Counteridenticals*

Alex Kocurek

Forthcoming in *The Philosophical Review*.

Abstract. A counteridentical is a counterfactual with an identity statement in the antecedent. While counteridenticals generally seem non-trivial, most semantic theories for counterfactuals, when combined with the necessity of identity and distinctness, attribute vacuous truth conditions to such counterfactuals. In light of this, one could try to save the orthodox theories either by appealing to pragmatics or by denying that the antecedents of alleged counteridenticals really contain identity claims. Or one could reject the orthodox theory of counterfactuals in favor of a hyperintensional semantics that accommodates non-trivial counterpossibles. In this paper, I argue that none of these approaches can account for all the peculiar features of counteridenticals. Instead, I propose a modified version of Lewis's counterpart theory, which rejects the necessity of identity, and show that it can explain all the peculiar features of counteridenticals in a satisfactory way. I conclude by defending the plausibility of contingent identity from objections.

§1 Introduction

A ***counteridentical*** is a counterfactual with an identity statement embedded in the antecedent. For example:

(*Bet*) If I were you, I would bet on that horse.

(*Einstein*) If Alyssa were Einstein, she would have aced the test.

(*Frege*) If Hesperus were not Phosphorus, Frege would need a different example.

(*Superman*) If Superman were not Superman, we would all be dead by now.

(*Give Up*) If you were anyone else, you would have given up.

We use counterfactuals such as these in ordinary conversation all the time. When we do use them, we often end up communicating something meaningful and informative. And yet, the standard views on counterfactuals and identity predict that most ordinary counteridenticals, such as the ones above, are trivial and uninformative.

*For extensive feedback on this paper, I would like to thank Shamik Dasgupta, Tyrus Fisher, Melissa Fusco, Wes Holliday, Carina Kauf, Annina Loets, John MacFarlane, Seth Yalcin, and two anonymous reviewers. For comments on early drafts of this paper, I would like to thank Adam Bradley, Peter Hawke, Ethan Jerzak, Alex Kerr, Geoff Lee, Antonia Peacocke, Rachel Rudolph, and Alberto Tassoni. I would also like to thank Richard Lawrence, Line Mikkelsen, Daniel Nolan, Mike Rieppel, and Sarah Zobel for their helpful comments and suggestions regarding this topic. Special thanks to Michael Deigan for pointing me to the passage of Lewis 1973, which inspired me to work on this topic. I am also grateful for the comments and discussion from the conferences where versions of this paper were presented, including the Berkeley-Stanford-Davis philosophy conference held at the University of California, Berkeley in 2015, the Conditionals at the Crossroads of Semantics and Pragmatics conference held at the University of Konstanz in 2016, and an invited talk for the LLEMMMa group at University of California, Davis in 2017.

To illustrate, consider the semantics for counterfactuals defended by Stalnaker [1968] and Lewis [1973], according to which a counterfactual is true just in case its consequent holds at all the closest possible worlds satisfying the antecedent. On the Lewis-Stalnaker semantics, counterfactuals with impossible antecedents, or ***counterpossibles***, are vacuously true (irrespective of how we spell out the notion of closeness) since there are no possible worlds satisfying the antecedent of such counterfactuals. But it is generally thought, following Kripke [1971, 1980], that facts about identity are necessary. That is, the following two principles are widely accepted:[1]

NecId. $a = b \vDash \Box(a = b)$ and $a \neq b \vDash \Box(a \neq b)$ for any names a and b.

NecId Quantified. $\vDash \forall x \forall y \ (x = y \supset \Box(x = y))$ and $\vDash \forall x \forall y \ (x \neq y \supset \Box(x \neq y))$.

If we add these principles to the Lewis-Stalnaker semantics, counteridenticals such as the ones above will be rendered vacuously true. This result generalizes to a large number of other semantic theories of counterfactuals discussed in the literature.[2]

There are several ways one might address this puzzle. One could try to save the orthodox theories, for example, by providing an alternative explanation for the felt non-vacuity of counteridenticals. Perhaps one could explain the felt non-vacuity of counteridenticals via pragmatics (§ 3). Or perhaps one could deny that sentences like (*Bet*), (*Einstein*), and so on really do contain identity claims (§ 4). Alternatively, one could replace the orthodox semantics counterfactuals with a hyperintensional semantics that can tolerate non-trivial counterpossibles (§ 5). Finally, one could abandon the Kripkean view on identity and hold that identity facts are contingent (§ 6).

In this paper, I argue that the most promising theory that explains the felt non-vacuity of counteridenticals is one that rejects the Kripkean view of identity. Here is how I do it. First, I make some observations about counteridenticals in § 2 and argue that none of the more conservative theories of counteridenticals can adequately explain all of this data in § 3–5. Then I develop a theory of counteridenticals in § 6 using counterpart theory (in the spirit of Lewis 1973, p. 43) that can adequately explain this data. Thus, I defend a Lewisian brand of contingent identity. After presenting what I take to be Lewis's theory of counteridenticals, I present a counterexample to this theory in § 7 involving the substitution of identicals and propose a version of counterpart theory that avoids this counterexample. Finally, I conclude in § 8 by attempting to defuse an objection against contingent identity and clarify where contingent identity theorists and their opponents disagree.

§2 Five Exhibits

Counteridenticals are puzzling in part because they do not seem to sit well with the standard views on counterfactuals and identity. But that is not all. Counteridenticals also display a number of peculiar linguistic behaviors that demand an explanation. So before

[1] While one could consistently hold one of these principles without the other, it is reasonable to suppose that they will stand or fall together. So I will set aside views which accept one but reject the other.

[2] The semantic proposals of Kratzer [1981, 1989, 2012]; Galles and Pearl [1998], and Lycan [2001] all render counterpossibles semantically vacuous. For von Fintel [2001]; Gillies [2007]; Starr [2014], and Willer [2015], counterpossibles violate a possibility presupposition.

we examine the different theories of how to interpret counteridenticals, I will first state some observations about counteridenticals that any adequate theory of their interpretation ought to account for.

Exhibit A: Explicit Identity. Replacing the 'were' in the antecedent of a counteridentical with the more explicit 'were identical to' generally does not seem to result in an equivalent sentence. Consider the following two sentences, for example:

(*Button*) If I were you, I would not push that button.

(*Button*=) If I were identical to you, I would not push that button.

It feels as though (*Button*) and (*Button*=) are not exactly saying the same thing. For some, (*Button*=) sounds marked, though it is unclear what to make of this, since (a) the intuition is not universally held, and (b) the oddity of (*Button*=) might stem from the fact that we do not often go around asserting counterfactuals with explicit identity statements in them. Still, it is quite plausible that at the very least (*Button*) and (*Button*=) are not equivalent—despite the fact that 'is' and 'is identical to' seem synonymous.

Exhibit B: Equivalence Properties. Identity is an equivalence relation. That means it is reflexive, symmetric, and transitive. Yet the relation in the antecedents of counteridenticals does not seem to be an equivalence relation. For one thing, the order of the terms seems to matter.[3] For example, the following do not seem equivalent:

(*Bet*) If I were you, I would bet on that horse.

(*Bet Rev*) If you were me, I would bet on that horse.

Moreover, the following sentences do not seem to be necessarily true (indeed, they generally seem false):

(*Trans*) If I were you and you were Pope Francis, I would be Pope Francis.

(*Sym*) If I were you, you would be me.

Both of these observations suggest that the relation in the antecedent of these counterfactuals is not an equivalence relation—even though identity most certainly is.

Exhibit C: Reflexive Pronouns. Consider an example due to Lakoff [1970, p. 248]:[4]

(*Hate*) If I were you and you were me, I would hate you.

It does not seem like (*Hate*) is equivalent to:

(*Self-Hate*) If I were you and you were me, I would hate myself.

[3] See Reboul 1996, p. 172.

[4] See Reboul 1996 and Arregui 2007 for discussion of similar examples.

Rather, it seems that 'I' and 'you' denote different people in the consequent of (*Hate*), despite the fact that they are apparently identified in the antecedent. Even more striking is the fact that the following sentences do not seem equivalent:

(*Vote Me*) If I were you, I would vote for me.

(*Vote Myself*) If I were you, I would vote for myself.

To illustrate, if Alice, who is running for mayor, asserted (*Vote Me*) to Beth, it would sound like Alice was saying "You should vote for me." But if she instead asserted (*Vote Myself*), it would sound like she was saying "You should vote for yourself." Yet why should these have different interpretations if 'me' and 'myself' necessarily codenote?

Exhibit D: Indicatives. Our next observation comes from the indicative versions of counteridenticals, which I will call ***identity indicatives*** for lack of a better term. Consider the indicativized versions of (*Bet*), (*Einstein*), and so on:

(*Bet Ind*) #If I am you, I will bet on that horse.

(*Einstein Ind*) #If Alyssa is Einstein, she aced the test.

(*Frege Ind*) #If Hesperus is not Phosphorus, Frege needed a different example.

(*Superman Ind*) #If Superman is not Superman, we are all dead by now.

(*Give Up Ind*) #If you are anyone else, you gave up.

In general, unless one is in a special context (which we will discuss in a moment), these indicatives sound marked. Note that it does not matter whether any of these examples are past-directed, present-directed, or future-directed: they all sound equally bad.[5]

To be clear, identity indicatives do not *always* sound marked, even if we know the antecedents are false. Here are some examples (with context in the brackets):

(*Plato*) [*A professor of ancient philosophy is talking about Zeno's arguments against the plurality of things in the Parmenides.*]

If you are Plato, you will not be convinced by Zeno's argument against plurality.

(*Venus*) [*An astronomer is trying to reason with a crackpot who claims that the first astronomical object in the sky in the evening is not Venus.*]

If Hesperus is not Venus, then every single telescope used by astronomers must be broken (except yours, of course).

(*PA*) [*A logician is trying to explain to their class that* PA (*Peano Arithmetic) is capable of proving every true* Σ^0_1*-sentence in the language of arithmetic.*]

If $2 + 2$ is 5, then PA can prove the sentence $\ulcorner 2 + 2 = 5 \urcorner$.

[5] Future-directed conditionals might be an exception. (*Bet Ind*) seems more acceptable in contexts where one is trying to predict or explain another's behavior (cf. (*Plato*) below). I suspect this is because future-directed indicatives have a counterfactual reading, so (*Bet Ind*) can be interpreted as (*Bet*) in certain contexts. By contrast, it sounds much worse to say "If I am you, I (already) bet on that horse."

(*Future*) [*Someone from the future builds a time machine and travels back in time to meet their younger self. Their younger self says:*]

If I am you, then I must be a genius!

Be that as it may, this does not undermine the need to explain the general infelicity of identity indicatives. (*Plato*), (*Venus*), (*PA*), and (*Future*) all seem to be very peculiar in their own unique way and each seems to require its own special treatment.[6] Even if one grants that identity indicatives are not always infelicitous, it is still quite striking that they sound infelicitous in a large range of cases.

Exhibit E: Propositional Attitude Reports. Finally, the observations above all apply straightforwardly to propositional attitude reports, dream reports, and other similar environments. Consider an example due to Cumming [2008, p. 529].[7] Biron (a perfect logician, we may assume) is attending a masked ball with Katherine, Rosaline, and Maria. While Biron is away, Katherine, Rosaline, and Maria switch masks. Now, Katherine has Rosaline's mask, Rosaline has Maria's, and Maria has Katherine's. Biron returns to the group, unaware of the switch.

In this case, the following seem true:

(*Kat is Rosa*) Biron believes Katherine is Rosaline.

(*Kat ≠ Rosa*) Biron does not believe that Katherine is identical to Rosaline.

This illustrates that 'is' and 'is identical to' are not equivalent in belief ascriptions. Moreover, the following also seem true:

(*Rosa is Maria*) Biron believes Rosaline is Maria.

(*Maria is Kat*) Biron believes Maria is Katherine.

But if we naïvely applied transitivity to (*Kat is Rosa*) and (*Rosa is Maria*), we would get the false conclusion that:

(*Kat is Maria*) Biron believes Katherine is Maria.

Likewise, if we applied symmetry to (*Kat is Rosa*), we would incorrectly predict that:

(*Rosa is Kat*) Biron believes Rosaline is Katherine.

So the relation in these belief ascriptions does not seem to be an equivalence relation. Finally, Lakoff's famous Brigitte Bardot examples illustrate the difference between 'me' and 'myself' in dream reports:

(*Kiss Me*) I dreamt I was Brigitte Bardot and I kissed me.

[6] For instance, (*Plato*) seems to be an idiosyncrasy of English whose felicity is not universally cross-linguistic. I am told that in other languages (such as Danish, German, and Hebrew), this conditional sounds bad even in the stated context. Many thanks to Klaus Corcilius, Ravit Dotan, and Line Mikkelsen for pointing this out to me.

[7] Stalnaker [2008, p. 147] has a similar example.

(*Kiss Myself*) #I dreamt I was Brigitte Bardot and I kissed myself.

Nothing is special about believing and dreaming: similar data arises for most non-factive propositional attitude reports.[8] Thus, it would be quite surprising if the explanation for these observations surrounding counteridenticals did not apply to propositional attitude reports. A unified theory is to be expected.

§3 Pragmatics

It is very tempting to think that counteridenticals are, in some sense, just "loose talk". When someone says (*Bet*), for instance, often what they mean is something like "You should bet on that horse" or "If I were placing a bet, I would bet on that horse". One might think that while counteridenticals are literally vacuous, we often communicate something other than their literal content when we assert them. This suggests that interpreting counteridenticals requires a pragmatic approach, rather than a semantic one. According to ***the pragmatic theory***, we can explain the felt non-vacuity of counteridenticals just using pragmatics, without any revision to our semantics or metaphysics.

How would a pragmatic explanation for the felt non-vacuity of counteridenticals go? One approach, ***the Gricean approach***, would be to appeal to Gricean principles. Perhaps when one hears someone assert a counteridentical like (*Bet*), they quickly reject the literal interpretation of that phrase in order to be charitable to the speaker, since asserting (*Bet*) flouts Grice's maxim of quantity. They then search for an alternative interpretation of what was meant by appealing to salient facts about the individuals involved. Compare this to other cases involving vacuous conditionals that sound non-vacuous such as:

(*Lose*) If you lose, you lose.

In asserting such a conditional, one would thereby flout Grice's maxim of quantity. Thus, one would naturally interpret a speaker who asserted (*Lose*) as saying something other than what they literally asserted (such as "Stop worrying so much!").

Unfortunately, this Gricean explanation of the felt non-vacuity of counteridenticals will not work. Felt strengthenings that arise from such pragmatic effects typically disappear in downward-entailing environments such as negations or doubt ascriptions.[9] So while (*Lose*) might not sound vacuous in conversation, each of the following sound marked:

(*Not Lose*) #It is not the case that if you lose, you lose.

(*Doubt Lose*) #I doubt that if you lose, you lose.

By contrast, counteridenticals sound fine even in these environments:

(*Not Bet*) It is not the case that if I were you, I would bet on that horse.

(*Doubt Bet*) I doubt that if I were you, I would bet on that horse.

[8] Factive propositional attitude reports behave like identity indicatives. For example, "Biron knows Katherine is Rosaline" sounds marked in the masked ball example.

[9] See, for instance, Gazdar 1979; Chierchia 1999, and Fox 2007.

This suggests that the felt non-vacuity of counteridenticals is not explained by standard Gricean mechanisms.

Another option would be to treat counteridenticals as metaphors. Metaphors have a number of the features that counteridenticals have. For example, while (*Juliet*) is readily heard metaphorically, it is very difficult to hear (*Juliet*=) metaphorically even when one is in a poetic mood:

(*Juliet*) Juliet is the sun.

(*Juliet*=) #Juliet is identical to the sun.

Likewise, order matters in metaphor. While the sentence below can be interpreted metaphorically, its metaphorical interpretation is very different from that of (*Juliet*):

(*Juliet Rev*) The sun is Juliet.

Thus, one might plausibly think that the antecedents of counteridenticals are usually interpreted metaphorically, rather than literally.[10]

However, even if the metaphor approach is appropriate for interpreting counteridenticals, it hardly seems to be the right kind of theory for interpreting belief ascriptions involving identity. Consider again the masked ball example from § 2 (Exhibit E). According to the metaphor approach, Biron only believes metaphorically that Katherine is Rosaline, just as Romeo only believes metaphorically that Juliet is the sun. But this metaphorical interpretation of what Biron believes does not seem plausible. It is not that Biron believes Katherine is *like* Rosaline in certain respects; he might believe that even before the mask switching takes place (and *vice versa*). Rather, Biron is in a real case of identity confusion. He is *literally* confusing Katherine with Rosaline. So if we want a unified theory of attitude ascriptions and counteridenticals, the metaphor approach will not do.

A final approach would be to bootstrap a pragmatic theory of counteridenticals from a more general pragmatic theory of counterpossibles. One pragmatic theory of counterpossibles due to Emery and Hill [2016] appeals to Gricean principles, such as the maxim of quantity. Another proposal due to Vetter [2016] is to interpret non-vacuous counterpossibles epistemically, rather than metaphysically or circumstantially. Finally, Williamson [2016] argues that the felt non-vacuity of counterpossibles is due to an overreliance on a heuristic that says to treat counterfactuals with the same antecedent but inconsistent consequents as inconsistent. This heuristic is reliable for ordinary counterfactuals with possible antecedents, but it leads us astray when the antecedent is impossible. So perhaps one of these theories could explain why counteridenticals do not feel vacuous.

Alas, none of these pragmatic accounts can explain all of the observations surrounding counteridenticals from § 2 either. Take Exhibit A (explicit identity). What would explain why (*Button*) and (*Button*=) do not sound equivalent?

(*Button*) If I were you, I would not push that button.

(*Button*=) If I were identical to you, I would not push that button.

[10] See Lakoff 1996. For a general discussion about metaphor, see Glanzberg 2007.

Gricean principles do not seem to help. If interpreters sought a non-literal interpretation of (*Button*) according to Gricean principles, then there is no reason why they would not also seek the same non-literal interpretation of (*Button=*) according to the same principles. Appealing to an epistemic interpretation of the counterfactual does not help either. For even if (*Button*) and (*Button=*) had an epistemic interpretation (which already seems unlikely), there does not seem to be any reason for giving (*Button*) a different epistemic interpretation from (*Button=*)—what would the other epistemic interpretation be?[11] And finally, while the heuristic Williamson described might provide an insightful error theory for counterpossibles, it does not explain how and why we interpret 'were' and 'were identical to' differently.

It is also worth flagging that none of the three pragmatic approaches above have anything to say about Exhibit C (reflexive pronouns). Gricean maxims, metaphors, and the pragmatics of counterpossibles seem to say nothing about why, for instance, the 'me' and the 'myself' receive apparently different interpretations in the following sentences:

(*Vote Me*) If I were you, I would vote for me.

(*Vote Myself*) If I were you, I would vote for myself.

All of these reasons for being skeptical of the prospects of the pragmatic theory are defeasible. Perhaps there is a systematic pragmatic explanation for Exhibits A–E. Or perhaps there just *is no* systematic explanation for these observations. But I do not want to dwell on the matter here. My inclination is to take these criticisms as preliminary reasons to explore semantic accounts of counteridenticals to see if they can clear things up first.

§4 Predication

Exhibits A–C suggest that sentences such as (*Bet*), (*Einstein*), and so on are not really counterfactuals with *identity* claims in the antecedent. So as not to beg the question, let us call counterfactuals such as (*Bet*), (*Einstein*), and so on that seem to have an identity statement in the antecedent ***alleged counteridenticals***. Could it be that alleged counteridenticals are not *genuine* counteridenticals? And if not, what are they?

It is very tempting to reach for something like the idea that 'if I were you' just means the same thing as 'if I were in your position'. For instance, one has the feeling that when one asserts (*Bet*), one is trying to say something like the following:

(*Bet Position*) If I were in your position, I would bet on that horse.

Thus, one might think we can analyze counteridenticals as disguised ***counterpositionals***, i.e., counterfactuals whose antecedents involve talk of positions.[12]

[11] One idea is that 'identical' in (*Button=*) tends to be interpreted as 'qualitatively identical'. Thus, 'if I were identical to you' means something like 'if I were qualitatively identical to you', whereas 'if I were you' does not. (Thanks to an editor for raising this possibility.) While this might be one interpretation of (*Button=*), I do not think this is the most natural interpretation of (*Button=*). Rather, (*Button=*) seems to be equivalent to something like the following:

(*Button Same*) If you and I were the same person, I would not push that button.

This suggests that 'identical' in (*Button=*) is being interpreted as 'numerically identical'.

[12] Reboul [1996, p. 162] seems to defend a view like this.

However, the suggestion that 'if I were you' is synonymous with 'if I were in your position' as it stands is too simplistic. For one thing, if it were correct, we should expect counteridenticals to be equivalent to their corresponding counterpositionals. But while this equivalence often holds, it does not always hold as the examples below demonstrate:

(*Earth*) If Earth were Jupiter, it would be hundreds of times bigger.

(*Earth Pos*) ?If Earth were in Jupiter's position, it would be hundreds of times bigger.

(*Nixon*) If Nixon were Meir, he would be a woman.[13]

(*Nixon Pos*) ?If Nixon were in Meir's position, he would be a woman.

($Id > \neg Pos$) If I were you, I would not be in your position.

($Pos > \neg Pos$) #If I were in your position, I would not be in your position.

More importantly, even if the suggestion were correct, one still would need to answer the question of whether the antecedents in alleged counteridenticals contain identity claims. If the answer is yes, then the apparent equivalence between (*Bet*) and (*Bet Position*) is a datum that needs to be explained by a theory of counteridenticals: why does embedding an identity claim in the antecedent have the same effect as embedding talk of positions? If the answer is no, then one is left with the following question: what *is* the logical form of alleged counteridenticals?

Here is another way of stating this question. Syntactically, 'were' is a copula. The copulas of English can be split into roughly three categories: equative, predicational, and specificational.[14] The first two correspond to the familiar distinction between the 'is' of identity and the 'is' of predication. Equative copulas are the ones that occur in identity claims, such as the 'is' in "Cicero is Tully" and "Hesperus is Phosphorus". Predicational copulas are used in attributing properties and relations to objects, such as the 'is' in "Liya is tall" and "Ellen is dating José." Specificational copulas are copulas that are used to specify who fulfills a certain description. For instance, the 'is' in "The 44th president of the United States is Barack Obama" is specificational. With this taxonomy for copulas, we can phrase our question more precisely as follows: what kind of copula is the 'were' in alleged counteridenticals? Is it equative, predicational, or specificational?

The proposal examined in this section is that the 'were' is predicational. The idea is that the term in the postcopular clause is, despite appearances, standing for a property rather than an object—in particular, it is standing for a contextually salient property that the referent of the term actually has. So in 'if I were you', the 'were you' is really picking out a contextually salient property of the listener of the context and ascribing that property to the speaker. The true logical forms of alleged counteridenticals involve predications, not identifications. Call this ***the predicate theory*** of counteridenticals.

There is already independent motivation for the claim that apparently referential terms such as names can sometimes act as predicates.[15] Consider the following sentences:

[13]This example is due to Pollock [1976, p. 6].

[14]This taxonomy comes from Mikkelsen 2005, p. 1806, which cites Higgins 1979, pp. 204–293. This taxonomy is not entirely uncontroversial; see Mikkelsen 2005, pp. 1813–1814 for further discussion.

[15]See, for instance, Burge 1973; Rieppel 2013, and Fara 2015 for discussion.

(*Alfred*) Every Alfred that I have met is a butler.

(*Sarah*) Three Sarahs walked into a bar.

Here, names are being used as predicates that stand for the property of being called by that name. As Rieppel [2013, p. 438] points out, names can also be used to pick out properties other than being called by that name. For instance:

(*Beyoncé*) She is a real Beyoncé.

(*Teresa*) You are no Mother Teresa.

Even indexicals can sometimes act as predicates, as in:

(*Son*) Just look at your son: he is a little you!

So it is possible that the antecedents of alleged counteridenticals are predicative expressions. The predicate theory, therefore, deserves a closer look.

Exhibits A–C as well as Exhibit E seem to provide support for the predicate theory. First take Exhibit A (explicit identity). According to the predicate theory, while real counteridenticals are regimented in our formal language using '=', alleged counteridenticals should be regimented using a special predicate—below, we will use '*be-*' for this special predicate—whose interpretation is as follows: *a be-b* is true just in case the referent of *a* has the contextually salient property that the referent of *b* actually has. Thus, there is no reason to expect (*Button*) and (*Button=*) to be equivalent.

Next, take Exhibit B (equivalence properties). Again, without further argument, there is no reason to suppose that '*be-*' picks out an equivalence relation. Thus, the closest worlds where (*a be-b*) and (*b be-c*) are true do not have to be worlds where (*a be-c*) is true, and the closest worlds where (*a be-b*) is true do not have to be worlds where (*b be-a*) is true. So we should not expect (*Bet*) to be equivalent to (*Bet Rev*), or for (*Trans*) and (*Sym*) to be necessarily true on the predicate theory.

Now take Exhibit C (reflexive pronouns). On the predicate theory, (*Hate*) and (*Self-Hate*) can be regimented respectively as follows:[16]

$$(I\ \textit{be-you} \wedge \textit{you}\ \textit{be-I}) > \textit{Hate}(I, \textit{you}) \tag{1}$$

$$(I\ \textit{be-you} \wedge \textit{you}\ \textit{be-I}) > \textit{Hate}(I, I). \tag{2}$$

Since *I be-you* and *you be-I* do not imply $I = \textit{you}$, there is no reason to think the '*I*' and the '*you*' in the consequent of (1) must corefer or that (1) and (2) are equivalent.

Finally, take Exhibit E (attitude reports). Arguably, the predicate theorist will treat belief ascriptions like (*Kat is Rosa*) in the same way that they treat alleged counteridenticals. So the difference between (*Kat is Rosa*) and (*Kat ≠ Rosa*) can be regimented as follows:

$$\textit{biron believes}\ (\textit{kat be-rosa}) \tag{3}$$

$$\textit{biron believes}\ (\textit{kat} \neq \textit{rosa}). \tag{4}$$

Despite how convenient it would be for the predicate theory to be correct, there is also strong evidence against the predicate theory. We now turn to examining that evidence.

[16] I use '*I*' throughout for the first-person pronoun, regardless of case.

Problem 1: Predicate Coordination. There are a number of grammatical tests that have been developed independently in the literature on copular clauses for determining whether or not a copula has a predicational reading. These tests suggest the 'were' in alleged counteridenticals does not have a predicational reading. To illustrate, I will discuss the simplest (and, to my mind, the most compelling) such test, viz., predicate coordination.[17]

Predicates can typically coordinate with other predicates. For example, observe that the following sentence sounds grammatical, despite the fact that the 'is' only occurs once:

(*Coord*) He is smart, kind, and brave.

The grammaticality of this sentence is explained by the fact that the copulas in "He is smart", "He is kind", and "He is brave" are all predicational. So there is no need to repeat the 'is' multiple times. By contrast, the following sentence is ungrammatical:

(*No Coord*) *Cicero is smart, well-read, and Tully.

This can be explained by the fact that the 'is' in "Cicero is Tully" only has an equative reading. Thus, the copula in (*No Coord*) cannot simultaneously coordinate 'smart', 'well-read', and 'Tully'. This is true even for specificational copulas:

(*Still No Coord*) *The 44th president of the United States is smart, well-read, and Barack Obama.

This suggests that one can test whether the copula in a copular clause has a predicational reading by seeing whether it can be used to coordinate the postcopular expression with other canonically predicative expressions.

Observe that this test does not just test for whether the postcopular expression is a referential term. For instance, observe that the following sentence is grammatical:[18]

(*Mayor*) She is ambitious, driven, and the mayor of Oakland.

The fact that this sentence is grammatical is evidence that the 'is' in "She is the mayor of Oakland" admits a predicational reading, and that 'is the mayor of Oakland' functions semantically as a predicate, *even though* 'the mayor of Oakland' can also function as a referential term (as in "The mayor of Oakland is ambitious").

Also, observe that this test works even in embedded constructions. In particular, it still works even in the antecedents of counterfactuals:

(*Coord*>) If he were smart, kind, and brave, he would have made a great leader.

(*No Coord*>) *If Cicero were smart, well-read, and Tully, he would have been more famous.

(*Mayor*>) If she were ambitious, driven, and the mayor of Oakland, she would have been more popular.

[17] Two other tests yielding the same conclusion involve pseudoclefts and the ability to drop predicational copulas under 'consider'. See Higgins 1979; Rothstein 1995; Mikkelsen 2004, 2005, and Rieppel 2013.

[18] From Rieppel 2013, p. 422.

The ungrammaticality of (*No Coord>*) is no accident. The same observation can be made for other alleged counteridenticals. Here are a few examples to illustrate:

(*Bet No Coord*) *If I were rich, clever, and you, I would bet on that horse.

(*Frege No Coord*) *If Hesperus were not big, bright, and Phosphorus, Frege would need a different example.

(*Give Up No Coord*) *If you were ambitious, confident, and anyone else, you would have given up.

This suggests that the 'were' in alleged counteridenticals is not predicational after all.

To be clear, predicate coordination is not a test of whether a copula has some syntactic feature. Everyone agrees on the syntactic type of 'you': it is a noun phrase (NP). Rather, it is a test of whether a postcopular expression is predicative, i.e., has semantic type $\langle e, t \rangle$. The test *utilizes* surface syntax to determine the semantic type of the postcopular expression, and so in that sense is a "syntactic" test. But what it is *testing for* is a semantic feature of the postcopular expression—viz., what its semantic type is—not a syntactic feature. And this test suggests that the postcopular expression in the antecedent of an alleged counteridentical does not denote a property.

Problem 2: Indicatives. It is not clear that the predicate theory can explain Exhibit D (identity indicatives). The most natural explanation for why identity indicatives generally sound unacceptable or marked is that identity indicatives are not exceptional: the received view is that *most* indicatives whose antecedents are known to be false sound unacceptable or marked.[19] Call this the ***E***pistemic ***P***ossibility ***P***resupposition:

EPP. If the antecedent of an indicative conditional is known or believed with certainty by an agent to be false, then that conditional should sound marked to that agent.

So, we should expect to find that identity indicatives generally sound marked because we generally know that their antecedents are false. By contrast, if we do not know the relevant (non-)identities, then the corresponding identity indicatives sound fine.

Let us assume that something like ***EPP*** is what explains the infelicity of identity indicatives (if the predicate theorist rejects this, the onus is on them to supply an alternative explanation). Then it seems as though the antecedents of their counterfactual counterparts must also contain identity statements. After all, consider (*Einstein*) and (*Einstein Ind*) side by side:

(*Einstein*) If Alyssa were Einstein, she would have aced the test.

(*Einstein Ind*) #If Alyssa is Einstein, she aced the test.

Arguably, the only difference in logical form between (*Einstein*) and (*Einstein Ind*) is in the flavor of the conditional: the former is a counterfactual, while the latter is an indicative. Thus, it is plausible to think that the antecedents of (*Einstein*) and (*Einstein Ind*) have the

[19] Stalnaker [1976, pp. 145–146], Warmbrōd [1983, pp. 250–251], Jeffrey and Edgington 1991, p. 189, and Bennett [2003, pp. 54–57], for instance, argue for this view.

same logical form. But the reason (*Einstein Ind*) sounds marked is that we know that Alyssa is not Einstein. This knowledge looks like knowledge of a nonidentity claim: what we know is that Alyssa is not *identical to* Einstein. So the antecedent of (*Einstein Ind*) seems to be an identity statement, suggesting that the antecedent of (*Einstein*) is too.

The most natural response on behalf of the predicate theorist would be to claim that the 'is' in "Alyssa is Einstein" is also predicational. Thus, counteridenticals and identity indicatives would get a uniform predicational treatment, and the infelicity of (*Einstein Ind*) would be explained by appealing to the fact that we generally know that Alyssa does not have the contextually salient property Einstein has.

However, this puts the predicate theorist in a bind. In order for this style of explanation to work, the predicate theorist needs for the property picked out by 'is Einstein' in a given context to be one that the interpreter knows (or believes with certainty) that *only* Einstein has in that context. Otherwise, it might be epistemically possible that Alyssa has that contextually salient property of Einstein's (even if it is not epistemically possible that Alyssa is identical to Einstein), in which case (*Einstein Ind*) would not generally sound marked.

But the assumption that 'is Einstein' picks out a property that is known to only be had by Einstein is incredibly unrealistic. If a property of a person or object is contextually salient, then that means the speaker has to be aware of the person or object having this property. But it is highly implausible to think that speakers are always aware that such properties are uniquely had by that person or object. One can assert and understand (*Einstein*), even if all one knows about Einstein is that he is a famous physicist.[20] The predicate theorist cannot readily explain why that would be.

Given all of this, I would like to tentatively conclude that the copula in the antecedents of counteridenticals are equative and therefore alleged counteridenticals are genuine ones.[21] I am open to reconsidering this if no other viable alternative emerges from the rubble. After all, on the whole, the predicate theory is still a rather good theory of counteridenticals, and I suspect many will be sympathetic to it. So the predicate theory is a fairly good fallback, in case other theories do not succeed. However, I will argue that we can do better.

[20] Kripke [1980, pp. 80–82] made a similar point against descriptivism about names. Note, however, that the predicate theory is not committed to descriptivism. The objection only targets the claim that for a speaker to interpret or meaningfully use a sentence such as (*Einstein*), there must be some contextually salient property that the speaker believes is uniquely had by Einstein.

[21] An anonymous reviewer suggests that one could grant this but postulate some other mechanism that reinterprets 'you' in 'if I were you' as a description. This suggests counteridenticals might be connected to descriptive indexicals. To account for the latter, for instance, Sæbø [2015] postulates a covert substitution operator at the level of logical form that substitutes an individual concept (a function from worlds to individuals) associated with a term with a unique contextually salient individual concept. The presence of this operator is pragmatically triggered when the literal interpretation of a sentence would be absurd or irrelevant. Perhaps one could use such an operator to give an account of counteridenticals.

Without being more specific about the nature of this operator and its associated pragmatics, the claim that there is a covert substitution operator at the level of logical form strikes me as difficult to evaluate. Concretely, I have two worries about this approach. First, it cannot explain Exhibits A, C, and D. For instance, the same pragmatic forces that would lead us to reinterpret 'if I were you' arguably should lead us to interpret 'if I were identical to you' in the same way, so it is unclear why there would be a felt difference between (*Button*) and (*Button=*). Likewise, it is unclear why the 'me' in (*Vote Me*) is open to reinterpretation while the 'myself' in (*Vote Myself*) is not, or why identity indicatives are not reinterpreted. Second, such a view does not resolve Problem 2 above. Why should we expect a speaker to come to a unique contextually salient individual concept that is substituted for the individual concept associated with 'Einstein' in (*Einstein*)?

§5 Hyperintensionality

In § 3–4, I argued against the most promising strategies for defending the orthodox views on counterfactuals and identity from the challenges posed by counteridenticals. Even so, all else being equal, we should prefer a theory that is as conservative as possible with respect to the orthodoxy. Arguably, the most conservative revision of these theories is to replace the standard semantics for counterfactuals with a semantics that can tolerate non-trivial counterpossibles. In this section, I will assess the plausibility of this approach.

Recently, there has been renewed interest in counterpossibles. On the orthodox semantic theories of counterfactuals, such as the Lewis-Stalnaker semantics, counterpossibles are vacuously true. This is because these semantic theories satisfy the replacement of necessary equivalents (where $\equiv$ is the material biconditional, and $C[A/B]$ is any formula that results from replacing some occurrences of A in C with B):

Replacement. $\Box(A \equiv B) \vDash C \equiv C[A/B]$ if A and B have the same free variables.

In words, ***Replacement*** says that if A and B are necessarily equivalent, then one can replace A with B in any complex formula *salva veritate*. Now, if A is impossible, that means A and $C \wedge \neg C$ are necessarily equivalent. So according to ***Replacement***, a counterpossible of the form $A > C$ is equivalent to $(C \wedge \neg C) > C$, which we may assume is vacuously true.

But at first glance, this seems at odds with intuition. For instance, consider the following counterpossibles:

(*Salmon*) If Sam were a salmon, he would be happy.

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

(*Intuitionism*) If intuitionistic logic were correct, logicians would not be surprised.

Each of these counterfactuals seems to be contingent, i.e., neither necessary nor impossible. So the fact that most orthodox theories of counterfactuals render all counterpossibles vacuously true is not obviously a good feature.

There are roughly two competing theories of how to address this concern. First, there is ***the vacuity theory***, which argues that counterpossibles are indeed vacuous, and that any apparent non-vacuity counterpossibles seem to display should be accounted for purely pragmatically.[22] Second, there is ***the non-vacuity theory***, which argues that counterpossibles are not always vacuous and that we need to revise the orthodox theories of counterfactuals in order to accommodate their non-vacuity. For instance, one common revision proposed by non-vacuity theorists is to take the standard Lewis-Stalnaker semantics and add *impossible worlds* to the models.[23]

The debate between the vacuity theorist and the non-vacuity theorist is by no means settled. Still, it is worth seeing whether one of these theories can account for the felt non-vacuity of counteridenticals. If the vacuity theorist is right, then any felt non-vacuity of

[22] Lewis [1973, pp. 24–26], Bennett [2003, pp. 229–231], and Williamson [2007, pp. 171–175], for example, argue that counterpossibles are vacuous.

[23] See Cohen 1987, 1990; Mares 1997; Nolan 1997; Goodman 2004; Vander Laan 2004; Krakauer 2012; Bjerring 2013; Brogaard and Salerno 2013; Kment 2014, and Berto et al. 2017 for this approach.

counterpossibles must be explained pragmatically. Unfortunately, we already saw in § 3 that such pragmatic theories cannot explain all of the data surrounding counteridenticals. So let us see if the non-vacuity theory can do better. Let ***the hyperintensional theory*** of counteridenticals be the view that the felt non-vacuity of counteridenticals can be explained in exactly the same way we explain the felt non-vacuity of counterpossibles, viz., via a revision of the orthodox semantics of counterfactuals.

The hyperintensionalist can easily accommodate Exhibits B, D, and E. Start with Exhibit B (equivalence properties). Consider again the following sentences:

(*Trans*) If I were you and you were Pope Francis, I would be Pope Francis.

(*Sym*) If I were you, you would be me.

The naïve way of regimenting (*Trans*) and (*Sym*) would be with = and $>$ as follows:

$$(I = you \wedge you = pope) > I = pope \tag{5}$$

$$I = you > you = I. \tag{6}$$

The hyperintensional theory has no trouble falsifying these sentences. Anyone who thinks some counterpossibles are false must reject the following principle:

Strict Entailment. $\Box(A \supset C) \models A > C$.

For since $\neg \Diamond A \models \Box(A \supset C)$ for any A and C, if ***Strict Entailment*** held, it would follow that $\neg \Diamond A \models A > C$, which just states that all counterpossibles are vacuously true. Thus, anyone who thinks some counterpossibles are false must reject ***Strict Entailment***.

But without ***Strict Entailment***, there is no reason to suppose either (*Trans*) or (*Sym*) are true. For even though $\Box((I = you \wedge you = pope) \supset I = pope)$ is valid, still it does not follow that (5) is valid without ***Strict Entailment***. Likewise, even though $\Box(I = you \supset you = I)$ is valid, it does not follow that (6) is valid. Similarly, the hyperintensionalist can differentiate between (*Bet*) and (*Bet Rev*) as follows:

(*Bet*) If I were you, I would bet on that horse.

$$I = you > Bet(I, that\ horse) \tag{7}$$

(*Bet Rev*) If you were me, I would bet on that horse.

$$you = I > Bet(I, that\ horse) \tag{8}$$

Since the hyperintensionalist rejects ***Replacement***, even if $\Box(I = you \equiv you = I)$ is valid, it does not follow that one can swap $I = you$ for $you = I$ in a counterfactual *salva veritate*.

Unlike the predicate theory, the hyperintensional theory has no difficulty explaining Exhibit D (identity indicatives). After all, they agree that sentences like (*Bet*), (*Einstein*), and so forth are genuine counteridenticals. So they are free to appeal to ***EPP*** to explain why the indicativized counterparts of these counterfactuals sound marked. In addition, Exhibit E (attitude reports) can be accommodated so long as the hyperintensionalist supplies us with a hyperintensional semantics for attitude reports (though the success of this approach will depend on the details of the semantics they supply).

Unfortunately, the hyperintensional theory of counteridenticals, like the predicate theory, faces a number of problems, which we will now examine.

Problem 1: Regimentation. It is less than clear that the hyperintensionalist can account for Exhibits A and C. First, consider Exhibit A (explicit identity). It is not clear how the hyperintensionalist even *regiments* the difference between (*Button*) and (*Button*=):

(*Button*) If I were you, I would not push that button.

(*Button*=) If I were identical to you, I would not push that button.

For it seems (*Button*) and (*Button*=) in the hyperintensional theory are both regimented as:

$$(I = you) > Bet(I, that\ horse). \tag{9}$$

Second, consider Exhibit C (reflexive pronouns). The hyperintensionalist might be able to accommodate the felt difference between (*Hate*) and (*Self-Hate*) as follows:

(*Hate*) If I were you and you were me, I would hate you.

$$(I = you \wedge you = I) > Hate(I, you) \tag{10}$$

(*Self-Hate*) If I were you and you were me, I would hate myself.

$$(I = you \wedge you = I) > Hate(I, I) \tag{11}$$

But it is not clear they can regiment the difference between (*Vote Me*) and (*Vote Myself*):

(*Vote Me*) If I were you, I would vote for me.

(*Vote Myself*) If I were you, I would vote for myself.

True, the hyperintensionalist could respond by emphasizing that ***Replacement*** fails for counterfactuals on their view. So even though "I am you" and "I am identical to you" are necessarily equivalent, and even though "I vote for me" and "I vote for myself" are necessarily equivalent, it will not follow, on their view, that we can replace one for the other *salve veritate* in counterfactual environments.

However, this move faces two problems. First, even if one concedes this point to the hyperintensionalist, it is still not entirely clear how to represent these differences at the level of logical form. For instance, what is the difference in logical form between 'is' and 'is identical to'? Perhaps these are distinct but necessarily coextensive relations just like being triangular and being trilateral. But this does not seem very plausible in the case of identity without independent motivation for such a view.

Second, postulating that counterfactuals are sensitive to these grammatical differences seems *ad hoc* unless one maintains that counterfactuals are generally sensitive to even minor grammatical differences. But this seems implausible. Consider the following:

(*Con*$_{PA}$ *Active*) If PA were able to prove its own consistency, logicians would be surprised.

(*Con*$_{PA}$ *Passive*) If the consistency of PA were provable in PA, logicians would be surprised.

The only difference between (*Con*$_{PA}$ *Active*) and (*Con*$_{PA}$ *Passive*) is that the former's antecedent is in active voice while the latter's antecedent is in passive voice. Arguably, this grammatical difference does not matter: (*Con*$_{PA}$ *Active*) and (*Con*$_{PA}$ *Passive*) are equivalent full stop. So the hyperintensionalist owes us an account of which grammatical differences the truth conditions of counterfactuals are sensitive to and why. But no account that differentiates (*Button*) and (*Button=*) and differentiates (*Vote Me*) and (*Vote Myself*) but equates (*Con*$_{PA}$ *Active*) and (*Con*$_{PA}$ *Passive*) seems forthcoming.

Problem 2: Closure. Consider the following sentence:[24]

(*LEM*) If Kripke were an earthworm, the law of excluded middle would still hold.

Many non-vacuity theorists think (*LEM*) is true: while a violation of a *metaphysical* law is bad, a violation of a *logical* law is much worse. One way to put this is in terms of counterpossibles: even if something that is merely metaphysically impossible had obtained, all the laws of logic would still hold. This can be codified in the following principle:

Closure. If $\not\models \neg(A \wedge B)$ and $B \models C$, then $A > B \models A > C$.

In words, if the constituents of a counterfactual are jointly consistent, then all the laws of logic can be applied to the consequent, even if the antecedent is metaphysically impossible.

Closure is a fairly plausible counterfactual principle.[25] However, if ***Closure*** holds, then the hyperintensionalist will no longer be able to accommodate the non-vacuity of (*Trans*) and (*Sym*), or the observation that (*Hate*) seems not to imply (*Self-Hate*). To show this, we need to assume the following almost unanimously accepted principle:

Triviality. $\models A > A$.

First, consider the counterfactual (*Trans*) again:

$$(I = you \wedge you = pope) > I = pope. \tag{5}$$

If the hyperintensionalist accepts both ***Closure*** and ***Triviality***, then (5) will come out valid on their view. Even though $((I = you \wedge you = pope) \wedge I = pope)$ might be metaphysically impossible (in some contexts), it is certainly not logically inconsistent on any reasonable notion of logical consistency. Thus, $(I = you \wedge you = pope) > (I = you \wedge you = pope)$ entails (5) by ***Closure***. But then (5) is valid by ***Triviality***. Likewise, the same reasoning shows that (*Sym*), i.e., (6), is also valid if ***Closure*** and ***Triviality*** hold:

$$I = you > you = I. \tag{6}$$

As for (*Hate*) and (*Self-Hate*), we need to assume an additional (but uncontroversial) counterfactual principle, which essentially says that conjunction introduction and elimination hold in the consequents of counterfactuals:

[24] This example is a modified version of an example from Goodman 2004, p. 36.

[25] In fact, many non-vacuity theorists, such as Krakauer [2012, pp. 76–78], Bjerring [2013, p. 335], and Brogaard and Salerno [2013, pp. 652–653], adopt an even stronger principle, namely if $\not\models \neg A$ and $B \models C$, then $A > B \models A > C$. For our purposes, we only need ***Closure***.

Counterfactual Conjunction. $\models [A > (B \wedge C)] \equiv [(A > B) \wedge (A > C)]$.

If ***Closure***, ***Triviality***, and ***Counterfactual Conjunction*** all hold, then (*Hate*) entails (*Self-Hate*), i.e., (10) $\models$ (11):[26]

$$(I = you \wedge you = I) > Hate(I, you) \tag{10}$$
$$(I = you \wedge you = I) > Hate(I, I). \tag{11}$$

So the hyperintensionalist needs to reject ***Closure*** in order to maintain their ability to accommodate the observations that speak in their favor. This is unfortunate because ***Closure*** is quite plausible as a principle of non-vacuous counterpossibles, and as far as I am aware, there are no other convincing counterexamples to such a principle.

Now, some non-vacuity theorists think that no interesting counterfactual inference is valid (not even ***Triviality***!).[27] So perhaps such a hyperintensionalist would reject ***Closure***. But even if that is right, there is still a lingering conceptual worry for the hyperintensionalist. While it might be true that, technically speaking, the hyperintensional theory can invalidate (*Trans*) and (*Sym*), thereby *accommodating* the felt lack of implication from (*Hate*) to (*Self-Hate*), it does not quite *explain* either of these observations.

After all, not everything goes when one counterfactually supposes something that is metaphysically impossible. Just because it is impossible for Kripke to be an earthworm, that does not mean if Kripke were an earthworm, the laws of physics would break down. But then why would the laws of identity all of the sudden fail had the particular identity facts been different from what they actually are? Would not identity still be transitive if I were you and you were Pope Francis? Would not identity still be symmetric if I were you? Although the flexibility of the hyperintensional theory gives it the ability to accommodate all of these observations above, it also makes it alarmingly straightforward to create a just-so story for nearly any observation one wants to accommodate.

Despite all of the problems that the hyperintensional theory faces, it could still turn out that counteridenticals are just a very special class of counterpossibles. The above remarks do not definitively rule that out. But they do suggest that they are dissimilar enough to warrant questioning the claim that they should be treated in the same way. So I want

[26] Here is the proof. Observe first that $(I = you \wedge you = I) \wedge Hate(I, you) \models Hate(I, I)$. Moreover, the sentence $(I = you \wedge you = I) \wedge Hate(I, you)$ is logically consistent. Hence, by ***Closure***, the first of the following sentences implies the second:

$$(I = you \wedge you = I) > ((I = you \wedge you = I) \wedge Hate(I, you))$$
$$(I = you \wedge you = I) > ((I = you \wedge you = I) \wedge Hate(I, I)).$$

So by ***Triviality*** and ***Counterfactual Conjunction***, (10) implies the first of these sentences. And by ***Counterfactual Conjunction***, the second of these sentences implies (11). Putting everything together, (10) $\models$ (11).

[27] Cohen [1990, p. 131] and Nolan [1997, p. 554] argue for such a view. For instance, let A, B, and C be such that $\not\models \neg(A \wedge B)$ and $B \models C$. We one might think we can refute $A > B \models A > C$ by adding a premise of the form "If A were true, ***Closure*** would have failed". If that is right, then ***Closure*** does not universally hold. In response, I argue that which counterfactual principles are valid depends on the context. In most contexts, we are not willing to tolerate such wild logical impossibilities. So in those contexts, ***Closure*** should still hold. Yet, (*Trans*) and (*Sym*) will still sound false in those contexts, and (*Hate*) will still not imply (*Self-Hate*) in those contexts. So I do not think such considerations help the hyperintensionalist here, even if they are correct (which may also be disputed).

to (again tentatively) conclude that the hyperintensional theory of counteridenticals is incorrect. Counteridenticals reveal nothing about the tolerance counterfactuals have toward impossibilities. Rather, they reveal the contingency of identity.

§6 Counterpart Theory

In § 3–5, I argued against the most promising attempts to resolve the puzzles surrounding counteridenticals without giving up the necessity of identity. I will now show that we can give a satisfactory account of counteridenticals if we abandon the necessity of identity.

Many will find the very notion of contingent identity unacceptable. I suspect many will have been so thoroughly convinced by the arguments from Kripke 1971, 1980 against contingent identity that they will view my endorsement of it as a *reductio* of my previous arguments. At the end of the day, there is little I can say to convince someone who has fully embraced the necessity of identity (though see § 8). But though I think the standard objections to contingent identity can all be addressed, I also think the best defense of it comes from an adequate explanation of counteridenticals. So before we reject the notion of contingent identity as such, I think we should first see how a theory countenancing it can be put to good work.

In this section, I will develop a theory of counteridenticals using counterpart theory, as extensively discussed by Lewis [1968, 1971, 1973, 1986]. To start, I will briefly review the basics of counterpart theory and explain how counterpart theory can also act as a theory of counteridenticals, following a suggestion from Lewis 1973, p. 43. Lewis never discusses any of the puzzles or observations surrounding counteridenticals that we have discussed in previous sections, but I will show how one might deal with such puzzles in the framework Lewis suggests. While the theory sketched in this section is a solid first approximation of the theory I ultimately endorse, I will argue in § 7 that it is prone to counterexample and thus needs (modest) revision.

According to the orthodox semantics for modality, the claim "Alice could have been tall" is true just in case there is a possible world where Alice is tall. Lewis, however, thinks this is incorrect because he holds that objects are world-bound—that is, no object can only exist at more than one possible world. What makes it the case, according to Lewis, that Alice could have been tall is not that there is a possible world where Alice *herself* is tall, but rather that there is a possible world where a *counterpart* of Alice is tall.

More generally, let us say a term t is ***free in*** A if it is either a constant or a free variable in A. Then where $t_1, \ldots, t_n$ are the free terms in A, according to Lewis [1973, p. 40], $\Diamond A$ is true just in case at some world, some counterparts of the objects denoted by $t_1, \ldots, t_n$ at that world satisfy A. Similarly for counterfactuals. For simplicity, let us suppose that A and C have the same free terms $t_1, \ldots, t_n$. Then according to Lewis [1973, p. 42], $A > C$ is true just in case at all the closest worlds where some counterparts of the objects denoted by $t_1, \ldots, t_n$ satisfy A, those counterparts also satisfy C at that world.

Counterparthood, for Lewis, is determined by considerations of similarity. A counterpart of Alice at a world w is any object that is at least as similar to Alice at the actual world as any other object is at w. On this understanding of counterparthood, there is no reason why Alice should have exactly one counterpart at every world. Maybe at some worlds, she

has multiple counterparts, while at others she has none. Maybe at some worlds, she shares a counterpart with someone else. Moreover, different notions of similarity will generate different counterpart relations. And this last point is the key to Lewis's theory of counteridenticals [Lewis, 1973, p. 43]:

> For a familiar illustration of the need for counterpart relations stressing different respects of comparison, take 'If I were you . . . '. The antecedent-worlds are worlds where you and I are vicariously identical; that is, we share a common counterpart. But we want him to be in *your* predicament with *my* ideas, not the other way around. He should be your counterpart under a counterpart relation that stresses similarity of predicament; mine under a different counterpart relation that stresses similarity of ideas.

This paragraph exhausts Lewis's discussion of counteridenticals. As far as I am aware, he never develops this idea in any further detail in subsequent work. So in what follows, I will attempt to reconstruct what Lewis's theory of counteridenticals might have looked like in more detail. I will call this theory ***Lewis's theory*** for brevity. But this reconstruction is just an educated guess at what his theory would have looked like had it been more fully developed in light of the data discussed above, and one should bear in mind that Lewis himself may or may not have agreed with the theory that follows.

The idea behind "Lewis's theory" is to attach counterpart relations to terms so that we know which kind of counterpart to look at in modal contexts. We do this by adding to each term an index, the value of which will be a counterpart relation determined by context. Then we use the truth conditions above, except we quantify over the counterparts via the relations associated with the indices at that context. For instance, we can formalize (*Bet*) with counterpart indices as follows:

$$(I^1 = you^2) > Bet(I^1, that\ horse^3). \tag{12}$$

So (12) is true if in all the closest worlds where a C_1-counterpart of the speaker is a C_2-counterpart of the listener, that counterpart will bet on every C_3-counterpart of that horse (though generally in such worlds, there will only be one such counterpart). Context will determine which counterpart relation to associate to a given counterpart index. In an ordinary context in which (*Bet*) would be asserted, C_1 might be the belief-counterpart relation, C_2 the predicament-counterpart relation, and C_3 the body-counterpart relation.

Just as it does not matter whether one uses the variable x_1 or x_{81} when regimenting "Someone is tall" in first-order logic, so too it does not matter whether one attaches the counterpart index 1 or 81 to I in (12). What is important is not the particular index attached to the variable but rather which indices match. This, as we will see, is what explains the data that we have encountered surrounding counteridenticals.

How does one determine which counterpart indices on which terms should match? Lewis never really gives us a systematic answer to the question. In Lewis 1971, p. 209, he suggests that somehow the sense of a term determines which counterpart relations to attach to it. But that does not really give us much of a guide to regimenting a given English sentence into this language with counterpart indices. Later in Lewis 1986, pp. 254–263, he seems to suggest that determining how to associate counterpart relations to terms is just a

highly context-sensitive process, and that there is not much more to say about the matter. But one might worry that this renders Lewis's theory too flexible to make any meaningful predictions.

Below I will propose and motivate some general rules and guidelines for regimenting a given English sentence into its logical form with counterparts. In doing so, I will show how counterpart theory can explain all of the puzzling features of counteridenticals we have encountered in previous sections. While I do not have a deterministic recipe for going from English into loglish, I hope that what I say will at least constrain Lewis's theory enough so as to make concrete predictions about counteridenticals and related discourse.

Perhaps the most useful guideline for regimenting English sentences into counterpart theory comes from Lewis 1971, p. 210, where he observes that which counterpart relation is attached to a term is sometimes determined explicitly with special clauses modifying the term. For instance, it seems like in an ordinary context in which (*Bet*) is asserted, one could have equally said:

(*Bet Qualified*) If I (with my beliefs) were you (as a bettor), I (with my beliefs) would bet on that horse (being the fastest horse here).

Phrases like 'with my beliefs' and 'as a bettor' can be used to make explicit which counterpart relation attaches to which terms.

We can use this observation to guide our regimention of a given English sentence, as asserted in a given context, into counterpart theory. So when regimenting an English sentence into counterpart theory, one can follow this general procedure. First, regiment the English sentence without any counterpart indices. Second, go term by term and determine what phrase of the form 'as ___', 'with ___', 'being ___', or '*qua* ___' would be appropriate to modify that term with in that context.[28] Finally, go back and assign counterpart indices to terms and counterpart relations to counterpart indices that would make sense of those phrases. In other words:

The Qualifier Test. To determine how to assign counterpart indices to terms in a sentence, modify each term with qualifier phrases like 'as ___', 'with ___', 'being ___', or '*qua* ___' and fill in the blanks accordingly. Assign a separate index to each distinct way of filling in the blanks.

This process is not totally precise, but it is precise enough to be a helpful guide to regimenting English sentences in counterpart theory. We will use this test extensively to motivate the regimentation principles below.

Let us start by applying ***The Qualifier Test*** to Exhibit C (reflexive pronouns):

(*Vote Me*) If I were you, I would vote for me.

(*Vote Myself*) If I were you, I would vote for myself.

The reason these sound different is that the 'I' and the 'myself' in the consequent of (*Vote Myself*) seem to refer to the same person, whereas the 'I' and the 'me' in the consequent of (*Vote Me*) seem to refer to different people, in some previously inexplicable sense.

[28] I do not mean to imply these phrases are the only kinds of phrases that can be used for this purpose. I will focus on these phrases for simplicity and concreteness.

Using mismatching counterpart indices, we can now explicate the sense in which the 'I' and 'me' refer to different people. The 'I' and the 'me' *do* refer to different people: they refer to different counterparts of the speaker. That is, (*Vote Me*) is regimented as:

$$I^1 = you^2 > VoteFor(I^1, I^3). \tag{13}$$

This is in accord with ***The Qualifier Test***, since (*Vote Me*) seems to be saying something like the following:

(*Vote Me Qualified*) If I (with my beliefs and preferences) were you (*qua* voter), I (with my beliefs and preferences) would vote for me (*qua* candidate).

Given that C_1 is the preference-counterpart relation, C_2 is the voting-counterpart relation, and C_3 is the candidate-counterpart relation, (13) is true just in case in all the closest worlds where some preference-counterpart of mine is a voting-counterpart of yours, that counterpart votes for every candidate-counterpart of mine. And in general, in such worlds, my preference-counterpart and my candidate-counterpart will come apart.

Contrast this with how (*Vote Myself*) is regimented:

$$I^1 = you^2 > VoteFor(I^1, I^1). \tag{14}$$

Using the same assignment of counterpart relations to counterpart indices, (14) is true just in case in all the closest worlds where some preference-counterpart of mine is a voting-counterpart of yours, that counterpart votes for itself. Because the counterpart indices on the two occurrences of I in the consequent match, they must denote the same person, just as two occurrences of a single variable denote the same object.

This seems true for reflexive pronouns in general.[29] This leads me to postulate the following regimentation rule:

Reflexive Pronouns. By default, the counterpart indices that attach to reflexive pronouns must match with the name or personal pronoun it is anaphoric on.

This principle for regimenting reflexive pronouns can also explain why (*Hate*) does not imply (*Self-Hate*):

(*Hate*) If I were you and you were me, I would hate you.

(*Self-Hate*) If I were you and you were me, I would hate myself.

Suppose I just broke your violin, but because you are such a forgiving person, you do not hold it against me. It seems like one way of filling out (*Hate*) in that context would be:

(*Hate Qualified*) If I (with my temperament) were you (with your broken violin) and you (with your temperament) were me (being the one who broke the violin), I (with my temperament) would hate you (having broken the violin).

[29] I say "in general", since one might want to regiment sentences like "He was not himself that day" so that 'he' and 'himself' attach to different counterpart relations. I am more inclined to treat these are metaphorical and literally false; but settling the matter would take us too far afield for now.

In that case, ***The Qualifier Test*** would recommend we regiment (*Hate*) as:

$$(I^1 = you^2 \wedge you^1 = I^3) > Hate(I^1, you^1). \tag{15}$$

But by the same token, we should regiment (*Self-Hate*) as the following according to ***Reflexive Pronouns***:

$$(I^1 = you^2 \wedge you^1 = I^2) > Hate(I^1, I^1). \tag{16}$$

This seems to conform to our intuitions. What is striking about (*Hate*) was that the 'I' and the 'you' in the consequent seem to refer to different objects, despite the fact that they are identified in the antecedent. Whereas the 'I' in the consequent seems to refer to the same person as the *first* 'I' in the antecedent, the 'you' in the consequent seems to refer to the same person as the *second* 'you' in the antecedent. By contrast, the 'I' and 'myself' in the consequent of (*Self-Hate*) seem to refer to the same person. What is more, it seems as though the second conjunct in the antecedent is not redundant in either counterfactual, in that it does not say the same thing as the first. When counterpart theory is equipped with regimentation principles like ***The Qualifier Test*** and ***Reflexive Pronouns***, these intuitions are easily corroborated.

Why does ***Reflexive Pronouns*** hold? Ultimately, I think ***Reflexive Pronouns*** is to be explained by syntax. Not every logical form that we can write down in the counterpart-theoretic language will correspond to a natural language sentence. A binding theory should be able to supply an explanation for why rules like ***Reflexive Pronouns*** hold.[30] But for now, I do not want to get too caught up in working out the details. My focus instead is on how counterpart theory can regiment these logical forms.[31]

The fact that 'I were you' and 'you were me' in (*Hate*) and (*Self-Hate*) are formalized differently in counterpart theory also helps us explain Exhibit B (equivalence properties). This is illustrated by (*Bet*) and (*Bet Rev*):

(*Bet*) If I were you, I would bet on that horse.

(*Bet Rev*) If you were me, I would bet on that horse.

Now, it is true that our formalization of (*Bet*), viz., (12), is equivalent to:

$$(you^2 = I^1) > Bet(I^1, that\ horse^3). \tag{17}$$

But the difference between (*Bet*) and (*Bet Rev*) in our formal language is not represented by a difference in the order of the terms. Rather, it is represented by a difference in which counterpart indices match.

To see what the formalization of (*Bet Rev*) should really be, it helps to consider an example. Suppose Alpha and Beta are about to compete in a horse race. Alpha is an expert

[30] For work along these lines, see Percus and Sauerland 2003 and Arregui 2007. It should be noted that the principle ***Reflexive Pronouns*** is essentially a counterpart-theoretic version of the principle of c-command governing the licensing of reflexive pronouns in general.

[31] Many thanks to Carina Kauf, Sarah Zobel, and an anonymous reviewer for encouraging me to clarify this point.

rider, while Beta is a mere novice. Beta's horse, Carrot, happens to be a top-notch horse—perhaps the best in the race—but Alpha knows that does not compensate for Beta's naïveté. Beta, in a stroke of hubris, tells Alpha that she should bet on Carrot. Alpha coolly replies, "If you were me, I would."[32]

In this case, ***The Qualifier Test*** seems to yield something like the following:

(*Bet Rev Qualified*) If you (as Carrot's rider) were me (with my riding skills), I (as a bettor) would bet on Carrot (being the fastest horse here).

So applying ***The Qualifier Test***, it seems like we should formalize (*Bet Rev*) as follows:

$$(you^1 = I^2) > Bet(I^3, that\ horse^4). \tag{18}$$

In (*Bet*), the person picked out by 'I' in the antecedent seems to be the same person that is picked out by 'I' in the consequent. This is captured in (12) by the fact that I in the antecedent and the consequent have the same counterpart index. By contrast, in (*Bet Rev*), the 'me' in the antecedent seems to pick out someone different from the person picked out by 'I' in the consequent. This is captured in (18) by the fact that I in the antecedent and the consequent have different counterpart indices.

This illustrates that the formal language presented here differs from English in a somewhat artificial respect. While the order of terms in identity statements matters in English, the order of terms in identity statements in our formal language is less important than which indices in the sentence match. This is not a bug but a feature of our language: this allows us to more clearly represent an aspect of the truth conditions of counteridenticals that the order of terms was representing in English sentences (compare this to the feature of formal languages regimenting English—but not of English itself—that the scope of a conjunction or disjunction is marked by parentheses). By convention, we assume that when we formalize English counteridenticals into our formal language, we keep the order of the terms as they are in English and distribute counterpart indices to the terms accordingly. But this is just a matter of convenience. We could have just as well formalized (*Bet*) with (17) rather than (12), just as we could have formalized (*Bet*) using counterpart indices 43, 11, and 238 instead of 1, 2, and 3. Nothing hinges on which convention we adopt, so long as the relevant counterpart indices match in the end.

The observation that (*Trans*) and (*Sym*) generally sound false is also accounted for with mismatching indices:

(*Trans*) If I were you and you were Pope Francis, I would be Pope Francis.

(*Sym*) If I were you, you would be me.

One gets the feeling that the two instances of 'you' in the antecedent of (*Trans*) do not refer to the same person, and that the 'you' in the antecedent of (*Sym*) does not refer to the same person as the 'you' in the consequent. Counterpart theory can regiment this intuition about (*Trans*) and (*Sym*) respectively as follows:

$$(I^1 = you^2 \wedge you^1 = pope^3) > (I^1 = pope^3) \tag{19}$$

[32] Lakoff [1996, p. 96] claims that sentences like (*Bet Rev*) are "decidedly strange"; Reboul [1996, p. 172] similarly claims the felicity of such sentences is somewhat questionable. This example seems to count against this judgment; rather, I think their apparent infelicity should be explained by appeal to relevance.

$$(I^1 = you^2) > (you^1 = I^2). \tag{20}$$

This does not conflict with the fact that identity is an equivalence relation, since the differing counterpart indices allow the denotation of the terms to come apart in modal contexts. Thus, you^1 and you^2 in (19) or (20) need not pick out the same object.

Contrast (*Trans*) and (*Sym*) with the following sentences:

(*Trans*=) If I were identical to you and you were identical to Pope Francis, I would be identical to Pope Francis.

(*Sym*=) If I were identical to you, you would be identical to me.

It seems that (*Trans*=) and (*Sym*=) (unlike (*Trans*) and (*Sym*)) are necessarily true: they just illustrations of the fact that identity is an equivalence relation. This motivates the following regimentation principle:[33]

Explicit Identity. In general, the occurrences of terms flanking explicit identity phrases (such as 'identical (to)' or 'the same such-and-such (as)') must all have matching counterpart indices in those positions.

Using ***Explicit Identity***, (*Trans*=) and (*Sym*=) can be regimented as:

$$(I^1 = you^1 \wedge you^1 = pope^1) > (I^1 = pope^1) \tag{21}$$

$$(I^1 = you^1) > (you^1 = I^1). \tag{22}$$

Because the counterpart indices on the terms match, we can safely apply transitivity and symmetry to the identity claims, in which case both (21) and (22) come out valid.

Consider how this applies to Exhibit A (explicit identity):

(*Button*) If I were you, I would not push that button.

(*Button*=) If I were identical to you, I would not push that button.

Using ***Explicit Identity***, the difference between (*Button*) and (*Button*=) can be explained by a difference in the counterpart index on '*you*':

$$(I^1 = you^2) > \neg\, Push(I^1, that\ button^3) \tag{23}$$

$$(I^1 = you^1) > \neg\, Push(I^1, that\ button^3). \tag{24}$$

Since $I^1 = you^2$ and $I^1 = you^1$ are not necessarily equivalent, (23) and (24) need not be necessarily equivalent either.

One might worry that if we follow ***Explicit Identity*** and regiment "a is b" as $a^1 = b^2$ and "a is identical to b" as $a^1 = b^1$, then we would falsely predict that 'is' and 'is identical to' are not equivalent. Indeed, according to counterpart theory, $a^1 = b^2$ is not necessarily

[33] This principle is corroborated by the fact that 'if I were you' is typically equivalent to 'if you and I were the same person'; see footnote **11**. As mentioned above, I suspect further investigation into the syntax of natural language will reveal why ***Explicit Identity*** holds. But I must set that aside for future work.

equivalent to $a^1 = b^1$ in general (i.e., $\not\models \Box(a^1 = b^2 \equiv a^1 = b^1)$). But then why do 'is' and 'is identical to' seem synonymous outside modal environments?

Note that this is going to be a problem for any theory that accounts for Exhibit A—that is, on any theory that distinguishes (*Button*) and (*Button=*) while treating 'is' and 'is identical to' as equivalent outside modal environments. In particular, the predicate theory also predicts that there would be a difference between 'is' and 'is identical to' even outside the scope of a modal. The way the predicate theory would handle this would be to postulate that the predicate *'be-b'* must pick out a property that b and only b has at the actual world. We saw some worries with making such a stipulation in § 4 (Problem 2). And in any case, the stipulation seems *ad hoc* in that it is not motivated or explained by any other part of the predicate theory.

Counterpart theory can address this concern by distinguishing between *logical* equivalence (truth at all contexts) and *necessary* equivalence (truth at all possible worlds).[34] Consider an analogy with the operator @, where @A is true at a world iff A is true at the actual world. Are A and @A equivalent? The answer depends on what one means by 'equivalent'. On the one hand, A and @A are logically equivalent insofar as they hold at all the same contexts: A is true at a context iff A is true at the designated actual world of the context, and that is true iff @A is true at that context. On the other hand, A and @A are not necessarily equivalent: if A is contingently true at the actual world, then @A is true at every possible world, though A need not be.

A similar story can be told about 'is' and 'is identical to' in counterpart theory. Recall that counterparthood is determined by similarity. But arguably, nothing is actually as similar to an object as that object is to itself.[35] Hence, for any counterpart relation C, one would expect that an object's actual C-counterpart will always be itself. Therefore, $a^k = a^n$ will be true at every context ($\models a^k = a^n$), though it will not be true at every possible world ($\not\models \Box(a^k = a^n)$) except when $k = n$.[36] From this, one can show that $a^k = b^n$ and $a^k = b^k$ are logically equivalent ($\models a^k = b^n \equiv a^k = b^k$), and hence "$a$ is b" is logically equivalent to "a is identical to b", even though they are not necessarily equivalent ($\not\models \Box(a^k = b^n \equiv a^k = b^k)$). In other words:

Actual Identity. Counterpart indices do not affect the denotation of a term that is not within the scope of a modal—i.e., they do not affect the truth value of actual identity claims.

Not only does this accord with our observations regarding explicit identity claims, it is also directly motivated by the very concept of a counterpart.

[34] See, for instance, Crossley and Humberstone 1977; Kaplan 1977, and Davies and Humberstone 1980.

[35] Note we need not assume with Lewis [1986, pp. 229–230] that similarity has to be understood in purely qualitative terms (Lewis [1968, 1971, 1973] does not assume this either). Thus, even if I have a qualitative duplicate somewhere, it is not as similar to me as I am to me (for example, it is not located in this very spot). Still, given objects are world-bound, determining other worldly counterparts will arguably be a purely qualitative matter. Thanks to an anonymous reviewer for encouraging me to say more about this issue.

[36] Henceforth, I assume $\models$ means preservation of truth at all contexts. Also, note that while Lewis [1968] held that the only counterpart of an object at the world where it exists is itself, he later came to reject this constraint [Lewis, 1986, pp. 231–232]. While I disagree, we do not need to settle the issue as long as we find some plausible justification for accepting $\models a^k = a^n$.

Using ***Actual Identity***, we can explain Exhibit D (identity indicatives). Notice that generally speaking, indicatives of the form $A \rightarrow C$ and $@A \rightarrow C$ are logically equivalent since A and $@A$ are logically equivalent (even though A and $@A$ are not necessarily equivalent). This is a fact about epistemic modals more generally: epistemic modals obey the replacement of logical equivalents, whereas metaphysical modals do not. For instance, "Alice knows that it is raining" is equivalent to "Alice knows that it is raining at the actual world". Thus, if a speaker knows that A is false, they will generally find $@A \rightarrow C$ marked in accordance with ***EPP***. Likewise, $(a^1 = b^2) \rightarrow C$ and $(a^1 = b^1) \rightarrow C$ are logically equivalent because $a^1 = b^2$ and $a^1 = b^1$ are logically equivalent (even though $a^1 = b^2$ and $a^1 = b^1$ are not necessarily equivalent). So if a speaker knows that $a^1 = b^1$ is false, they will generally find $(a^1 = b^2) \rightarrow C$ marked in accordance with ***EPP***.

Finally, counterpart theory can give an account of Exhibit E (attitude reports).[37] For instance, the difference between (*Kat is Rosa*) and (*Kat ≠ Rosa*) can be formulated as:

$$\textit{biron believes } (\textit{kat}^1 = \textit{rosa}^2) \tag{25}$$

$$\textit{biron believes } (\textit{kat}^1 \neq \textit{rosa}^1). \tag{26}$$

Say C_1 is the person-counterpart relation and C_2 is the appearance-counterpart. Then (25) is true iff in all the worlds compatible with what Biron believes, every person-counterpart of Katherine is an appearance-counterpart of Rosaline. This seems to be exactly what one is trying to say when they assert (*Kat is Rosa*) in relation to this example.[38] Thus, the counterpart theorist has an adequate account of all of Exhibits A–E.

§7 Substitution of Identicals

So far, Lewis's theory of counteridenticals seems to hold up fairly well. However, I will now consider one problem (regarding the substitution of identicals) for Lewis's theory as stated.[39] Consider the following argument:

(*Hesp*) If Hesperus were not identical to Phosphorus, Hesperus would be the second planet from the sun.

(*Hesp = Phos*) Hesperus is identical to Phosphorus.

(*Phos*) If Hesperus were not identical to Phosphorus, Phosphorus would be the second planet from the sun.

[37] See, for instance, van Rooij 2006 for a counterpart-theoretic semantics for belief ascriptions. See Aloni 2005; Cumming 2008, and Ninan 2012 for related approaches to belief ascriptions.

[38] The intuitive interpretation of the counterpart indices in attitude reports may not be quite the same as in metaphysical modals. Rather than using similarity to characterize counterparts, a more natural approach for belief ascriptions would be to use something like acquaintance relations [Lewis, 1979; Ninan, 2012].

[39] This problem is closely related to one raised against Lewis's counterpart-theoretic treatment of necessity by Hazen [1979, pp. 328–329] regarding counterparts of pairs.

Using the regimentations rules from § 6, this argument is regimented as follows (in this example, it would be natural interpret both C_1 and C_2 as the physical counterpart relation):

$$h^1 \neq p^1 > Sec(h^1, s^2) \tag{27}$$

$$h^1 = p^1 \tag{28}$$

$$\therefore h^1 \neq p^1 > Sec(p^1, s^2). \tag{29}$$

This argument is intuitively invalid; the premises could be true while the conclusion was false. This is especially apparent if we add the premise "If Hesperus were not identical to Phosphorus, Hesperus and Phosphorus would not be equidistant from the sun." Unfortunately, Lewis's counterpart theory renders this argument valid.

The problem is essentially that Lewis's theory quantifies over all combinations of counterparts that satisfy the antecedent. For instance, (27) is true just in case at all the closest worlds where a C_1-counterpart of Hesperus is distinct from a C_1-counterpart of Phosphorus, that C_1-counterpart of Hesperus is the second planet from every C_2-counterpart of the sun. But Hesperus *is* Phosphorus. So (27) is true just in case at all the closest worlds where Hesperus/Phosphorus has two distinct C_1-counterparts, the first of them is the second planet from every C_2-counterpart of the sun. But the order of quantification does not matter; we could have easily chosen the "second" of these C_1-counterparts to be the first. So at all such worlds, *every* C_1-counterpart of Hesperus/Phosphorus is the second planet from every C_2-counterpart of the sun. The same applies to (29): given (*Hesp = Phos*), (*Hesp*) is true iff (*Phos*) is true.

This suggests Lewis's theory needs revision. Fortunately, such a revision can be supplied using a distinction Lewis [1986] introduces between a *possible world* and a *possibility*[40] (though Lewis surprisingly never incorporated these revisions into his theory of counterfactuals).[41] Lewis [1986, p. 231] illustrates the distinction as follows:

> I might have been one of a pair of twins. I might have been the first-born one, or the second-born one. These two possibilities involve no qualitative difference in the way the world is…They differ in respect of 'cross-identification'; that is, they differ in what they represent, *de re*, concerning someone…But they are not two worlds. They are two possibilities within a single world.

In terms of counterpart theory, a possibility is a world plus a way of assigning counterparts. More technically, we can think of a possibility as a world-assignment pair, where the assignments map names and variables to objects. A possibility $\langle v, h\rangle$ is ***accessible*** to a possibility $\langle w, g\rangle$ if for each term t, $h(t)$ at v is a counterpart of $g(t)$ at w. Then $\Diamond A$ is true at a possibility $\langle w, g\rangle$ iff there is an accessible possibility $\langle v, h\rangle$ such that A is true at $\langle v, h\rangle$. So for example, "Lewis might have been one of a pair of twins" is true relative to a possibility $\langle w, g\rangle$ iff there is a possibility $\langle v, h\rangle$ accessible to $\langle w, g\rangle$ (i.e., where $h(lewis)$ at v is a counterpart of $g(lewis)$ at w) such that $h(lewis)$ is one of a pair of twins at v.

[40] See Russell 2013 and Schwarz 2013 for further development of this distinction.

[41] It should be noted that the the counterpart-theoretic truth conditions for counterfactuals given in Lewis 1986, where he makes the distinction we are about to discuss, is the same as those discussed in § 6.

With this distinction between possibilities and possible worlds in place, we can state the revision needed to counterpart theory as follows.[42] Originally, we said a counterfactual of the form $A > C$ (where $t_1, \ldots, t_n$ are the free terms in A and C) is true at *possible world* w iff at all the closest *possible worlds* to w where some counterparts (of the right kind) of the objects denoted by $t_1, \ldots, t_n$ satisfy A, those counterparts also satisfy C. Now, we will say $A > C$ is true at a *possibility* $\langle w, g \rangle$ iff at all the closest *accessible possibilities* to $\langle w, g \rangle$ that satisfy A, C is also satisfied. The formal details are made precise in § A.

Let us see now how this distinction can be used to invalidate the argument above. Suppose (*Hesp*) and (*Hesp = Phos*) are true at a possibility $\langle w, g \rangle$. That means that at all the closest possibilities where two distinct C_1-counterparts of Hesperus/Phosphorus are represented as being Hesperus and Phosphorus respectively, the counterpart that is represented as being Hesperus is the second planet from the C_2-counterpart represented as being the sun. But that does not guarantee (and in fact conflicts with the claim) that at all the closest such possibilities, the counterpart that is represented as being Phosphorus is the second planet from the C_2-counterpart represented as being the sun. After all, at such possibilities, different counterparts are represented as being Hesperus and Phosphorus respectively. So the inference is blocked.

§8 Contingent Identity

Counterpart theory provides a fruitful and comprehensive theory of counteridenticals—or so I have argued. In my view, that alone is a sufficient reason to adopt counterpart theory and to reject the necessity of identity. However, I take it others will need more convincing. A complete defense of the contingency of identity would require a paper all of its own, and many of the standard objections to contingent identity are dealt with elsewhere.[43] Still, I will conclude by attempting to defuse one worry about the form of contingent identity I am defending. In doing so, I hope to at least shed some light on where contingent identity theorists and their opponents disagree.[44]

Some might find it difficult to imagine how two individuals could have been one or

[42] This proposal closely resembles the semantics for belief ascriptions in Aloni 2005; van Rooij 2006; Cumming 2008, and Ninan 2012. It is also related to the solution due to Hazen [1979, p. 334] to the problem in footnote **39**. For a closely related view, see Holliday and Perry 2014.

[43] See Gibbard 1975; Lewis 1986; Stalnaker 1986, and Schwarz 2013, 2014.

[44] Note that according to the counterpart theorist, ***NecId*** is ambiguous between two principles:

NecId Mix. $a^n = b^k \models \Box(a^n = b^k)$ and $a^n \neq b^k \models \Box(a^n \neq b^k)$ for any names a and b and any counterpart indices n and k.

NecId Match. $a^n = b^n \models \Box(a^n = b^n)$ and $a^n \neq b^n \models \Box(a^n \neq b^n)$ for any names a and b and any counterpart index n.

Adding ***NecId Match*** to counterpart theory is equivalent to requiring each counterpart relation C to be an injective function, i.e., every object has at most one C-counterpart at every world and no two distinct objects share a C-counterpart at any world. Adding ***NecId Mix*** on top of that is equivalent to requiring, in addition, that there is effectively only one counterpart relation (or more accurately, all the counterpart relations are coextensive). Combining counterpart theory with ***NecId Mix*** would seem to defeat the whole purpose of using counterpart theory. Whether or not the counterpart theorist should endorse ***NecId Match***, however, is up for debate. In what follows, I focus on ***NecId Mix***, since this is the version of ***NecId*** that the counterpart theorist must reject. But for the record, I think the counterpart theorist should also reject ***NecId Match***.

how one individual could have been two. For instance, it seems intuitive that Aristotle is necessarily Aristotle, i.e., Aristotle could not have failed to have been Aristotle. However, the contingent identity theorist seems to reject this intuition: on their view, Aristotle might not have been Aristotle. And that, to many, will seem unacceptable.

It is true that the formula $\Box(a^1 = a^2)$—i.e., the regimentation of the claim that Aristotle is necessarily Aristotle—is falsifiable on counterpart theory. However, the "intuition" that Aristotle is necessarily Aristotle should not be conflated with another intuition, viz., that Aristotle is necessarily self-identical. On the Kripkean view, these are synonymous. But on counterpart theory, they are not. Given that 'self-identical' means 'is identical to itself', the claim that Aristotle is necessarily self-identical is regimented as $\Box(a^1 = a^1)$, which is a theorem of counterpart theory. Thus, when citing this intuition against counterpart theory, one must take care not to conflate the claim that Aristotle is necessarily Aristotle with the claim that Aristotle is necessarily self-identical.

Even with this in mind, I think the non-vacuity of counteridenticals puts some pressure on this "intuition" that Aristotle is necessarily Aristotle. Consider the sentence:

(*Aristotle*) Aristotle might not have been Aristotle; no one would be interested in philosophy today if Aristotle were not Aristotle.

To my ears, not only does (*Aristotle*) sound felicitous, it sounds like something that might be said in an ordinary conversation about Greek philosophy. If that is right, then we should not hastily accept as an unquestionable datum that Aristotle is necessarily Aristotle.

Now, one could respond, as Kripke [1971, p. 149] does, that this is just loose talk: when someone says (*Aristotle*), what they mean is that Aristotle might have been a different *kind* of person, not that he might have been a different person. Thus, while (*Aristotle*) might be false strictly speaking, we can nevertheless communicate something true by asserting it.

Everyone—even the counterpart theorist—agrees that (*Aristotle*) could be used to communicate that Aristotle could have been a different kind of person. The interesting question is *why* (*Aristotle*) can be used to communicate this. According to the Kripkean, sentences such as (*Aristotle*) are metaphorical: while they are literally false, they are true in some non-literal sense (we raised problems for this view in § 3). As a result, such sentences do not reveal anything interesting about the nature of identity. By contrast, according to counterpart theory, (*Aristotle*) is literally true, and it implies that Aristotle could have been a different sort of person. The counterpart theorist takes the non-vacuity of such sentences at face value and holds that they do reveal something interesting about identity.

Here is another way to state the difference. On the view I am defending here, the property of being Rosaline, say, is not just some primitive irreducible property that Rosaline and only Rosaline has. Rather, the property of being Rosaline is determined by a certain role that Rosaline occupies. The identity facts are not independent facts over and above facts about such roles.[45] This is why to say Biron believes Katherine is Rosaline is to say that in all of Biron's belief-worlds, the person that occupies the Rosaline-role is Katherine. This is also why we can generally paraphrase counteridenticals in terms of talk of positions. It is not that counteridenticals are "just loose talk" for talk of positions—that would

[45]For discussion of a related view, see Dasgupta 2009; Russell 2013, and Shumener 2017.

make such paraphrasing seem like a happy coincidence. Rather, identity facts *just are* facts about roles or "positions".

One should note that this is a metaphysical issue, not just a semantic one. As a good Kripkean would rightfully point out, identity claims are not synonymous with claims about roles. This can be established very easily with a modal argument: intuitively, Rosaline could have failed to be the person who occupies the role Rosaline in fact occupies. Thus, 'Rosaline' and 'the person who occupies the role Rosaline in fact occupies' are not equivalent expressions. The counterpart theorist does not deny this. Identity *facts* are facts about roles or positions, just as heat facts are facts about mean molecular kinetic energy. But that does not make identity *claims* synonymous with role or position claims any more than it makes claims about heat synonymous with claims about mean molecular kinetic energy. Thus, the two sides disagree not just on semantics, but also on metaphysics: the Kripkean holds that identity facts are irreducible whereas the counterpart theorist holds that identity facts are reducible to facts about roles.

§9 Conclusion

Counteridenticals exhibit a number of puzzling features. In addition to their apparent non-vacuity, there is evidence suggesting both that their antecedents do contain identities and that they do not. We saw that any attempt to resolve this tension while maintaining the necessity of identity proved unsatisfactory. Instead, I proposed we give up the necessity of identity and showed how to explain these features in a unified and systematic way using the tools of counterpart theory.

I have argued at length for a particular view on the nature of identity. Yet I have not said much about its philosophical significance on questions regarding the nature of persons, persistence, and belief ascriptions—not because I think it has nothing to say about these topics, but simply because doing so would require separate treatment. Nor have I spent much time elaborating on the ways in which this view complicates our picture of belief, communication, and the divide between the necessary and the *a priori*. Again, that must be saved for another time. My goal in this paper was simply to establish that the thesis of contingent identity, despite the overwhelming consensus to the contrary, is here to stay.

§A Formal Details

In this appendix, I will lay out the formal details of the version of counterpart theory sketched in § 7. My approach to counterpart theory largely follows Kracht and Kutz 2001, 2005 and Schwarz 2014. While I will use the Lewis-Stalnaker semantics as my base semantics for concreteness, this is not essential. We could have just as easily used any of the other standard semantic accounts of conditionals in the literature.

The syntax for our formal language is defined recursively as follows:

$$A ::= P(y_1^{k_1}, \ldots, y_n^{k_n}) \mid (y_1^{k_1} = y_2^{k_2}) \mid \neg A \mid (A \wedge A) \mid \Box A \mid (A > A) \mid (A \to A) \mid \forall x\, A$$

where $x, y_1, \ldots, y_n$ are variables and $k_1, \ldots, k_n \in \mathbb{N}$ are counterpart indices. We define $\vee$, $\supset$, $\equiv$, $\Diamond$, and $\exists$ in the usual way. For simplicity, we will treat names as unbound variables,

though adding constants does not significantly change the formal details below.

In what follows, if $f : \mathbb{N} \to X$ is a function, we may write "$f(k)$" as "f_k".

Definition A.1 *(Structure).* A ***structure*** is a tuple $\mathcal{S} = \langle W, D, d \rangle$ where W is a nonempty set (of worlds), D is a nonempty set (of objects) disjoint from W, and $d : D \to W$ is a total function from objects to worlds. We will define $D_w \coloneqq \{\alpha \in D \mid d(\alpha) = w\}$.

Definition A.2 *(Counterpart Relation).* A ***counterpart relation on $\mathcal{S}$*** is a binary relation $C \subseteq (D \times W)^2$ such that for all $\alpha \in D$, and all $w, v \in W$, there is a $\beta \in D$ such that $\langle \alpha, w \rangle \, C \, \langle \beta, v \rangle$. We will define $C^v(\alpha, w) \coloneqq \{\beta \in D \mid \langle \alpha, w \rangle \, C \, \langle \beta, v \rangle\}$.

Definition A.3 *(Counterpart Structure).* A ***counterpart structure*** is a tuple $\mathcal{C} = \langle \mathcal{S}, C \rangle$ where $\mathcal{S}$ is a structure and C maps every $k \in \mathbb{N}$ to a counterpart relation C_k on $\mathcal{S}$.

In Definition A, we require that every object-world pair $\langle \alpha, w \rangle$ have a counterpart at any world. This is for technical simplicity, so that we do not have to decide now how to handle truth-value gaps. One should bear in mind that this does *not* say that every object necessarily exists, since an object's counterpart at a world may not exist at that world.

Definition A.4 *(Variable Assignment).* Let $\mathcal{C}$ be a counterpart structure and let VAR be the set of variables. A ***variable assignment on $\mathcal{C}$*** is a map $g : \mathbb{N} \to (\mathsf{VAR} \to D)$. If $\alpha \in D$, define g^x_α as the result of reassigning $g_k(x) = \alpha$ for each k. We define $\mathsf{VA}(\mathcal{C})$ as the set of all variable assignments on $\mathcal{C}$.

Definition A.5 *(Possibility).* Let $\mathcal{C}$ be a counterpart structure. A ***possibility in $\mathcal{C}$*** is a pair $\langle w, g \rangle$, where $w \in W$ and $g \in \mathsf{VA}(\mathcal{C})$. We will write "$\langle w, g \rangle$" as "$w_g$". We define $\mathsf{Pos}^{\mathcal{C}}$ as the set of possibilities in $\mathcal{C}$. A possibility w_g is ***centered*** if for all $k, n \in \mathbb{N}$ and all $x \in \mathsf{VAR}$, $g_k(x) = g_n(x)$. We will let $\mathsf{CenPos}^{\mathcal{C}}$ be the set of centered possibilities in $\mathcal{C}$. Define $\twoheadrightarrow^{\mathcal{C}} \subseteq (\mathsf{Pos}^{\mathcal{C}})^2$ so that $w_g \twoheadrightarrow^{\mathcal{C}} v_h$ iff $\forall k \in \mathbb{N} \, \forall x \in \mathsf{VAR}$: $\langle h_k(x), v \rangle \, C_k \, \langle g_k(x), w \rangle$.

Definition A.6 *(Selection Functions).* Let $\mathcal{C}$ be a counterpart structure. A ***selection function over $\mathcal{C}$*** is a function $f : (\wp(\mathsf{Pos}^{\mathcal{C}}) \times \mathsf{Pos}^{\mathcal{C}}) \to \wp(\mathsf{Pos}^{\mathcal{C}})$. A selection function f is ***aligned*** if for any $S \subseteq \mathsf{Pos}^{\mathcal{C}}$ and any $w_g \in \mathsf{Pos}^{\mathcal{C}}$, $f(S, w_g) \subseteq \mathsf{CenPos}^{\mathcal{C}}$.

Definition A.7 *(Models).* A ***model over $\mathcal{C}$*** is a pair $\mathcal{M} = \langle \mathcal{C}, f_>, f_\to, I \rangle$ where $\mathcal{C}$ is a counterpart structure, $f_>$ is a selection function over $\mathcal{C}$, $f_\to$ is an aligned selection function over $\mathcal{C}$, and $I(P, w) \subseteq D^n$ for each $w \in W$ and each n-place predicate P.

Of course, as it stands, we would need to constrain the selection functions more than we have for $>$ and $\rightarrow$ to be more realistic. But we do not need to get into these issues here; even the bare minimal semantics sketched below suffices for our purposes.

Definition A.8 *(Satisfaction).* Let $\mathcal{M}$ be a model and let $w_g \in \mathsf{Pos}^{\mathcal{C}}$. We define the ***satisfaction relation*** $\Vdash$ recursively as follows:

$$
\begin{array}{lcl}
\mathcal{M}, w_g \Vdash P(y_1^{k_1}, \ldots, y_n^{k_n}) & \Leftrightarrow & \langle g_{k_1}(y_1), \ldots, g_{k_n}(y_n)\rangle \in I(P, w) \\
\mathcal{M}, w_g \Vdash y_1^{k_1} = y_2^{k_2} & \Leftrightarrow & g_{k_1}(y_1) = g_{k_2}(y_2) \\
\mathcal{M}, w_g \Vdash \neg A & \Leftrightarrow & \mathcal{M}, w_g \nVdash A \\
\mathcal{M}, w_g \Vdash A \wedge B & \Leftrightarrow & \mathcal{M}, w_g \Vdash A \text{ and } \mathcal{M}, w_g \Vdash B \\
\mathcal{M}, w_g \Vdash \Box A & \Leftrightarrow & \forall v_h \in \mathsf{Pos}^{\mathcal{C}}\colon\ w_g \twoheadrightarrow v_h \ \Rightarrow\ \mathcal{M}, v_h \Vdash A \\
\mathcal{M}, w_g \Vdash A > C & \Leftrightarrow & \forall v_h \in f_{>}(|A|^{\mathcal{M}, w_g}, w_g)\colon\ \mathcal{M}, v_h \Vdash C \\
\mathcal{M}, w_g \Vdash A \rightarrow C & \Leftrightarrow & \forall v_h \in f_{\rightarrow}(||A||^{\mathcal{M}, w_g}, w_g)\colon\ \mathcal{M}, v_h \Vdash C \\
\mathcal{M}, w_g \Vdash \forall x\, A & \Leftrightarrow & \forall \alpha \in D_v\colon\ \mathcal{M}, w_{g^x_\alpha} \Vdash A
\end{array}
$$

where:

$$
\begin{aligned}
|A|^{\mathcal{M}, w_g} &:= \{v_h \in \mathsf{Pos}^{\mathcal{C}} \mid w_g \twoheadrightarrow v_h \text{ and } \mathcal{M}, v_h \Vdash A\} \\
||A||^{\mathcal{M}, w_g} &:= |A|^{\mathcal{M}, w_g} \cap \mathsf{CenPos}^{\mathcal{C}}.
\end{aligned}
$$

Definition A.9 *(Consequence).* We will say Γ ***entails*** A, $\Gamma \models A$, if for all models $\mathcal{M}$ and for all $w_g \in \mathsf{CenPos}^{\mathcal{C}}$, if $\mathcal{M}, w_g \Vdash B$ for all $B \in \Gamma$, then $\mathcal{M}, w_g \Vdash A$. We will say A is ***valid***, $\models A$, if $\varnothing \models A$.

One can verify that $\models (a^k = b^n \equiv a^k = b^k)$ and $\models a^k = a^n$, while $\not\models \Box(a^k = b^n \equiv a^k = b^k)$ and $\not\models \Box(a^k = a^n)$ (though $\models \Box(a^k = a^k)$). Notice also that if $\models A \equiv A'$ and $\models C \equiv C'$, then $\models (A \rightarrow C) \equiv (A' \rightarrow C')$. Thus, $\models ((a^k = b^n) \rightarrow C) \equiv ((a^k = b^k) \rightarrow C)$, and $\models (a^k \neq a^n) \rightarrow C$. (Vacuity is the Lewis-Stalnaker simulacrum for semantic defectiveness.) For counterfactuals, we can only infer $\models (A > C) \equiv (A' > C')$ if $\models \Box(A \equiv A')$ and $\models \Box(C \equiv C')$, assuming $f_{>}(S, w_g) \subseteq S$ for any $S \subseteq \mathsf{Pos}^{\mathcal{C}}$. Indeed, $\not\models ((a^k = b^n) > C) \equiv ((a^k = b^k) > C)$ and $\not\models (a^k \neq a^n) > C$, even though $\Box \neg A \models A > C$.

Note that our semantics technically validates $\Box \neg A \models A \rightarrow C$. This is because we assumed for simplicity that $\Box$ and $\rightarrow$ quantify over the same possibilities. If this is undesirable, it could easily be fixed by separating $\mathsf{Pos}^{\mathcal{C}}$ into metaphysical and epistemic possibilities in our models. Then we could restrict $\Box$ and $>$ to only quantify over the metaphysical possibilities while restricting $\rightarrow$ to only quantify over the epistemic possibilities.

References

Aloni, Maria. 2005. "Individual Concepts in Modal Predicate Logic." *Journal of Philosophical Logic* 34:1–64.

Arregui, Ana. 2007. "Being Me, Being You: Pronoun Puzzles in Modal Contexts." In E Puig-Waldmüller (ed.), *Proceedings of Sinn und Bedeutung 11*, 31–45. Barcelona: Universitat Pompeu Fabra.

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

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.

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

Burge, Tyler. 1973. "Reference and Proper Names." *The Journal of Philosophy* 70:425–439.

Chierchia, Gennaro. 1999. "Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface." In A Belletti (ed.), *Structures and Beyond: The Cartography of Syntactic Structure*, 39–103. Oxford: Oxford University Press.

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.

Crossley, John N and Humberstone, Lloyd. 1977. "The Logic of 'Actually'." *Reports on Mathematical Logic* 8:11–29.

Cumming, Samuel. 2008. "Variabilism." *The Philosophical Review* 117:525–554.

Dasgupta, Shamik. 2009. "Individuals: An Essay in Revisionary Metaphysics." *Philosophical Studies* 145:35–67.

Davies, Martin and Humberstone, Lloyd. 1980. "Two Notions of Necessity." *Philosophical Studies* 38:1–30.

Emery, Nina and Hill, Christopher S. 2016. "Impossible Worlds and Metaphysical Explanation: Comments on Kment's Modality and Explanatory Reasoning." *Analysis* 77:134–148.

Fara, Delia Graff. 2015. "Names Are Predicates." *The Philosophical Review* 124:59–117.

Fox, Danny. 2007. "Free Choice and the Theory of Scalar Implicatures." In Uli Sauerland and Penka Stateva (eds.), *Presupposition and Implicature in Compositional Semantics*, 71–120. London: Palgrave Macmillan.

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

Gazdar, Gerald. 1979. *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Gibbard, Allan. 1975. "Contingent Identity." *Journal of Philosophical Logic* 4:187–221.

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

Glanzberg, Michael. 2007. "Metaphor and Lexical Semantics." *Baltic International Yearbook of Cognition, Logic and Communication* 3:1–47.

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

Hazen, Allen P. 1979. "Counterpart-Theoretic Semantics for Modal Logic." *The Journal of Philosophy* 76:319–338.

Higgins, Roger Francis. 1979. *The Pseudo-Cleft Construction in English*. New York: Routledge.

Holliday, Wesley H and Perry, John. 2014. "Roles, Rigidity, and Quantification in Epistemic Logic." In Alexandru Baltag and Sonja Smets (eds.), *Johan van Benthem on Logic and Information Dynamics*, 591–629. Cham: Springer International Publishing.

Jeffrey, Richard and Edgington, Dorothy. 1991. "Matter-of-Fact Conditionals." *Aristotelian Society Supplementary Volume* 65:161–210.

Kaplan, David. 1977. "Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and Other Indexicals." In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, 481–563. Oxford: Oxford University Press.

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

Kracht, Marcus and Kutz, Oliver. 2001. "The Semantics of Modal Predicate Logic I: Counterpart-Frames." In Frank Wolter, Heinrich Wansing, Maarten de Rijke, and Michael Zakharyaschev (eds.), *Advances in Modal Logic*, 299–320. Singapore: CSLI Publications.

—. 2005. "The Semantics of Modal Predicate Logic II. Modal Individuals Revisited." 60–97. Wellesley, MA: AK Peters.

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.

Kripke, Saul A. 1971. "Identity and Necessity." In Milton Karl Munitz (ed.), *Identity and Individuation*, 135–164. New York: New York University Press.

—. 1980. *Naming and Necessity*. Harvard University Press.

Lakoff, George. 1970. "Linguistics and Natural Logic." *Synthese* 22:151–271.

—. 1996. "Sorry, I'm Not Myself Today: The Metaphor System for Conceptualizing the Self." In Gilles Fauconnier and Eve Sweetser (eds.), *Spaces, Worlds, and Grammar*, 91–123. Chicago: University of Chicago Press.

Lewis, David K. 1968. "Counterpart Theory and Quantified Modal Logic." *The Journal of Philosophy* 65:113–126.

—. 1971. "Counterparts of Persons and Their Bodies." *The Journal of Philosophy* 68:203–211.

—. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

—. 1979. "Attitudes De Dicto and De Se." *The Philosophical Review* 88:513–543.

—. 1986. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

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.

Mikkelsen, Line. 2004. *Specifying Who: On the Structure, Meaning, and Use of Specificational Copular Clauses*. Ph.D. thesis, University of California, Santa Cruz.

—. 2005. "Copular Clauses." In Claudia Maienborn, Klaus von Heusinger, and Paul Portner (eds.), *Semantics An International Handbook of Natural Language Meaning*, 1805–1829. Berlin: Mouton de Gruyter.

Ninan, Dilip. 2012. "Counterfactual Attitudes and Multi-Centered Worlds." *Semantics and Pragmatics* 5:1–57.

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

Percus, Orin and Sauerland, Uli. 2003. "Pronoun Movement in Dream Reports." In Makoto Kadowaki and Shigeto Kawahara (eds.), *Proceedings of NELS 33*. University of Massachusetts, Amherst.

Pollock, John L. 1976. *Subjunctive Reasoning*. Dordrecht: Reidel.

Reboul, Anne. 1996. "If I Were You, I Wouldn't Trust Myself: Indexicals, Ambiguity and Counterfactuals." *Acts of the 2nd International Colloquium on Deixis "Time, Space and Identity"* 151–175.

Rieppel, Michael. 2013. "The Double Life of 'The Mayor of Oakland'." *Linguistics and Philosophy* 36:417–446.

Rothstein, Susan. 1995. "Small Clauses and Copular Constructions." In Anna Cardinaletti and Maria T Guasti (eds.), *Small Clauses*, 27–48. New York: Academic Press.

Russell, Jeffrey Sanford. 2013. "Possible Worlds and the Objective World." *Philosophy and Phenomenological Research* 90:389–422.

Sæbø, Kjell Johan. 2015. "Lessons from Descriptive Indexicals." *Mind* 124:1111–1161.

Schwarz, Wolfgang. 2013. "Contingent Identity." *Philosophy Compass* 8:486–495.

—. 2014. "Counterpart Theory and the Paradox of Occasional Identity." *Mind* 123:1057–1094.

Shumener, Erica. 2017. "The Metaphysics of Identity: Is Identity Fundamental?" *Philosophy Compass* 12:e12397–13.

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

—. 1976. "Indicative conditionals." In William L Harper, Robert C Stalnaker, and Glenn Pearce (eds.), *Ifs*, 193–210. Dordrecht: Springer Netherlands.

—. 1986. "Counterparts and Identity." *Midwest Studies in Philosophy* 11:121–140.

—. 2008. *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

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

van Rooij, Robert. 2006. *Attitudes and Changing Contexts*. Ph.D. thesis, University of Stuttgart.

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

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

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. 1983. "Epistemic Conditionals." *Pacific Philosophical Quarterly* 64:249–265.

Willer, Malte. 2015. "Simplifying Counterfactuals." In Thomas Brochhagen, Floris Roelofsen, and Nadine Theiler (eds.), *Proceedings of the 20th Amsterdam Colloquium*, 428–437. Amsterdam: ILLC.

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

—. 2016. "Counterpossibles in Metaphysics." In Brad Armour-Garb and Fred Kroon (eds.), *Philosophical Fictionalism*. Oxford: Oxford University Press.