \supset a \neq b) \vDash \phi(a) \boxrightarrow \phi(b).$$

Likewise, we have:

$$a = b, \neg\, \mathsf{APK}(\phi(b) \supset a \neq b) \vDash \phi(b) \boxrightarrow \phi(a).$$

Putting these together:

$$\begin{aligned} &a = b, \neg\, \mathsf{APK}(\phi(a) \supset a \neq b), \neg\, \mathsf{APK}(\phi(b) \supset a \neq b) \\ \vDash\ &(\phi(a) \boxrightarrow \phi(b)) \wedge (\phi(b) \boxrightarrow \phi(a)). \end{aligned}$$

Using an application of ***Antecedent Replacement*** and ***A Priori Substitution in Consequent***, we obtain ***A Priori Substitution***. ■

References

Alonso-Ovalle, Luis. 2006. *Disjunction in Alternative Semantics*. Ph.D. thesis, University of Massachusetts Amherst.

—. 2008. "Alternatives in the Disjunctive Antecedents Problem." In Charles B Chang and Hannah J Haynie (eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 42–50. Somerville, MA.

Bennett, Jonathan Francis. 2003. *A Philosophical Guide to Conditionals*. Oxford University Press.

Bernstein, Sara. 2016. "Omission impossible." *Philosophical Studies* 173:2575–2589.

Berto, Francesco, French, Rohan, Priest, Graham, and Ripley, David. 2017. "Williamson on Counterpossibles." *Journal of Philosophical Logic* 17:1–21.

Bjerring, Jens Christian. 2013. "On Counterpossibles." *Philosophical Studies* 168:327–353.

Briggs, Rachael. 2012. "Interventionist counterfactuals." *Philosophical Studies* 160:139–166.

Brogaard, Berit and Salerno, Joe. 2013. "Remarks on Counterpossibles." *Synthese* 190:639–660.

Cohen, Daniel H. 1987. "The Problem of Counterpossibles." *Notre Dame Journal of Formal Logic* 29:91–101.

—. 1990. "On What Cannot Be." In Jon Michael Dunn and Anil Gupta (eds.), *Truth or Consequences: Essays in Honor of Nuel Belnap*, 123–132. Dordrecht: Springer Netherlands.

Edgington, Dorothy. 2008. "Counterfactuals." *Proceedings of the Aristotelian Society* 108:1–21.

Ellis, Brian, Jackson, Frank, and Pargetter, Robert. 1977. "An Objection to Possible-World Semantics for Counterfactual Logics." *Journal of Philosophical Logic* 6:355–357.

Fine, Kit. 1975. "Critical Notice of David Lewis's *Counterfactuals*." *Mind* LXXXIV:451–458.

—. 2012. "A Difficulty for the Possible Worlds Analysis of Counterfactuals." *Synthese* 189:29–57.

Frege, Gottlob. 1892. "Über Sinn und Bedeutung." *Zeitschrift für Philosophie und philosophische Kritik* 25–50. Translated in 1948 as: "Sense and Reference." *The Philosophical Review* 57: 209–230.

Galles, David and Pearl, Judea. 1998. "An Axiomatic Characterization of Causal Counterfactuals." *Foundations of Science* 3:151–182.

Gillies, Anthony S. 2007. "Counterfactual Scorekeeping." *Linguistics and Philosophy* 30:329–360.

Goodman, Jeffrey. 2004. "An Extended Lewis/Stalnaker Semantics and the New Problem of Counterpossibles." *Philosophical Papers* 33:35–66.

Goodman, Nelson. 1947. "The Problem of Counterfactual Conditionals." *The Journal of Philosophy* 44:113–128.

Kallestrup, Jesper. 2009. "Conceivability, Rigidity and Counterpossibles." *Synthese* 171:377–386.

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

Kocurek, Alexander W. 2018. "Counteridenticals." *The Philosophical Review* 127:323–369.

Kolodny, Niko and MacFarlane, John. 2010. "Ifs and Oughts." *Journal of Philosophy* 107:115–143.

Krakauer, Barak. 2012. *Counterpossibles*. Ph.D. thesis, University of Massachusetts, Amherst.

Kratzer, Angelika. 1981. "Partition and Revision: The Semantics of Counterfactuals." *Journal of Philosophical Logic* 10:201–216.

—. 1989. "An Investigation of the Lumps of Thought." *Linguistics and Philosophy* 12:607–653.

—. 2012. *Modals and Conditionals*. Oxford: Oxford University Press.

Lewis, David K. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

—. 1977. "Possible-world Semantics for Counterfactual Logics: A Rejoinder." *Journal of Philosophical Logic* 6:359–363.

Lewis, Karen S. 2017. "Counterfactual Discourse in Context." *Noûs* 30:329–27.

Loewer, Barry M. 1976. "Counterfactuals with Disjunctive Antecedents." *The Journal of Philosophy* 73:531–537.

Lycan, William G. 2001. *Real Conditionals*. Oxford: Clarendon Press.

Mares, Edwin D. 1997. "Who's Afraid of Impossible Worlds?" *Notre Dame Journal of Formal Logic* 38:516–526.

Mckay, Thomas and van Inwagen, Peter. 1977. "Counterfactuals with Disjunctive Antecedents." *Philosophical Studies* 31:353–356.

Moss, Sarah. 2012. "Updating as Communication." *Philosophy and Phenomenological Research* LXXXV:225–248.

Nolan, Daniel. 1997. "Impossible Worlds: A Modest Approach." *Notre Dame Journal of Formal Logic* 38:535–572.

Nute, Donald. 1975. "Counterfactuals and the Similarity of Words." *The Journal of Philosophy* 72:773–778.

—. 1978. "Simplification and Substitution of Counterfactual Antecedents." *Philosophia* 7:317–325.

Popper, Karl. 1959. "On Subjunctive Conditionals With Impossible Antecedents." *Mind* LXVIII:518–520.

Santorio, Paolo. Forthcoming. "Alternatives and Truthmakers in Conditional Semantics." *Journal of Philosophy* .

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In *IFS*, 41–55. Dordrecht: Springer Netherlands.

Starr, William B. 2014. "A Uniform Theory of Conditionals." *Journal of Philosophical Logic* 43:1019–1064.

Vander Laan, David A. 2004. "Counterpossibles and Similarity." In Frank Jackson and Graham Priest (eds.), *Lewisian Themes*, 258–275. Oxford: Clarendon Press.

Veltman, Frank. 2005. "Making Counterfactual Assumptions." *Journal of Semantics* 22:159–180.

Vetter, Barbara. 2016a. "Counterpossibles (Not Only) for Dispositionalists." *Philosophical Studies* 173:2681–2700.

—. 2016b. "Williamsonian Modal Epistemology, Possibility-Based." *Canadian Journal of Philosophy* 46:766–795.

von Fintel, Kai. 2001. "Counterfactuals in a Dynamic Context." In Michael Keystowicz (ed.), *Ken Hale: A Life in Language*, 123–152. Cambridge, MA: MIT Press.

Warmbrōd, Ken. 1981. "Counterfactuals and Substitution of Equivalent Antecedents." *Journal of Philosophical Logic* 10:267–289.

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

—. 2017. "Counterpossibles in Semantics and Metaphysics." *Argumenta* 2:195–226.