Counterlogicals as Counterconventionals*

Alexander W. Kocurek and Ethan Jerzak
Forthcoming in *Journal of Philosophical Logic*

Abstract

We develop and defend a new approach to counterlogicals. Non-vacuous counterlogicals, we argue, fall within a broader class of counterfactuals known as *counterconventionals*. Existing semantics for counterconventionals (developed by Einheuser (2006) and Kocurek et al. (2020)) allow counterfactuals to shift the interpretation of predicates and relations. We extend these theories to counterlogicals by allowing counterfactuals to shift the interpretation of logical vocabulary. This yields an elegant semantics for counterlogicals that avoids problems with the usual impossible worlds semantics. We conclude by showing how this approach can be extended to counterpossibles more generally.

1 The Difficult Question

Cohen (1990, p. 131) asked a difficult question:

(1) If the failure of modus ponens implied the failure of modus tollens, and modus ponens were to fail, would modus tollens fail too?

It seems we could go either way. On the one hand, modus ponens is exactly the form of reasoning whose validity is being denied in the antecedent of (1). So to answer affirmatively seems to require reasoning with the very inference rule we are supposing to be invalid. On the other hand, modus ponens is *actually* valid. So a negative answer to (1) would constitute an actual counterexample to a seemingly valid form of counterfactual reasoning, viz., the schema $((\phi \rightarrow \psi) \wedge \phi) \boxdot\!\!\rightarrow \psi$.

The question is hard because the antecedent of this counterfactual is, in a broad sense, logically impossible. Such counterfactuals are known as ***counterlogicals***, and they are the topic of this paper. Some other examples:

*We are grateful to Bob Beddor, Kelly Gaus, Jens Kipper, Daniel Nolan, Dave Ripley, Rachel Rudolph, Alex Sandgren, Zeynep Soysal, James Walsh, and two anonymous referees for their helpful comments. This paper was presented at the Richard Wollheim Society (2018), the Melbourne Logic Seminar (2018), the Central APA (2019), the Cornell Workshop in Linguistics & Philosophy (2019), the Australian National University (2020), the Faculty Reading Group at the National University of Singapore (2019), and Zeynep Soysal's hyperintensionality seminar at the University of Rochester (2020). We are grateful to the audience members of all these venues for their feedback.

(2) If France were both a monarchy and not a monarchy, the revolutionists would be very confused.

(3) If the Liar sentence were both true and not true, the moon would be made of green cheese.

(4) If Aristotle were not self-identical, his argument for the law of non-contradiction in *Metaphysics Gamma* would have failed.

Our question is this: Do counterlogicals like these have non-trivial truth conditions? And if so, what are they? In other words, what is the correct semantics for counterlogicals?

There is disagreement in the counterfactuals literature on this question. On the one hand, examples such as (1)–(4) suggest that the truth conditions of counterlogicals are not trivial: some seem true, others false, and for others it is hard to say. That suggests that they do not all mean the same thing. On the other hand, it is unclear how to reason under a logically impossible counterfactual supposition without leading to a trivializing explosion. As Lewis (1973, p. 24) famously put it, "Confronted by an antecedent that is not really an entertainable supposition, one may react by saying, with a shrug: If that were so, anything you like would be true!"

We will develop and defend a middle ground. Counterlogicals have non-trivial truth conditions, but only insofar as they are interpreted as *counterconventionals*—insofar, that is, as their antecedents trigger a shift in the meanings of the logical vocabulary in the consequent. So when a counterlogical receives a non-trivial reading, it is because words like 'not', 'or', and so on get reinterpreted according to another logic. As a result, its antecedent ends up expressing a completely coherent proposition. Without such a convention shift, counterlogicals are vacuous. So there is a kernel of truth on both sides of existing debate—a kernel which we preserve while discarding the concomitant husks.

Here is an outline of what is to come. We start by reviewing the standard impossible worlds semantics for counterlogicals, where counterfactuals range over logically impossible worlds (§2). We observe that this semantics is prone to a Quinean "changing the subject" objection: seemingly non-vacuous counterlogicals really just involve changing the meaning of the logical vocabulary (§3). In response, we propose to treat non-vacuous counterlogicals as counterconventionals, which shift the conventions used to interpret their constituents. This phenomenon also occurs in non-logical examples (§4).

Our proposal is motivated by the view—inspired by Carnap (1937) and Ayer (1952), and recently defended by Kouri Kissel (2018, 2019)—that disputes over logic are metalinguistic negotiations (§5). We call this view ***logical expressivism***. It holds that claims about logical validity are expressions of the logical commitments of the speaker, rather than factual assertions about the world. In this spirit, we extend the expressivist semantics for counterconventionals developed by Einheuser (2006) and

Kocurek et al. (2020) to allow counterfactuals to shift the interpretation of logical vocabulary (§6). This yields an elegant, unified, and well-behaved semantics for counterlogicals that does not fall prey to the Quinean objection. We conclude by suggesting that this view of counterlogicals extends to counterpossibles more generally (§7). Some technical results are proved in a brief appendix (§A).

A quick note before we get started. We assume throughout that our background logic is classical. Thus, as a first pass, a counterfactual is a counterlogical iff the negation of its antecedent is classically valid. This is largely for ease of exposition; our framework is easily adaptable to non-classical background logics (see §6).

2 Impossible Worlds Semantics

Counterlogicals are often considered a subspecies of ***counterpossibles***, i.e., counterfactuals with impossible antecedents.[1] Here are some purported counterpossibles that are not counterlogicals:

(5) If water were hydrogen peroxide, life would not exist.

(6) If Sam were a salmon, she would have wings.

(7) If Socrates had existed without his singleton, he would be sad.

There are two main views about the truth conditions of counterpossibles. According to ***vacuism***, counterpossibles all have vacuous truth conditions.[2] Vacuism is predicted by the standard Lewis-Stalnaker semantics for counterfactuals. On this semantics, a counterfactual of the form $\phi \boxrightarrow \psi$ is true iff all the closest possible worlds where ϕ is true are worlds where ψ is true. So if there are no possible worlds where ϕ is true, it is trivially the case that "all" of the closest such worlds are worlds where ψ is true. Hence, the Lewis-Stalnaker semantics validates the following principle:

Vacuism. $\neg \Diamond \phi \vDash \phi \boxrightarrow \psi$.

Because it is a consequence of the orthodox theory of counterfactuals, vacuism is considered the orthodox position on counterpossibles.

[1] There are exceptions involving indexical vocabulary such as "actually". On some natural notions of logical consequence, $\phi \equiv @\phi$ is a logical truth while $\Box(\phi \equiv @\phi)$ is not. In that case, counterfactuals of the form $(\phi \wedge \neg @\phi) \boxrightarrow \psi$ would technically be counterlogicals but not counterpossibles. Such counterfactuals do not raise any serious problems for the standard theories of counterfactuals in the literature, though: a simple two-dimensional semantics can easily explain why $(\phi \wedge \neg @\phi) \boxrightarrow \psi$ is not vacuous. For now, we ignore these cases and continue to treat counterlogicals as counterpossibles.

[2] Popper (1959) is perhaps one of the first to defend this view (also in Popper 2005, p. 461), though the view is predominantly associated with Lewis (1973, pp. 24–25). Other contemporary defenders of this view include Lycan (2001); Bennett (2003); Emery and Hill (2016); Williamson (2007, 2017). Lewis (1973, pp. 25–26) notes the possibility of a vacuist view on which all counterpossibles are vacuously false. Vetter (2016) defends an intermediate position, on which counterpossibles are vacuously true unless they are given an epistemic interpretation.

Yet the tides are turning: ***non-vacuism***, the view that some counterpossibles have non-vacuous truth conditions, has gained converts.[3] One of its primary virtues is that it accords with our intuitive judgments. We do not judge each of (5)–(7) the same way, even when we know their antecedents are impossible: (5) seems true, (6) seems false, and (7) could be true or false depending on Socrates's psychology.

Vacuists have responded with a number of pragmatic explanations of these judgments.[4] But there is a relatively straightforward semantic explanation available to non-vacuists. We simply need to enrich the Lewis-Stalnaker semantics with *impossible worlds*, which can be represented as sets of sentences, viz., the sentences that are true at those impossible worlds. On this semantics, $\phi \boxdot\!\!\rightarrow \psi$ is true just in case all the closest worlds where ϕ is true—whether or not those worlds are possible—are worlds where ψ is true. This allows counterpossibles to be non-vacuous while preserving much of the spirit of the orthodox semantics.

Here is how that would go in a bit more detail.[5] The language $\mathcal{L}$ of this semantics can be summarized in Backus-Naur form as follows:

$$\phi ::= p \mid \neg\phi \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid (\phi \rightarrow \phi) \mid \Box\phi \mid \Diamond\phi \mid (\phi \boxdot\!\!\rightarrow \phi).$$

An ***impossible worlds model*** is a tuple of the form $\mathcal{I} = \langle W, P, f, V\rangle$, where W is a non-empty set of *worlds*, $P \subseteq W$ is a non-empty set of *possible worlds* (thus we define $I := (W - P)$ as the set of *impossible worlds*), $f\colon \wp(W) \times W \rightarrow \wp(W)$ is a *selection function*, and V is a *valuation function*, mapping every atomic sentence p and possible world $w \in P$ to either 0 or 1, and every formula ϕ and impossible world $w \in I$ to either 0 or 1. The satisfaction relation $\Vdash_{\text{i}}$ is defined as follows. If $w \in I$, then $\mathcal{I}, w \Vdash_{\text{i}} \phi$ iff $V(\phi, w) = 1$. If $w \in P$, then $\Vdash_{\text{i}}$ is defined recursively:

$$\begin{aligned}
\mathcal{I}, w \Vdash_{\text{i}} p &\quad\Leftrightarrow\quad V(p, w) = 1\\
\mathcal{I}, w \Vdash_{\text{i}} \neg\phi &\quad\Leftrightarrow\quad \mathcal{I}, w \nVdash_{\text{i}} \phi\\
\mathcal{I}, w \Vdash_{\text{i}} \phi \wedge \psi &\quad\Leftrightarrow\quad \mathcal{I}, w \Vdash_{\text{i}} \phi \text{ and } \mathcal{I}, w \Vdash_{\text{i}} \psi\\
\mathcal{I}, w \Vdash_{\text{i}} \phi \vee \psi &\quad\Leftrightarrow\quad \mathcal{I}, w \Vdash_{\text{i}} \phi \text{ or } \mathcal{I}, w \Vdash_{\text{i}} \psi\\
\mathcal{I}, w \Vdash_{\text{i}} \phi \rightarrow \psi &\quad\Leftrightarrow\quad \mathcal{I}, w \Vdash_{\text{i}} \phi \text{ only if } \mathcal{I}, w \Vdash_{\text{i}} \psi\\
\mathcal{I}, w \Vdash_{\text{i}} \Box\phi &\quad\Leftrightarrow\quad \text{for all } v \in P\text{: } \mathcal{I}, v \Vdash_{\text{i}} \phi\\
\mathcal{I}, w \Vdash_{\text{i}} \Diamond\phi &\quad\Leftrightarrow\quad \text{for some } v \in P\text{: } \mathcal{I}, v \Vdash_{\text{i}} \phi\\
\mathcal{I}, w \Vdash_{\text{i}} \phi \boxdot\!\!\rightarrow \psi &\quad\Leftrightarrow\quad \text{for all } v \in f([\![\phi]\!]^{\mathcal{I}}, w)\text{: } \mathcal{I}, v \Vdash_{\text{i}} \psi,
\end{aligned}$$

where $[\![\phi]\!]^{\mathcal{I}} := \{w \in W \mid \mathcal{I}, w \Vdash_{\text{i}} \phi\}$. Consequence is the preservation of satisfaction at *possible* worlds: $\Gamma \models_{\text{i}} \phi$ iff for every impossible worlds model $\mathcal{I} = \langle W, P, f, V\rangle$

[3] Downing (1959) is an early critic of vacuism. Other non-vacuists include Cohen (1987, 1990); Mares (1997); Nolan (1997); Goodman (2004); Vander Laan (2004); Kim and Maslen (2006); Krakauer (2012); Brogaard and Salerno (2013); Kment (2014); Bernstein (2016); Jenny (2016); Berto et al. (2018).

[4] See Emery and Hill 2016; Vetter 2016; Williamson 2017.

[5] Our presentation of this semantics roughly follows that of Berto et al. 2018 with some minor changes. See also French et al. 2020.

and every $w \in P$, if $\mathcal{I}, w \Vdash_{\text{¡}} \Gamma$, then $\mathcal{I}, w \Vdash_{\text{¡}} \phi$. This guarantees that $\models_{\text{¡}}$ obeys all of the laws of classical logic (in fact, **S5**).

The non-vacuist may want to place constraints on f to make the truth conditions for counterfactuals more realistic. Some common constraints include:

Success. $f(X, w) \subseteq X$.

Weak Centering. If $w \in X$, then $w \in f(X, w)$.

Strangeness of Impossibility. If $X \cap P \neq \varnothing$, then $f(X, w) \subseteq P$.

These constraints correspond respectively to the following principles:[6]

Identity. $\models \phi \boxrightarrow \phi$.

Modus Ponens. $\phi \boxrightarrow \psi, \phi \models \psi$.

Necessity Condition. $\Diamond \phi, \Box \psi \models \phi \boxrightarrow \psi$.

Though some of these principles may seem plausible, they are contested even amongst non-vacuists. So we will not impose them for the sake of neutrality.[7] By contrast, **Vacuism** corresponds to a relatively strong requirement, viz., that impossible worlds are never close:

No Close Impossibilities. If $X \cap P = \varnothing$, then $f(X, w) = \varnothing$.

Thus, the standard vacuist semantics can be viewed as a special case of the more general impossible worlds semantics.

There is certainly more to say about this debate. But we find the semantic explanation provided by non-vacuists to be simple, elegant, and plausible. So for now, we will table this debate and simply assume non-vacuism is true (though we return to this issue in §7).

Even amongst non-vacuists, there is disagreement over counterlogicals. As with counterpossibles, there seem to be two main views, though we do not know of an established name for either. According to what we will call ***counterlogical vacuism***, counterlogicals all have vacuous truth conditions, even though counterpossibles generally do not.[8] According to what we will call ***counterlogical non-vacuism***, counterlogicals, like other counterpossibles, do not all have vacuous truth conditions.[9]

[6] See French et al. 2020 for details. While they assume impossible worlds are logically consistent and maximal, at least these correspondence results carry over to our more general setting.

[7] For instance, Nolan (1997, p. 555) suggests that 'if nothing were true, nothing would be true' is a potential counterexample to $\models \phi \boxrightarrow \phi$. Moreover, there are well-known counterexamples to modus ponens independently of counterpossibles; see, e.g., Briggs 2012.

[8] Defenders of counterlogical vacuism include Downing (1959); Goodman (2004); Kment (2014).

[9] Defenders of counterlogical non-vacuism include Cohen (1987, 1990); Mares (1997); Nolan (1997); Vander Laan (2004); Kim and Maslen (2006); Krakauer (2012); Brogaard and Salerno (2013); Berto et al. (2018).

Counterlogical vacuists will need to modify the impossible worlds semantics so as to rule out logically impossible worlds from the models without ruling out all impossible worlds. To do this, they will restrict the class of models to those where V determines the truth value of ϕ at impossible worlds in a logically coherent manner. There are many ways to do this,[10] but however they do it, they will want the following principle to be validated (without validating **Vacuism**):

Counterlogical Vacuism. If $\models \neg\, \phi$, then $\models \phi \mathrel{\Box\!\!\rightarrow} \psi$.

Counterlogical non-vacuists may accept the impossible worlds semantics as it stands. Since truth at impossible worlds is determined solely by the valuation function V, and since there are no constraints on how V may assign truth values to formulas at impossible worlds, impossible worlds models can contain worlds that are logically impossible. So the impossible worlds semantics already accommodates non-vacuous counterlogicals (though perhaps there are other modifications they would make).[11]

At first glance, our intuitive judgments seem to support counterlogical non-vacuism. We do not judge each of (2)–(4) the same way: (2) seems true, (3) seems false, and (4) seems to depend on the correct interpretation of Aristotle's *Metaphysics Gamma*. Again, we might try to explain these judgments via pragmatic mechanisms. But the impossible worlds semantics already provides a simple, elegant, and plausible explanation for the apparent non-vacuity of counterlogicals. If we are already willing to tolerate non-vacuous counterpossibles, then as a default, we ought to tolerate non-vacuous counterlogicals as well unless there were a strong reason not to.

3 Changing the Subject

Yet there is a strong reason to think counterlogicals are semantically vacuous even if other counterpossibles are not.

Consider the following example:

(8) If intuitionistic logic were the correct logic, then either the continuum hypothesis would be true or it would not be true.

[10] Here is one suggestion. Where $\mathcal{I} = \langle W, P, f, V\rangle$, say $w \in W$ is ***coherent in*** $\mathcal{I}$ if (i) $\{\phi \in \mathcal{L} \mid \mathcal{I}, w \Vdash_{\mathrm{i}} \phi\} \not\models_{\mathrm{i}} \bot$, (ii) for each $\phi \in \mathcal{L}$, either $\mathcal{I}, w \Vdash_{\mathrm{i}} \phi$ or $\mathcal{I}, w \Vdash_{\mathrm{i}} \neg\, \phi$, and (iii) $\mathcal{I}, w \Vdash_{\mathrm{i}} \phi \mathrel{\Box\!\!\rightarrow} \psi$ iff $f(\llbracket \phi \rrbracket^{\mathcal{I}}, w) \subseteq \llbracket \psi \rrbracket^{\mathcal{I}}$ for all $\phi, \psi \in \mathcal{L}$. Say $\mathcal{I}$ itself is ***coherent*** if w is coherent in $\mathcal{I}$ for all $w \in I$. Intuitively, coherent models are ones where every impossible world is the possible world of some model. We can then restrict consequence to the class of coherent models satisfying **Success**.

[11] Some counterlogical non-vacuists may want to impose a variant of **Strangeness of Impossibility**:

Strangeness of Inconsistency. If $\not\models \neg\, \phi$ and $\psi_1, \ldots, \psi_n \models \chi$, then $\phi \mathrel{\Box\!\!\rightarrow} \psi_1, \ldots, \phi \mathrel{\Box\!\!\rightarrow} \psi_n \models \phi \mathrel{\Box\!\!\rightarrow} \chi$.

In terms of closeness of worlds, this says that every logically consistent world is closer than every logically inconsistent world (though only roughly; more precisely, this says that if every closest ϕ-world satisfies the premises of a valid argument, where ϕ is logically consistent, then every closest ϕ-world satisfies its conclusion). We will not take a stand on **Strangeness of Inconsistency** here.

Presumably, the counterlogical non-vacuist will maintain that (8) is a false counterlogical if there ever was one.[12] After all, intuitionistic logic rejects the law of excluded middle. Since the continuum hypothesis is undecidable, it is a counterexample to the law of excluded middle in intuitionistic set theory.

But one might object that this temptation to call (8) false reflects a use-mention error. Recall a familiar point due to Kripke (1971, 1980): when we evaluate truth at a counterfactual scenario, we use *our* language, not the language adopted by speakers in that scenario. Thus, a world where speakers use the names 'Hesperus' and 'Phosphorus' to refer to distinct objects is not a world where Hesperus is Phosphorus; it is a world where people speak differently. So while (9a) is false, (9b) is true:

(9) a. If people had used 'Hesperus' and 'Phosphorus' to refer to different objects, the sentence 'Hesperus is Phosphorus' would be true.
b. If people had used 'Hesperus' and 'Phosphorus' to refer to different objects, Hesperus would be Phosphorus.

Since we are *using* 'Hesperus' and 'Phosphorus' in the consequent of (9b), those terms refer to what *we actually* refer to by them, viz., Venus. And since Venus would still be self-identical if people had spoken differently, (9b) is true.

Similarly, in (8), we are *using* the words 'not' and 'or' in the consequent, not *mentioning* them. True, in a world governed by intuitionistic logic, people would use the words 'not' and 'or' in such a way so as not to validate the law of excluded middle. That is, the following counterfactual might be false:

(10) If intuitionistic logic were the correct logic, then the sentence 'Either the continuum hypothesis is true or it is not true' would be true.

That is irrelevant, however: when we describe what is true at a counterfactual scenario, we describe it in *our* language, not the language of these hypothetical speakers. And since (let's suppose) we use 'not' and 'or' in a way that is governed by the law of excluded middle, a world where intuitionistic logic is correct is still a world where either the continuum hypothesis is true or it is not true. So unless we are changing the meaning of 'not' and 'or' in (8), it is true.

The same goes for any apparently non-vacuous counterlogical. Consider (3):

[12] We may want to distinguish counterfactuals such as (2) and (3), whose antecedents have a logically inconsistent form ($\phi \wedge \neg \phi$), from counterfactuals such as (1) and (8), whose antecedents do not have a logically inconsistent form but express a *metalogical* falsehood, e.g., about the laws of logic. The latter class of counterfactuals might be more accurately labeled ***countermetalogicals***. For discussion of this distinction, see Williamson 2017; Sandgren and Tanaka 2019. For ease of exposition, we continue to talk about countermetalogicals as though they are counterlogicals. As it turns out, countermetalogicals may be counterlogicals under certain formulations, e.g., with propositional quantifiers. For instance, if we formalized the law of excluded middle as $\forall p \, \Box(p \vee \neg p)$, then $\neg \forall p \, \Box(p \vee \neg p) \, \Box\!\!\rightarrow q$ would be a counterlogical, since $\forall p \, \Box(p \vee \neg p)$ is valid on the standard semantics for modal logic with propositional quantifiers (Fine, 1970).

(3) If the Liar sentence were both true and not true, the moon would be made of green cheese.

At first, this seems false, since a world where the Liar sentence is both true and not true is a world where the principle of explosion fails, and so not everything would be true. But what kind of world are we describing here? After all, we are not mentioning 'not' and 'and' in (3); we are using them. Following Kripke, then, the antecedent of (3) is interpreted according to how *we* use the terms 'not' and 'and'. If we interpret those words classically, the antecedent of (3) must be interpreted classically. And on that interpretation, the world being described by the antecedent makes no sense, just as it makes no sense to say that (literally, strictly speaking, in one and the same sense) that France is a monarchy and France is not a monarchy. It is literally *incoherent*: we cannot comprehend what such an absurd world is like. So either 'not' and 'and' do not have their classical meanings in (3)—in which case, its antecedent is not really logically impossible, in the classical sense—or (3) is vacuously true since contradictions imply everything.

We can bring out the objection by making explicit that we are interpreting the connectives in the antecedent classically. Contrast (3) with:

(11) If the Liar sentence were both true and not true according to a classical interpretation of negation and conjunction, the moon would be made of green cheese.

Unlike (3), we find it hard to deny (11). It is just part of the meaning of classical negation and conjunction that everything follows from a conjunction of a sentence and its negation. To deny (11) is not to deny classical logic but to simply fail to understand what classical logic *is*. If that is right, then the antecedents of counterlogicals such as (3) are not describing scenarios where contradictions are true under a classical interpretation.[13]

This argument seems specific to counterlogicals: it does not obviously generalize to all counterpossibles. Consider, for example:

(12) If water were hydrogen peroxide, life would go on as usual.

(12) seems false since hydrogen peroxide is toxic. Unlike with counterlogicals, the false reading of (12) does not seem to rely on a shift in the meaning of 'water', 'hydrogen peroxide', or 'life'—using those words as *we actually* use them, (12) still seems false. So this argument may give us a reason to think counterlogicals really are vacuous without thereby giving us a reason to think all counterpossibles are (though we think it still might; see §7).

This objection is reminiscent of the "changing the subject" objection against non-classical logics due to Quine (1970). Quine held that whenever the non-classical

[13]This distinction is similar to the distinction due to Sandgren and Tanaka (2019) between logically different worlds and worlds containing logical violations. We return to this point in §6.

logician denies some law of classical logic, they end up talking about different logical operations than the classical logician is talking about. To demonstrate this, Quine asks us to consider a fictitious city—call it "Deviantsville"—whose denizens speak a language like English except their use of 'and' and 'or' is non-standard. According to the denizens of Deviantsville:

$$\begin{aligned} (\phi \text{ and } \psi) &\nvdash_D \phi & \phi &\vdash_D (\phi \text{ and } \psi) \\ (\phi \text{ and } \psi) &\nvdash_D \psi & \psi &\vdash_D (\phi \text{ and } \psi) \\ (\phi \text{ or } \psi) &\vdash_D \phi & \phi &\nvdash_D (\phi \text{ or } \psi) \\ (\phi \text{ or } \psi) &\vdash_D \psi & \psi &\nvdash_D (\phi \text{ or } \psi). \end{aligned}$$

Clearly what is happening here is that denizens of Deviantsville are using the word 'and' to mean disjunction (i.e., what we mean by 'or') and 'or' to mean conjunction (i.e., what we mean by 'and'). It would be silly to characterize the denizens of Deviantsville as employing a highly non-classical logic in which conjunction elimination fails, where disjunctions entail their disjuncts, and so on.

Quine held that this is essentially what the non-classical logician is doing, albeit in disguise. The intuitionist uses the words 'not' and 'or' much like classical logicians, but those words in the intuitionist's mouth do not mean what classical logicians mean by them. It is just part of what 'not' and 'or' mean that they obey the law of excluded middle. "Here, evidently," Quine says, "is the deviant logician's predicament: when he tries to deny the doctrine he only changes the subject."

The same could be said of the counterlogical non-vacuist: they are just changing the subject. When they point to counterfactuals such as (8) as examples of non-vacuous counterlogicals, they are implicitly reinterpreting the logical connectives according to a non-classical logic in order to make sense of the antecedent. But that says nothing about these antecedents on their original, classical interpretation. Holding fixed what *we* mean by the logical connectives, (8) is true. And the same could be said about all counterlogicals. When a counterlogical appears non-vacuous, the speaker is implicitly reinterpreting the logical vocabulary to conform to the logic suggested by the antecedent.

To foreshadow our own approach to counterlogicals, we agree with Quine that the intuitionist and the classicist mean something different when they use 'not' and 'or'. As a consequence, (8) only receives a non-vacuous interpretation when we implicitly reinterpret the logical connectives. That does not mean, however, that the intuitionist "changes the subject" or that counterlogicals are vacuous. Instead, the subject matter of these disputes is precisely over which meanings to adopt and counterlogicals reveal the effects of different conventional choices.

4 Counterconventionals

Over the next few sections, we present our positive proposal of counterlogicals. In short, our proposal is to treat non-vacuous counterlogicals as counterconventionals,

which can shift the conventional interpretation of expressions in their scope.

Counterconventionals arise most naturally in conversational contexts where speakers disagree over how to use specific words or phrases. Such disagreements are known as ***metalinguistic negotiations***.[14] Very roughly, a metalinguistic negotiation is a dispute over *what counts as what* as opposed to *what things are like*.

Here is an example. In 2006, the International Astronomical Union (IAU) decided to redefine the word 'planet' so as to require all planets to "clear their orbital neighborhood" (meaning they had to be significantly more massive than any other object in their orbit). This was because astronomers discovered a number of planet-like objects (Ceres, Eris, Haumea, and Makemake) and were worried that future astronomical discoveries would lead to an explosion in the number of planets. Since Pluto does not clear its orbital neighborhood (its orbit crosses Neptune's), it was reclassified as a dwarf planet. Similarly for these other planet-like objects. This caused an outcry amongst the public. People were outraged that their beloved Pluto was no longer considered a planet. Even amongst the scientific community, there was disagreement over whether this was the right decision.

Now, consider the following dialogue (adapted from Kocurek et al. (2020)):

(13) **Alpha:** Pluto is a planet.

Beta: No it's not. Pluto is not a planet because it does not clear its orbital neighborhood.

Alpha: I don't accept the IAU's definition! Pluto is a planet, I don't care what the IAU says.

Beta: Look, I know that *you* think that Pluto is a planet, but there's a good reason the IAU disagrees. If Pluto were a planet, Ceres, Eris, Haumea, Makemake, and many other objects would be planets, too.

Clearly, Alpha and Beta are not disagreeing over the shape of Pluto's orbit or whether it crosses with Neptune's. Rather, they are disagreeing over how to classify Pluto. Alpha is, in some sense, advocating for classifying Pluto as a planet whereas Beta is advocating for the IAU's classification. So (13) is a metalinguistic negotiation.

Notice that Beta's last response here took the form of a counterfactual:

(14) If Pluto were a planet, Ceres, Eris, Haumea, Makemake, and many other objects would be planets, too.

There are two ways of reading (14). One reading of (14) is something like: if Pluto had cleared its orbital neighborhood, so would Ceres, Eris, and so on. On this reading, (14) is clearly false: a correction to Pluto's orbit so that it clears its orbital neighborhood would not (we may suppose) lead to a similar correction in the orbits of all these other objects.

[14] See Haslanger 2000, 2005; Plunkett and Sundell 2013; Burgess and Plunkett 2013a,b; Plunkett 2015; Hansen 2019.

But intuitively, this is not what Beta was trying to say. Rather, they were saying that if Pluto were *classified* as a planet, Ceres, Eris, etc. would be too. On this reading, (14) is true, and indeed, captures exactly the consideration that led IAU to redefine 'planet' in the first place.

Following Einheuser (2006), we call the first reading of (14) the ***countersubstratum*** reading and the second reading the ***counterconventional*** reading. On the countersubstratum reading of counterfactuals, we hold fixed the conventions used to interpret language and vary reality (the "substratum"), whereas on the counterconventional reading, it is the reverse. So on the countersubstratum reading of (14), we hold fixed the meaning of 'planet' and consider a world where Pluto's orbit is different, whereas on the counterconventional reading, we hold fixed Pluto's orbit but reinterpret 'planet' by dropping the orbital neighborhood condition.

Other examples illustrating the distinction:

(15) a. If Secretariat were an athlete, the world's fastest athletes would all be horses.
b. If pizza were a vegetable, our children would not be any healthier.
c. If *Game of Thrones* were a comedy, it would be quite a dark comedy.

Each of these sentences has a (true) counterconventional reading and a (false) countersubstratum reading. On the former reading, we hold fixed the intrinsic properties of the objects involved (Secretariat, pizza, *Game of Thrones*) and change what is meant by the target word ('athlete', 'vegetable', 'comedy'). On the latter reading, we hold fixed the meanings of the target words but vary the intrinsic properties of the objects involved.

Counterconventionals are what Einheuser calls ***c-monsters***, i.e., convention-shifting expressions.[15] Other examples of c-monsters include tense, attitude verbs, and dependency verbs. To illustrate with the Pluto example:

(16) a. Pluto used to be a planet, but it isn't any more.
b. Alpha thinks Pluto is a planet.
c. Whether or not Pluto is a planet depends on what definition the members of the IAU agree on.

Each of these sentences has a (true) convention-shifting reading and a (false) substratum-shifting reading. On the true, convention-shifting reading, the sentences concern how Pluto is classified; e.g., on this reading, (16a) says that Pluto's classification changed over time. On the false, substratum-shifting reading, the sentences concern Pluto's physical properties; e.g., on this reading, (16a) effectively says that Pluto's orbit changed over time.

There are even expressions like 'according to' or 'in x's sense' that seem to only

[15]This term is inspired by the notion of a ***monster*** from Kaplan 1977. In brief, whereas a monster is context-shifting expression, a c-monster is (in Kaplan's terms) a character-shifting expression.

shift conventions, lacking a substratum-shifting reading altogether:[16]

(17) a. According to Alpha's definition of 'planet', Pluto is a planet.
b. Pluto is a planet in Alpha's sense.

Thus, convention-shifting occurs in a wide variety of embedding expressions, not just counterfactuals.

To make these ideas more concrete, we sketch a semantics for counterconventionals due to Einheuser (2006) and Kocurek et al. (2020). The key idea is to introduce a shiftable convention parameter into indices of evaluation.[17] The convention parameter is essentially an interpretation function, assigning intensions (functions from worlds to extensions) to the non-logical vocabulary of the language. Counterfactuals are c-monsters when they shift this convention parameter.[18]

Let's spell this out in more detail. Where W is a set of worlds, a ***hyperconvention*** over W is a function $c : \mathtt{At} \to \wp(W)$ mapping each atomic sentence to a set of worlds. An ***index*** over W is a pair $\langle w, c\rangle$ where $w \in W$ and c is a hyperconvention over W. We let I_W be the set of all indices over W. A ***conventional model*** is a pair $\mathcal{C} = \langle W, f\rangle$ where $W \neq \varnothing$ is a set of worlds and $f : \wp(I_W) \times I_W \to \wp(I_W)$ is the selection function. As before, we may impose constraints on f if desired. The satisfaction relation $\Vdash_{\mathsf{c}}$ is defined relative to indices, rather than worlds:

$$
\begin{aligned}
\mathcal{C}, w, c \Vdash_{\mathsf{c}} p \quad &\Leftrightarrow \quad w \in c(p) \\
\mathcal{C}, w, c \Vdash_{\mathsf{c}} \neg\phi \quad &\Leftrightarrow \quad \mathcal{C}, w, c \nVdash_{\mathsf{c}} \phi \\
\mathcal{C}, w, c \Vdash_{\mathsf{c}} \phi \wedge \psi \quad &\Leftrightarrow \quad \mathcal{C}, w, c \Vdash_{\mathsf{c}} \phi \text{ and } \mathcal{C}, w, c \Vdash_{\mathsf{c}} \psi
\end{aligned}
$$

[16] Note that these are not equivalent to attitude reports such as (16b); both (17a) and (17b) could be true even if Alpha does not realize it.

[17] This idea is inspired by a similar idea due to Gibbard (2003) for modeling normative discourse and by to Barker (2002) and MacFarlane (2016) for modeling vagueness.

[18] A similar approach would be to use a two-dimensional framework (e.g., Crossley and Humberstone 1977; Stalnaker 1978; Davies and Humberstone 1980) and let the conventions be those used by speakers in the world-as-actual. Counterconventionals could then be modeled as "counteractuals" or "diagonalizing counterfactuals", i.e., counterfactuals that reset the world-as-actual parameter to be the world of evaluation. We do not opt for this approach for two reasons. First, counterconventional interpretations can arise even if there are no speakers in the relevant counterfactual scenario. Thus, this approach would incorrectly predict the following to sound marked:

(i) If Pluto were a planet, then even if there were no people, Ceres, Eris, etc. would be planets, too.

Second, this simple two-dimensionalist approach predicts that counterconventional readings shift the interpretation of 'actually' due to diagonalization. But then the following should sound false on the counterconventional reading, since 'actually' in the consequent refers to the world of evaluation:

(ii) If Pluto were a planet, there would be many more planets in the solar system than there actually are.

For these reasons, we think it's better to not tie this convention parameter to the world-as-actual.

$$
\begin{aligned}
C, w, c \Vdash_c \phi \vee \psi \quad &\Leftrightarrow \quad C, w, c \Vdash_c \phi \text{ or } C, w, c \Vdash_c \psi \\
C, w, c \Vdash_c \phi \rightarrow \psi \quad &\Leftrightarrow \quad C, w, c \Vdash_c \phi \text{ only if } C, w \Vdash_c \psi \\
C, w, c \Vdash_c \Box \phi \quad &\Leftrightarrow \quad \text{for all } v \in W\text{: } C, v, c \Vdash_c \phi \\
C, w, c \Vdash_c \Diamond \phi \quad &\Leftrightarrow \quad \text{for some } v \in W\text{: } C, v, c \Vdash_c \phi \\
C, w, c \Vdash_c \phi \boxrightarrow \psi \quad &\Leftrightarrow \quad \text{for all } \langle v, d \rangle \in f([\![\phi]\!]^C, w, c)\text{: } C, v, d \Vdash_c \psi,
\end{aligned}
$$

where $[\![\phi]\!]^C := \{\langle w, c \rangle \in I_W \mid C, w, c \Vdash_c \phi\}$. Consequence is the preservation of satisfaction over indices: $\Gamma \vDash_c \phi$ iff for every conventional model $C = \langle W, f \rangle$ and every $\langle w, c \rangle \in I_W$, if $C, w, c \Vdash_c \Gamma$, then $C, w, c \Vdash_c \phi$.

Unlike the standard Lewis-Stalnaker semantics or the impossible worlds semantics, counterfactuals can shift the interpretation of atomics in the conventional semantics. We assume, however, that $\Box$ and $\Diamond$ cannot. As a result, the conventional semantics is hyperintensional. In particular, **Vacuism** does not hold: $\neg \Diamond \phi \nvDash_c \phi \boxrightarrow \psi$. However, **Counterlogical Vacuism** does hold (at least given **Success**): if $\vDash_c \neg \phi$, then $[\![\phi]\!]^C = \varnothing$, and so $f([\![\phi]\!]^C, w, c) = \varnothing$, for all C. In that case, $\vDash_c \phi \boxrightarrow \psi$.

Note that counterfactuals can simultaneously shift the world and the hyperconvention. This is by design; some counterfactuals need to shift both. For instance:

(18) If Pluto were a planet, astronomers would be overwhelmed at having to keep track of all the planets.

(18) seems true. But it is not true on a countersubstratum reading, since astronomers would only have to keep track of one more planet if Pluto had cleared its orbital neighborhood. Nor is it true on a "pure" counterconventional reading where the hyperconvention is shifted but the "substratum" is held fixed, since astronomers are not actually overwhelmed.[19] So for (18) to receive a true reading, the counterfactual needs to shift the world and hyperconvention simultaneously.

Which reading (countersubstratum, counterconventional, or something in between) a counterfactual receives is a context-sensitive matter. Specifically, it depends on the features of the context that determine the selection function. Some contexts will prioritize holding fixed our conventions, while others will prioritize holding fixed the substratum. For instance, in a context where speakers are disputing Pluto's orbit, the meaning of 'planet' will likely be held fixed, whereas in a context like that of (13) where speakers are disputing how Pluto is classified, Pluto's physical characteristics will likely be held fixed. This is exactly the kind of context-sensitivity that counterfactuals are widely believed to exhibit.

With that said, we can regiment the difference between countersubstratums and counterconventionals directly in our object language with a few additions. The basic idea is to extend $\mathcal{L}$ to a hybrid logic where there are hybrid operators for each

[19] It is actually hard to find examples of pure counterconventionals in natural language, though there may indeed be some. See Kocurek et al. 2020, pp. 18–19 for discussion.

parameter of the index.[20] Hybrid logic, as it is typically understood, is an extension of basic modal logic to include operators and terms for talking about specific worlds directly in the object language. This involves making three additions to the simple propositional modal language:

(i) new atomic formulas $n_1, n_2, n_3, \ldots$ known as ***nominals***, which act like rigid names for specific worlds (so n stands for "the current world is n");

(ii) a unary operator $@_n$ ("according to n,...") for each nominal n, which resets the world of evaluation to be the value of n;

(iii) a binding operator $\downarrow n.$ ("letting n stand for the current world,...") for each nominal n, which resets the value of n to be the world of evaluation.

We will make these additions for both worlds and hyperconventions. This will allow us to regiment the idea of "holding fixed" one of the parameters.

More precisely, we extend $\mathcal{L}$ to a hybrid language $\mathcal{L}^{\mathsf{H}}$ with three kinds of atomics: regular proposition letters $p_1, p_2, p_3, \ldots$, state nominals $s_1, s_2, s_3, \ldots$ designating worlds, and interpretation nominals $i_1, i_2, i_3, \ldots$ designating hyperconventions. The formulas of $\mathcal{L}^{\mathsf{H}}$ are summarized in Backus-Naur form as with $\mathcal{L}$, except we add:

$$\phi ::= \cdots \mid s \mid i \mid @_s\, \phi \mid @_i\, \phi \mid \downarrow s.\phi \mid \downarrow i.\phi.$$

The definitions of hyperconventions, indices, and conventional models are exactly as before. A ***variable assignment*** over W is a function g that (i) maps each state nominal s to a world in W, and (ii) maps each interpretation nominal i to a hyperconvention over W. Satisfaction is now defined relative to indices and variable assignments. The semantic clauses for the old vocabulary are the same as before except relativized to variable assignments. For example, where $[\![\phi]\!]^{C,g} = \{\langle w, c\rangle \in I_W \mid C, w, c, g \Vdash_{\mathsf{c}} \phi\}$:

$$C, w, c, g \Vdash_{\mathsf{c}} \phi \boxdot\!\!\rightarrow \psi \quad \Leftrightarrow \quad \text{for all } \langle v, d\rangle \in f([\![\phi]\!]^{C,g}, w, c)\text{: } C, v, d, g \Vdash_{\mathsf{c}} \psi.$$

The clauses for the new vocabulary are as follows:

$$\begin{aligned}
C, w, c, g \Vdash_{\mathsf{c}} s &\quad\Leftrightarrow\quad g(s) = w \\
C, w, c, g \Vdash_{\mathsf{c}} i &\quad\Leftrightarrow\quad g(i) = c \\
C, w, c, g \Vdash_{\mathsf{c}} @_s\, \phi &\quad\Leftrightarrow\quad C, g(s), c, g \Vdash_{\mathsf{c}} \phi \\
C, w, c, g \Vdash_{\mathsf{c}} @_i\, \phi &\quad\Leftrightarrow\quad C, w, g(i), g \Vdash_{\mathsf{c}} \phi \\
C, w, c, g \Vdash_{\mathsf{c}} \downarrow s.\phi &\quad\Leftrightarrow\quad C, w, c, g^s_w \Vdash_{\mathsf{c}} \phi \\
C, w, c, g \Vdash_{\mathsf{c}} \downarrow i.\phi &\quad\Leftrightarrow\quad C, w, c, g^i_c \Vdash_{\mathsf{c}} \phi,
\end{aligned}$$

[20] For more on hybrid logic, see Areces and ten Cate 2007.

where g^s_w is the variable assignment just like g except $g^s_w(s) = w$, and similarly for g^i_c. Consequence is preservation of satisfaction over indices and variable assignments.

One benefit of adding hybrid operators is that they allow us to regiment what is often common ground in metalinguistic negotiations. While Alpha and Beta disagree over (19a), both agree to (19b).

(19) a. Pluto is a planet.
b. According to Alpha's use of the term, Pluto is a planet.

If we idealize a bit and assume that Alpha's mental state can be associated with a single hyperconvention,[21] then we can regiment (19b) using @: where a stands for Alpha's hyperconvention and p stands for (19a), (19b) becomes $@_a\, p$. And while $a \rightarrow (p \leftrightarrow @_a\, p)$ is a theorem of the conventional semantics, $p \leftrightarrow @_a\, p$ is not. So for Alpha, p and $@_a\, p$ coincide, since Alpha accepts a (i.e., Alpha accepts their own definition of 'planet'), whereas for Beta, they come apart since Beta does not accept a. Thus, hybrid logic allows us to articulate explicitly, in the object language, where disputants in a metalinguistic negotiation disagree.

These new hybrid expressions also allow us to regiment both countersubstratum and counterconventional readings. To regiment a countersubstratum reading, we can replace $\phi \mathrel{\Box\!\!\rightarrow} \psi$ with $\downarrow i.((\phi \wedge i) \mathrel{\Box\!\!\rightarrow} \psi)$, where i occurs nowhere in ϕ or ψ.[22] Intuitively, $\downarrow i.$ saves the current hyperconvention to nominal i, and then conjoining the antecedent with i ensures that this hyperconvention is "held fixed". Thus, $\downarrow i.((\phi \wedge i) \mathrel{\Box\!\!\rightarrow} \psi)$ is true at $\langle w, c\rangle$ iff for all the closest worlds v such that ϕ is true at $\langle v, c\rangle$, ψ is true at $\langle v, c\rangle$. Likewise, to regiment a (pure) counterconventional reading, we can replace $\phi \mathrel{\Box\!\!\rightarrow} \psi$ with $\downarrow s.((\phi \wedge s) \mathrel{\Box\!\!\rightarrow} \psi)$, where s occurs nowhere in ϕ or ψ. Thus, $\downarrow s.((\phi \wedge s) \mathrel{\Box\!\!\rightarrow} \psi)$ is true at $\langle w, c\rangle$ iff for all the closest hyperconventions d such that ϕ is true at $\langle w, d\rangle$, ψ is true at $\langle w, d\rangle$.

In the conventional semantics, counterpossibles are non-vacuous only on their c-monstrous readings. The countersubstratum readings of counterpossibles are all vacuous. So while $\neg \Diamond \phi \not\models_{\mathsf{c}} \phi \mathrel{\Box\!\!\rightarrow} \psi$, still $\neg \Diamond \phi \models_{\mathsf{c}} \downarrow i.((\phi \wedge i) \mathrel{\Box\!\!\rightarrow} \psi)$. If we wanted, we could always add impossible worlds to our models to render counterpossibles non-vacuous even on their countersubstratum readings. We will return to this possibility in §7. For now, what the conventional semantics demonstrates is that one can coherently maintain that counterpossibles are non-vacuous on their counterconventional reading without being committed to their non-vacuity on their countersubstratum reading.

[21] This idealization can be dropped by having interpretation nominals denote *sets* of hyperconventions; we set this complication aside for ease of exposition.

[22] This formulation assumes **Success**. A more general regimentation that works even in the absence of **Success** would be $\downarrow i.((\phi \wedge i) \mathrel{\Box\!\!\rightarrow} (\psi \wedge i))$.

5 Logical Expressivism

In the previous section, we argued that metalinguistic negotiations and counterconventionals are closely connected. The sentences that express metalinguistic disagreements are the very sentences that paradigmatically trigger counterconventional interpretations when embedded in the antecedents of counterfactuals. So our view that counterlogicals are counterconventionals pairs well with a view on which disputes over logic are metalinguistic negotiations. In this section, we motivate and elaborate this view (though we won't give a complete defense here).

One reason to think of disputes over logic as metalinguistic negotiations stems from Quine's changing the subject objection (§3). Start with the following plausible assumption (which we will take on board in what follows): it is constitutive of the meaning of the connectives that they obey the laws that they do.[23] Thus, when the intuitionist denies the law of excluded middle, they thereby interpret negation differently from the classicist. In that case, the question of how to interpret their disagreement immediately arises. Everyone agrees that the law of excluded middle holds *according to classical logic* but not *according to intuitionistic logic*. What, then, are the two sides disagreeing about?

According to a realist interpretation, they are disagreeing about whether the law of excluded middle is correct, *period*.[24] Just as Newtonian physics is a theory of the movement of material bodies, so too classical and intuitionistic logic are theories of negation, conjunction, and so on. No one doubted that the predictions of the Newtonian theory are what they are. What was at issue was whether they were correct. So too, what is at issue in disputes over logic is not what the predictions of each logic are, but which logic makes the correct predictions.

This does not yet address the Quinean worry, however. A logic is always the logic *of a language*. Before we can sensibly ask whether the law of excluded middle holds, for example, we must specify what language we are talking about (recall the case of Deviantsville). Quine's objection is that there is no way of answering that question without trivializing the debate. Since the laws governing the connectives are plausibly constitutive of their meaning, the classical and non-classical logician are necessarily making claims about different languages, and so talking past each other. They cannot disagree without losing the sameness of subject matter required

[23]This assumption is not without controversy. As an anonymous reviewer points out, most metasemantic theories allow for two speakers to associate the same meaning with an expression despite believing it to license different inferences. One who holds that "Amber is mean" follows from "Amber is a cat" may be a monster, but he does not thereby mean something different by 'cat'. Why should logical expressions be any different? Can't a non-classical logician weaken a law slightly (say, to account for semantic paradoxes), without thereby adopting an entirely different connective from the classical logicians, so long as the divergence remains relatively minor? Perhaps; in that case, whether a particular dispute about logic is a metalinguistic negotiation would depend on how radical the divergences are, and on whether the shared meaning builds in the laws (perhaps via externalist means).

[24]See Read 2006; Priest 2008; Russell 2008; Field 2009a; Keefe 2014 for discussion.

for genuine disagreement.

Another interpretation is that they are disputing whether the law of excluded middle holds for a specific natural language, such as English. That is, the dispute is over an empirical claim about the way people actually use 'not', 'or', etc., in English. The classicist says English speakers use these words in accordance with the law of excluded middle; the intuitionist denies this. This avoids the Quinean worry by construing opposing logics as competing descriptive theories about the same language.

Though many disputes over logic may seen this way, there are two reasons to find it unsatisfactory as a universal interpretation of all logical disputes. First, it does not explain why philosophers of logic would appeal to such esoteric matters as semantic paradoxes, the epistemology of mathematics, computational complexity, or quantum mechanics—matters mostly unknown to the vast majority of native speakers—to justify their positions. None of those considerations are relevant to the linguistic question of how ordinary speakers actually use logical vocabulary; that's a matter to be settled by surveys.[25]

Second, it fails to account for the normativity of logic.[26] Logic is the study of *correct* reasoning, not just the study of how people actually reason. The vast majority of a given linguistic community could commit the fallacy of affirming the consequent, and it would remain a fallacy. The fact that people tend to reason in a particular way may constrain our theory of correct reasoning, but it cannot completely determine it.

But there is a third option: they are engaging in a metalinguistic negotiation. Yes, the classicist and intuitionist associate different meanings with logical vocabulary, and so are disagreeing fundamentally about meaning. But that does not mean the two sides are "changing the subject", or talking past each other. They disagree over *how to use* the logical vocabulary, not over factual claims about how certain people *do* use that vocabulary. In other words, logical disagreements express differences over which inferences to count as valid.

This interpretation of logical disputes has close ties to earlier conventionalist views of logic, such as those of Carnap (1937) and Ayer (1952). More recently, Kouri Kissel (2018, 2019) has explicitly defended the idea that logical disputes are metalinguistic negotiations. In what follows, we lay out a version of this view, which we call ***logical expressivism***, on which logical assertions (e.g., that some inference is valid or that some logic is correct) are expressions of the logical commitments of the speaker, rather than factual claims either about an "objectively correct" logic, or about the patterns of reasoning a specific group happens to employ.

[25]Logicians do sometimes appeal to ordinary linguistic judgments to justify their position; see Ripley 2016 for discussion. A well-known example is the use of the paradoxes of material implication as a motivation for relevance logics (Anderson and Belnap, 1975; Mares, 2012).

[26]Not everyone agrees that logic is normative in any important sense—see Harman 1984 for well-known criticisms. See also, however, MacFarlane 2004; Field 2009b; Steinberger 2019a,b for defenses against Harman's skepticism.

Here is how we are thinking about the view. At any given time, a speaker adopts certain conventions for how to talk. In most conversations, speakers will converge on a sizeable set of conventions, though they may disagree over edge cases, e.g., whether to use 'planet' to include Pluto. Metalinguistic negotiations are disputes over such cases—that is, they concern what conventions to adopt rather than what the extra-conventional world is like.

There is no such thing as a *correct* or *incorrect* convention. Nature does not settle what the word 'planet' must mean; that's our job. As Frege (1892) taught us, "Nobody can be forbidden to use any arbitrarily producible event or object as a sign for something."

Still, some conventions may be *better* or *worse* for certain purposes. Some might fit more nicely with preexisting conventions, or be easier for us to use, or (in the bad case) be offensive or harmful to a particular group of people. The fact that we can decide which conventions to adopt does not imply that all conventions are on a par.[27] Speakers can meaningfully disagree over which conventions are better, or which should be used, even if there is another sense in which our linguistic conventions are arbitrary.

Logics are just conventions of a special sort: they are conventions governing the use of words like 'not', 'and', 'or', 'true', 'valid', and so on. Nothing in principle stops us from adopting a language governed by classical logic, intuitionistic logic, paraconsistent logic, or even the trivial logic where everything is valid. Even so, some logics are better than others and there can be reasonable disagreement over which is better for specific purposes.[28] Everyone will probably agree, for instance, that the trivial logic is a bad logic to adopt. But not everyone will agree that classical logic is the unique best logic for every application. There does not have to be a "correct logic" for there to be genuine disagreements over logic, just as there does not have to be a "correct convention" for there to be genuine disagreements over convention.

Consider, as a case study, a dispute one might find in a seminar on paradoxes:

(20) **Param:** The Liar sentence is both true and not true.

Clara: No it's not. The Liar sentence cannot be both true and not true because no sentence can.

Param: I don't accept the law of non-contradiction. The Liar paradox shows that some contradictions are true.

Clara: Look, I know it's tempting to blame the law of non-contradiction, but there's a good reason not to. Even if the Liar sentence were both true and not true, Curry's paradox would remain unresolved.

[27] For more on this point, see Haslanger 2000, 2005; McConnell-Ginet 2006, 2008; Plunkett and Sundell 2013; Plunkett 2015.

[28] Putnam (1969, pp. 231–232) defends quantum logic along these lines.

How should we think about what that disagreement in (20) is about? According to logical expressivism, it is not about anything purely factual. Param is not asserting that the Liar sentence is both true and not true according to the way English speakers tend to speak. Nor is he asserting that the Liar sentence is both true and not true in some objective, language-transcendent sense. Instead, he is best viewed as making a proposal to adopt a certain logical convention—that is, to use 'not', 'and', and 'true' in such a way as to make the Liar sentence count as both true and untrue.

To do this, he does not need to explicitly mention anything about words or conventions. Indeed, doing so would change the force of his assertions. When Param asserts (21a), he does not thereby assert (21b):

(21) a. The Liar sentence is both true and not true.
b. According to my logic, the Liar sentence is both true and not true.

Param and Clara agree on (21b), but disagree about (21a). Instead, Param *expresses* (21b) by asserting (21a), much like one expresses that one believes p by asserting p, without thereby asserting that one believes p. This is the sense in which the view is a form of expressivism.

Logical expressivism respects the spirit of Quine's changing the subject objection: two people cannot disagree about logic without thereby using words differently. But that does not mean they are, as Quine thought, simply talking past each other. Instead, their disagreement can be normative—they may be disagreeing about *how to use* the logical vocabulary. Logical disputants are engaging in a normative, non-trivial debate, centered on exactly the kinds of concerns (semantic paradoxes, vagueness, etc.) where the appropriateness of logical conventions is at stake. So logical expressivism answers Quine's challenge, while accounting both for the normativity of logic and for the types of considerations philosophers of logic raise in support of their views. These are all marks in its favor.

In addition to these theoretical virtues, logical expressivism is supported by linguistic data. Contrast the following:

(22) a. I consider Pluto to be a planet.
b. #I consider Pluto to be more than 10^{22} kg.

As Kennedy and Willer (2016) point out, the verb 'considers' carries a "counterstance-contingency" presupposition: $\ulcorner \alpha$ considers $\phi \urcorner$ is defined in a context only when ϕ is contingent on some extra-factual matter like speakers' tastes, or linguistic decisions such as how to define a predicate. Thus, (22a) sounds fine in the context of (13) because whether Pluto is (or counts as) a planet is contingent on whether we adopt the IAU's definition or the old definition. By contrast, (22b) sounds bad in the context of (13) since whether Pluto is more than 10^{22} kg cannot be resolved by stipulation.[29]

[29]This does not mean that (22b) sounds bad in every context. In some contexts, the mass of Pluto

In a context like (20), (23a) and (23b) sound fine:

(23) a. I consider the Liar sentence to be both true and not true.
b. I consider the law of non-contradiction to be valid.

This indicates that in the context of (20), whether the Liar is both true and not true, or whether the law of non-contradiction is valid, can be settled by stipulation. In other words, we can *choose* to adopt a logic on which contradictions can be true or on which the law of non-contradiction holds. Statements about logic are expressions of these sorts of choices.

This does not mean that the truth of *every* assertion about logic can be settled by fiat. Contrast (23a) and (23b) with:

(24) a. #I consider the Liar sentence to be both true and not true according to a paraconsistent interpretation.
b. #I consider the law of non-contradiction to be valid in classical logic.

These sound terrible in the context of (20). That is because Param and Clara do not disagree over what paraconsistent or classical logic are. It is not open to either of them to decide what entails what on classical logic or whether the Liar sentence is both true and not true on a paraconsistent interpretation (though perhaps in other contexts, one could dispute these things).

Nor does it mean that speakers are logically omniscient—that they can know every logical truth at will. Speakers can be (and often are) committed to a convention without realizing it. This can happen if the speaker makes the convention they adopt contingent on worldly facts. Someone could choose, for example, to adopt the IAU's definition of 'planet' without knowing what that definition is exactly. In that case, it would be weird to say (22a), since the speaker cannot simply decide that Pluto meets the IAU's definition of 'planet'.

The same goes for logical conventions. If a student in introductory logic is working on a problem set, it would sound weird for them to say:

(25) #I consider the law of excluded middle to be valid.

In this context, the student (even if just for the purposes of the problem set) is adopting the logic the problem set is asking about and is trying to determine what laws are valid according to it. This leaves no room for decision: it is not an option for them to choose what is valid according to that logic. Logical mistakes are possible even if logic is settled by convention.

Finally, logical expressivism does not imply that logicians engaging in disputes over logic must take themselves to be disputing how to use words, any more

might be something that can be settled at least partly by stipulation, e.g., if the boundaries of Pluto are unclear or if there is disagreement over how to define a kilogram. The point is just that in the context of (13), the mass of Pluto cannot be settled by stipulation.

than normative expressivism implies participants in a moral disagreement must take themselves to be non-cognitivists. The considerations that lead us to logical expressivism have less to do with the self-understanding of logicians, and more to do with reflections on the nature of logic as well as various kinds of linguistic data that have been used more generally in philosophy of language to motivate expressivism.

We do not claim at this point to have given definitive arguments in favor of logical expressivism. A complete defense will have to be taken up elsewhere. We have simply provided (to our mind, compelling) reasons to take logical expressivism seriously, and thus to take seriously the idea that many disputes over logic are best viewed as metalinguistic negotiations.

6 A Counterconventional Semantics for Counterlogicals

We now present our counterconventional semantics for counterlogicals, which we call the ***(logical) expressivist semantics***. It extends the conventional semantics from §4 to allow counterfactuals to shift the interpretation of both logical and non-logical vocabulary.[30] We then explain how this semantics addresses the Quinean objection.

The first step is relatively simple. Where W is a set of worlds, we redefine a ***hyperconvention*** over W to be a function c such that:

(i) for each $p \in \mathtt{At}$, $c(p) \subseteq W$

(ii) for each n-place $\triangle \in \{\neg, \wedge, \vee, \rightarrow, \Box, \Diamond\}$, $c(\triangle)\colon\ \wp(W)^n \rightarrow \wp(W)$.

So in addition to assigning a set of worlds to each atomic sentence, hyperconventions assign intensions to the logical connectives, where the intension of a connective is modeled as an operation on sets of worlds.[31] As before, an index is just a world-hyperconvention pair and the set of indices over W is I_W. We will call a hyperconvention ***classical*** if the following conditions are met for all $X, Y \subseteq W$:

$$\begin{aligned} c(\neg)(X) &= \overline{X} & \qquad c(\wedge)(X,Y) &= X \cap Y \\ c(\Box)(X) &= \{w \in W \mid X = W\} & \qquad c(\vee)(X,Y) &= X \cup Y \\ c(\Diamond)(X) &= \{w \in W \mid X \neq \varnothing\} & \qquad c(\rightarrow)(X,Y) &= \overline{X} \cup Y \end{aligned}$$

An index is ***classical*** if its hyperconvention is classical. We let CI_W be the set of classical indices over W. An ***expressivist model*** is a pair of the form $\mathcal{E} = \langle W, f \rangle$

[30] Similar semantic frameworks for knowledge and belief have been developed to address the problem of logical omniscience. See, e.g., Muskens 1991; Williamson 2009, for examples. To our knowledge, such approaches have not been extended to counterfactuals.

[31] Note that hyperconventions do not determine the interpretation of $\Box\!\!\rightarrow$. Although we would like to extend our semantics to allow for the interpretation of $\Box\!\!\rightarrow$ to shift, it is not clear how to do this consistently. We cannot simply have hyperconventions assign an interpretation to $\Box\!\!\rightarrow$, since $\Box\!\!\rightarrow$ denotes an operation on sets of indices, which themselves contain hyperconventions. See Kocurek 2020 for a proposal to avoid this problem.

where $W \neq \varnothing$ and $f\colon\ \wp(I_W) \times I_W \to \wp(I_W)$. The satisfaction relation $\Vdash_{\mathsf{e}}$ is defined as follows ($\triangle \in \{\neg, \wedge, \vee, \to, \Box, \Diamond\}$):

$$\begin{aligned}
\mathcal{E}, w, c \Vdash_{\mathsf{e}} p &\quad\Leftrightarrow\quad w \in c(p)\\
\mathcal{E}, w, c \Vdash_{\mathsf{e}} \triangle(\phi_1, \ldots, \phi_n) &\quad\Leftrightarrow\quad w \in c(\triangle)(\llbracket\phi_1\rrbracket^{\mathcal{E},c}, \ldots, \llbracket\phi_n\rrbracket^{\mathcal{E},c})\\
\mathcal{E}, w, c \Vdash_{\mathsf{e}} \phi \mathrel{\Box\!\!\to} \psi &\quad\Leftrightarrow\quad \text{for all } \langle v, d\rangle \in f(\llbracket\phi\rrbracket^{\mathcal{E}}, w, c)\text{: } \mathcal{E}, v, d \Vdash_{\mathsf{e}} \psi,
\end{aligned}$$

where $\llbracket\phi\rrbracket^{\mathcal{E}} := \{\langle w, c\rangle \in I_W \mid \mathcal{E}, w, c \Vdash_{\mathsf{e}} \phi\}$ and $\llbracket\phi\rrbracket^{\mathcal{E},c} = \{w \in W \mid \mathcal{E}, w, c \Vdash_{\mathsf{e}} \phi\}$. Note that if c is classical, then the clause for each member of $\{\neg, \wedge, \vee, \to, \Box, \Diamond\}$ is exactly the same as in the conventional semantics.

Consequence is the preservation of satisfaction over *classical* indices: $\Gamma \models_{\mathsf{e}} \phi$ iff for every expressivist model $\mathcal{E}$ and every $\langle w, c\rangle \in CI_W$, if $\mathcal{E}, w, c \Vdash_{\mathsf{e}} \Gamma$, then $\mathcal{E}, w, c \Vdash_{\mathsf{e}} \phi$. This guarantees that $\models_{\mathsf{e}}$ obeys classical logic. Of course, if we wanted, we could obtain a different base logic by redefining consequence accordingly. We could, for instance, obtain a Kleene base logic by defining Kleene indices by how they interpret the connectives (e.g., $c(\neg)$ is an operation where (i) $c(\neg)(X) \cap X = \varnothing$ and (ii) $c(\neg)(c(\neg)(X)) = X$), and then defining consequence as satisfaction-preservation over Kleene indices.[32] For ease of exposition, though, we stick with classical logic as our base logic.

Given this notion of consequence, the impossible worlds semantics and the logical expressivist semantics generate the same logic over $\mathcal{L}$—that is, $\models_{\mathsf{i}} = \models_{\mathsf{e}}$ (§A). Thus, anything the former can do, the latter can do as well. In particular, counterlogicals are not vacuously true on the expressivist semantics.

The two frameworks render counterlogicals non-vacuous in different ways, however. The impossible worlds semantics does it essentially by brute force: it simply introduces impossible worlds satisfying whatever sentences you like into the model. There are no rules governing how truth works at impossible worlds; truth is simply determined by fiat.

By contrast, the expressivist semantics generates the non-vacuity of counterlogicals in a more systematic fashion. Truth is always determined compositionally: the semantic value of a complex formula is always a function of the semantic value of its parts. It is just that which function that is can vary in the scope of counterfactuals. In effect, the expressivist semantics replaces *impossible worlds* with *possible worlds under different descriptions*.

This difference can be made manifest in a hybrid language. Recall from §4 that counterpossibles are only non-vacuous on their c-monstrous readings in the

[32] This raises the question: *which* logics can be represented by a hyperconvention? The answer depends on what "represent" means. Here is a natural suggestion: c represents a logic $\mathbf{L}$ in $\mathcal{E}$ just in case $\phi_1, \ldots, \phi_n \vdash_{\mathbf{L}} \psi$ iff for all w, if $\mathcal{E}, w, c \Vdash_{\mathsf{e}} \{\phi_1, \ldots, \phi_n\}$, then $\mathcal{E}, w, c \Vdash_{\mathsf{e}} \psi$. Then it can be shown that a logic is representable in any countably infinite model iff it is reflexive, transitive, monotonic, and congruential, i.e., validates the replacement of logical equivalents (see Kocurek 2020 for the proof). While this rules out the ability to represent certain substructural logics (e.g., non-monotonic logics), these limitations can be lifted by employing other notions of representability. See Kocurek 2020 for details. Thanks to an anonymous reviewer for raising this question.

conventional semantics their countersubstratum readings are all vacuously true. This was stated using hybrid operators: countersubstratums are counterfactuals of the form $\downarrow i.((\phi \wedge i) \boxrightarrow \psi)$, and $\neg \Diamond \phi \vDash_{\mathsf{c}} \downarrow i.((\phi \wedge i) \boxrightarrow \psi)$ even though in general $\neg \Diamond \phi \nvDash_{\mathsf{c}} \phi \boxrightarrow \psi$. Similarly, in the expressivist semantics, counterlogicals are only non-vacuous on c-monstrous readings. This can be stated using hybrid operators as well though doing so requires a bit more work.

To illustrate, compare (3) and (11) with their respective regimentations in $\mathcal{L}^{\mathsf{H}}$:

(3) If the Liar sentence were both true and not true, the moon would be made of green cheese.

$$(l \wedge \neg l) \boxrightarrow m$$

(11) If the Liar sentence were both true and not true according to a classical interpretation of negation and conjunction, the moon would be made of green cheese.

$$\downarrow i.(@_i(l \wedge \neg l) \boxrightarrow m)$$

Like the impossible worlds semantics, $(l \wedge \neg l) \boxrightarrow m$ is not valid in the expressivist semantics. This is because counterfactuals can shift the interpretation $\wedge$ and $\neg$ so that $l \wedge \neg l$ is true at some worlds. By contrast, $\downarrow i.(@_i(l \wedge \neg l) \boxrightarrow m)$ is valid in the expressivist semantics. This is because the antecedent of the counterfactual is forced by $@_i$ to be evaluated according to a classical hyperconvention.

Note that we cannot formalize countersubstratums as $\downarrow i.((\phi \wedge i) \boxrightarrow \psi)$ in the expressivist semantics, since the counterfactual can shift the interpretation of $\wedge$ (so $\phi \wedge i$ might be true at an index even if i is not). But we can formalize them in the expressivist semantics (given **Success**) as:

$$\downarrow i.(\downarrow j.@_i(\phi \wedge j) \boxrightarrow \psi).$$

In the conventional semantics, this is equivalent to our original formulation of countersubstratums, viz., $\downarrow i.((\phi \wedge i) \boxrightarrow \psi)$. The two come apart in the expressivist semantics, but it is easy to check that $\downarrow i.(\downarrow j.@_i(\phi \wedge j) \boxrightarrow \psi)$ has the same truth conditions in the expressivist semantics as $\downarrow i.((\phi \wedge i) \boxrightarrow \psi)$ has in the conventional semantics. Thus, we can state the idea that counterlogicals are only non-vacuous on their c-monstrous readings as follows: if $\vDash_{\mathsf{e}} \neg \phi$, then $\vDash_{\mathsf{e}} \downarrow i.(\downarrow j.@_i(\phi \wedge j) \boxrightarrow \psi)$.

We can now see how our expressivist semantics addresses the Quinean objection. The objection is correct in that non-vacuous counterlogicals involve a change in the meaning of the logical connectives. But that doesn't imply that all counterlogicals are vacuously true. Instead, the non-vacuous ones are simply counterconventionals.

Is this counterlogical non-vacuism or counterlogical vacuism? It is hard to say, for it draws a distinction that was not made when we introduced those labels. On

the one hand, **Counterlogical Vacuism** does not hold on the expressivist semantics, so in that sense, it is a non-vacuist view. On the other, proponents of counterlogical vacuism often argue as though they have specifically countersubstratum readings in mind, and our theory predicts that all of *those* are vacuous. So our view is best characterized as an intermediate position: it is non-vacuist about counterconventional readings of counterlogicals but vacuist about countersubstratum readings.

To elaborate, there is a kind of confusion around what exactly a counterlogical is. This is due to an ambiguity in the notion of a "logically impossible" antecedent. To explain, we use a distinction due to Sandgren and Tanaka (2019), but phrased slightly differently: distinguish claims that are *actually* logically impossible, i.e., according to the logic we actually adopt, from sentences that are *counterfactually* logically impossible, i.e., according to a counterfactual logic. For example, the antecedent of (3) is actually logically impossible, i.e., impossible by our lights, though not counterfactually impossible, i.e., impossible by the lights of the counterfactual logic in question. By contrast, the antecedent of (11) is counterfactually logically impossible: even by the lights of the counterfactual logic, it is absurd.[33]

When we talk about counterfactuals with "logically impossible antecedents", we must be careful to clarify the logic relative to which the antecedent is considered logically impossible. Our view is that counterfactuals whose antecedents are actually logically impossible are not automatically vacuous, but those whose antecedents are counterfactually logically impossible are. Counterlogical non-vacuism is right if we use actually logical impossibility to define counterlogicals, while counterlogical vacuism is right if we use counterfactual logical impossibility. Put in these terms, the view seems quite sensible: it simply states that for a counterlogical to be non-vacuous, its antecedent must be coherent by its own lights.

7 Counterpossibles

This paper has been a tale of two semantics. In the beginning, it was the impossible worlds semantics. There, non-vacuous counterlogicals were explained by introducing logically impossible worlds, modeled as inconsistent and/or incomplete sets of sentences, where the classical laws of logic need not apply. But this semantics is prone to a kind of Quinean changing the subject objection: in order for counterlogicals to be interpreted non-vacuously, we have to reinterpret the logical vocabulary so that their antecedents no longer express genuine logical impossibilities.

In response, we flipped this objection on its head and turned it into a theory.

[33] We still assume that ϕ is logically impossible according to a logic **L** if $\neg\,\phi$ is **L**-valid. But it is an open question how to understand the notion of logical impossibility in certain non-classical logics. For instance, in the strong Kleene 3-valued logic **K3**, no formula is valid. So by the current definition, nothing is logically impossible in **K3**. As another example, in the logic of paradox **LP**, every formula is true on some interpretation: even if ϕ is **LP**-valid, there are valuations satisfying $\neg\,\phi$. So whether the negation of a validity counts as a "logically impossibility" in **LP** is debatable. Thus, the very notion of logical impossibility might become more complicated in non-classical settings.

Counterfactuals, as we saw, are generally capable of shifting the conventional interpretation of their constituents, especially in metalinguistics negotiations. We then defended logical expressivism, according to which disputes over logic are best thought of as metalinguistic negotiations and logical assertions as expressions of the logical conventions speakers adopt.

This paved the way for a new, logical expressivist semantics. On this semantics, non-vacuous counterlogicals are explained in terms of shifting conventions governing the logical vocabulary. Counterlogicals, when interpreted non-vacuously, are simply counterconventionals. This vindicates the insight behind the Quinean objection—that some sort of change of meaning is taking place in counterlogicals—while preserving the felt non-vacuity of counterlogicals.

Thus, no separate theory of counterlogicals is required. We do not need to introduce logically impossible worlds to accommodate non-vacuous counterlogicals in our semantic theories. Non-vacuous counterlogicals come for free with a semantics for convention-shifting expressions, which is already needed to explain counterconventionals and other c-monstrous expressions.

Don't we still need impossible worlds, though? After all, we still have to deal with counterpossibles that are not counterlogicals, such as (5)–(7).

(5) If water were hydrogen peroxide, life would not exist.

(6) If Sam were a salmon, she would have wings.

(7) If Socrates had existed without his singleton, he would be sad.

The careful reader may have noticed that the expressivist semantics from §6 treats all counterpossibles similarly to how it treats counterlogicals. Like counterlogicals, counterpossibles are not vacuously true: $\neg \Diamond \phi \not\models_{\mathsf{e}} \phi \mathrel{\Box\!\!\rightarrow} \psi$. But also like counterlogicals, counterpossibles are only non-vacuous on counterconventional readings. On their countersubstratum readings, they are all vacuous: $\neg \Diamond \phi \models_{\mathsf{e}} \downarrow i.(\downarrow j.@_i(\phi \wedge j) \mathrel{\Box\!\!\rightarrow} \psi)$.

Yet it is far from obvious that the most natural interpretations of (5)–(7) are really counterconventional ones. We are not saying 'If water were *classified as* hydrogen peroxide,...' or 'If Sam *counted as* a salmon,...'. Rather, these counterpossibles seem non-vacuous even on their countersubstratum readings. If that is right, then we may need to add impossible worlds anyway to accommodate their apparent non-vacuity. But if impossible worlds are already needed to account for non-vacuous counterpossibles, why not just use the impossible worlds semantics for counterlogicals, too? Doesn't the fact that we need impossible worlds for other counterpossibles cut against giving a different treatment of counterlogicals?

We think not, for two reasons. First, the argument runs both ways. Convention-shifting is already needed to analyze counterconventionals, independently of counterpossibles. Impossible worlds are of no help in giving a systematic theory of counterconventionals since most counterconventionals are not counterpossibles (e.g., the

Pluto example, (14)). So the same argument can be used in favor of the expressivist semantics: if convention-shifting is already needed to account for counterconventionals, why not just use the expressivist semantics for counterlogicals, too?

Second, and more importantly, as we saw in §3, there are special reasons for questioning the impossible worlds semantics for counterlogicals that don't apply to other counterpossibles. In the case of counterlogicals, there is a good case to be made that we are reinterpreting the logical vocabulary in the scope of the counterfactual; we do not mean classical conjunction and negation by 'and' and 'not' in (3). *Maybe* this applies to other kinds of counterpossibles too, such as counteranalyticals ('If there were round squares,. . . ') or countermathematicals ('If Fermat's last theorem had failed,. . . '). It is less clear, however, that any change of meaning is taking place in countermetaphysicals such as (5)–(7).

At any rate, there are reasons to think that logic has a closer connection to meaning and convention than other kinds of metaphysical claims. Whether Sam is a salmon does not seem to be a matter of convention or subject to stipulation in the same way that the meanings of the logical connectives are. So it is not ad hoc to take a kind of disjunctive approach, using convention-shifting to model some counterpossibles, e.g., counterlogicals, while using impossible worlds to model others, e.g., countermetaphysicals.

With all that said, we are sympathetic to a more radical view of impossible worlds—one that does not give separate accounts of logical and non-logical impossibility, but instead views impossible worlds of *all* sorts through the lens of shifting conventions. On this view, impossible worlds are not separate entities from possible worlds; impossible worlds are just possible worlds *under different descriptions*. This would open the door to a novel account of counterpossibles, one that aims to strike a balance between vacuism and non-vacuism. In short: counterpossibles are only non-vacuous on counterconventional readings, whereas they are all vacuous on a countersubstratum reading. While we do not have space to defend this view in full, we nevertheless wish to close by expressing optimism for this more ambitious project.

This account of impossible worlds is not original to us. It is defended in various forms in the literature on imagining the impossible. Many philosophers have argued that we cannot imagine the impossible precisely because whenever we try to do so, we end up imagining a possible scenario that we simply describe in an impossible-sounding way.[34] For instance, Kripke (1971, 1980) claims that when we try to imagine a world where Hesperus is distinct from Phosphorus, we end up imagining something else—that 'Hesperus' and 'Phosphorus' are used to refer to distinct individuals, that the evening star and the morning star are distinct, etc. These are not *really* worlds where Hesperus, as we use the term, is distinct from Phosphorus, as we use the term. We confuse ourselves into thinking otherwise due to an implicit meaning change.

[34]For discussion, see Yablo 1993; van Inwagen 1998; Gregory 2007; Kung 2010, 2016.

But there is another way to see things. Some philosophers, such as Kung (2010, 2016), have held that it is actually quite easy to imagine the impossible: all it is to imagine the impossible is to imagine a scenario that represents some impossible proposition being true.[35] Imagination, on this view, involves constructing a kind of representation of a world. We never imagine a world directly but only under a description or mode of presentation. As Kung (2010, pp. 626–628) puts it, we attach "labels" to various items of our representation.

Moreover, this view of counterpossibles as counterconventionals is reminiscent of other views in the literature. Vetter (2016) argues that counterpossibles are only non-vacuous on epistemic, representation-sensitive readings of counterfactuals. More directly related, Locke (2019) defends the view that "metaphysical counterpossibles function to illustrate or express changes, or consequences of changes, to the actual constitutive rules that govern language use while remaining in the object language where terms are used rather than mentioned." (p. 8) While Locke does not subscribe specifically to the view that impossible worlds are possible worlds under different descriptions, his view that impossible worlds talk is an "object language resource for "mis-using" language" (p. 11) closely resembles the kind of view we are sketching here.

On this picture, even countermetaphysicals such as (5)–(7) can ultimately be viewed as counterconventionals. When evaluating (5), for instance, we imagine a scenario where our lakes and oceans are filled with hydrogen peroxide, which we stipulate to be "water". When evaluating (6), we consider a scenario involving a salmon that we (as the supposer) stipulate to be "Sam". When evaluating (7), we stipulate some set-like abstract objects, whose existence are not entailed by the existence of their members, to be "sets". In each case, we are shifting what we actually mean by the relevant expression ('water', 'Sam', or 'singleton'). For sure, there are details to be ironed out. But we think there is promise in this general strategy for understanding the underlying mechanics of counterpossibles.

Note this does not mean counterpossibles are *about* language in any meaningful sense. It just means that they involve shifts in language. So 'if water were hydrogen peroxide,. . . ' does not mean the same thing as 'if water were classified as hydrogen peroxide,. . . '. The former shifts the labels we actually use; the latter describes what labels are used in other counterfactual scenarios. Counterconventionals are no more about linguistic conventions than quantifiers are about variable assignments.

Let's end where we began: with Cohen's difficult question.

(1) If the failure of modus ponens implied the failure of modus tollens, and modus ponens were to fail, would modus tollens fail too?

We now have an answer: it depends. There are two readings of the question. On a countersubstratum reading, the answer is yes: holding fixed what we mean by

[35] See also Berto 2017.

'implies', any scenario where the failure of modus ponens implies the failure of modus tollens and where modus ponens fails must be a scenario where modus tollens fails. On a counterconventional reading, however, the answer is no: even though modus ponens is actually valid, it is not counterfactually valid, i.e., valid according to the counterfactual conventions expressed by the antecedent. Both readings are readily available, which explains why we feel pulled in both directions. The question was difficult to answer only because it was ambiguous.

A Appendix

In this appendix, we establish that the impossible worlds semantics and the expressivist semantics generate the same logic over $\mathcal{L}$, i.e., that $\models_{\mathsf{i}} = \models_{\mathsf{e}}$. To do this, we establish the following:

Theorem A.1.

(a) For any expressivist model $\mathcal{E} = \langle W, f \rangle$ and any $x \in I_W$, there is a impossible worlds model $\mathcal{E}^{\mathsf{i}} = \langle W^{\mathsf{i}}, P^{\mathsf{i}}, f^{\mathsf{i}}, V^{\mathsf{i}} \rangle$ and a $w \in W^{\mathsf{i}}$ such that for all $\phi \in \mathcal{L}$:

$$\mathcal{E}, x \Vdash_{\mathsf{e}} \phi \quad \Leftrightarrow \quad \mathcal{E}^{\mathsf{i}}, w \Vdash_{\mathsf{i}} \phi.$$

When $x \in CI_W$, we can take $w \in P^{\mathsf{i}}$.

(b) For any impossible worlds model $\mathcal{I} = \langle W, P, f, V \rangle$ and any $w \in W$, there is a expressivist model $\mathcal{I}^{\mathsf{e}} = \langle W^{\mathsf{e}}, f^{\mathsf{e}} \rangle$ and an $x \in I_{W^{\mathsf{e}}}$ such that for all $\phi \in \mathcal{L}$:

$$\mathcal{I}, w \Vdash_{\mathsf{i}} \phi \quad \Leftrightarrow \quad \mathcal{I}^{\mathsf{e}}, x \Vdash_{\mathsf{e}} \phi.$$

When $w \in P$, we can take $x \in CI_{W^{\mathsf{e}}}$.

Corollary A.2. For any $\Gamma \subseteq \mathcal{L}$ and $\phi \in \mathcal{L}$, $\Gamma \models_{\mathsf{i}} \phi$ iff $\Gamma \models_{\mathsf{e}} \phi$.

It is easiest to establish Theorem A.1(a) first.

Proof (*Theorem A.1(a)*): Suppose first that $x \in CI_W$. Define $\mathcal{E}^{\mathsf{i}} = \langle W^{\mathsf{i}}, P^{\mathsf{i}}, f^{\mathsf{i}}, V^{\mathsf{i}} \rangle$ as follows:

- $W^{\mathsf{i}} = I_W$
- $P^{\mathsf{i}} = W \times \{c_x\}$
- for each $y \in W^{\mathsf{i}}$ and $X \subseteq W^{\mathsf{i}}$, $f^{\mathsf{i}}(X, y) = f(X, y)$
- for each $y \in P^{\mathsf{i}}$, $V^{\mathsf{i}}(p, y) = 1$ iff $w_y \in c_y(p)$

- for each $y \notin P^{\mathsf{i}}$, $V^{\mathsf{i}}(\phi, y) = 1$ iff $\mathcal{E}, y \Vdash_{\mathsf{e}} \phi$.

Clearly, $\mathcal{E}^{\mathsf{i}}$ is a impossible worlds model and $x \in P^{\mathsf{i}}$. It suffices to show that for any ϕ and any $y \in I_W$:

$$\mathcal{E}, y \Vdash_{\mathsf{e}} \phi \quad \Leftrightarrow \quad \mathcal{E}^{\mathsf{i}}, y \Vdash_{\mathsf{i}} \phi.$$

If $y \notin P^{\mathsf{i}}$, then by construction, $\mathcal{E}^{\mathsf{i}}, y \Vdash_{\mathsf{i}} \phi$ iff $V^{\mathsf{i}}(\phi, y) = 1$ iff $\mathcal{E}, y \Vdash_{\mathsf{e}} \phi$. If $y \in P^{\mathsf{i}}$, then we proceed by induction. The atomic case holds by definition of V^{i}. The other cases are straightforward since $c_y = c_x$ is classical and since $P^{\mathsf{i}} = W \times \{c_x\}$.

Now suppose $x \notin CI_W$. Then we can define $\mathcal{E}^{\mathsf{i}}$ as above except now we take $P^{\mathsf{i}} = CI_W$. Then $\mathcal{E}, x \Vdash_{\mathsf{e}} \phi$ iff $\mathcal{E}^{\mathsf{i}}, x \Vdash_{\mathsf{i}} \phi$ by construction of V^{i}. ■

Theorem A.1(a) is not terribly surprising in retrospect. All it says is that anything that is i-valid is also e-valid. But i-validity is pretty weak without further constraints. One way to make that clear is to observe that, as far as the logic is concerned, counterfactuals behave exactly like distinct atomic sentences.

Definition A.3. An $\mathcal{L}$-formula is an ***S5-formula*** if it does not contain $\boxright$. An $\mathcal{L}$-formula is a ***counterfactual*** if its main connective is $\boxright$.

Proposition A.4. Let $\mathcal{M} = \langle P, i\rangle$ be an **S5**-model (where $i(p) \subseteq P$ for all $p \in \mathtt{At}$) and let $\Phi\colon P \to \wp(\mathcal{L})$ map every $w \in P$ to a set Φ_w of counterfactuals. Then there is an impossible worlds model $\mathcal{I} = \langle W, P, f, V\rangle$ such that for any $w \in P$:

(i) if ϕ is an **S5**-formula, then $\mathcal{I}, w \Vdash_{\mathsf{i}} \phi$ iff $\mathcal{M}, w \Vdash_{\mathbf{S5}} \phi$

(ii) if ψ is a counterfactual, then $\mathcal{I}, w \Vdash_{\mathsf{i}} \psi$ iff $\psi \in \Phi_w$.

Proof: WLOG, we may assume that P is disjoint from $\mathcal{L}$ and from $(\mathcal{L} \times P)$. Define $\mathcal{I} = \langle P \cup \mathcal{L} \cup (\mathcal{L} \times P), P, f, V\rangle$, where:

- for each $p \in \mathtt{At}$ and $w \in P$, $V(p, w) = 1$ iff $w \in i(p)$
- for each $\phi \in \mathcal{L}$ and $\alpha \in \mathcal{L}$, $V(\phi, \alpha) = 1$ iff $\alpha = \phi$
- for each $\phi \in \mathcal{L}$ and $\langle \alpha, w\rangle \in (\mathcal{L} \times P)$, $V(\phi, \langle \alpha, w\rangle) = 1$ iff $\alpha \boxright \phi \in \Phi_w$
- f is any selection function with the following property: if $X \cap \mathcal{L} = \{\alpha\}$ and $w \in P$, then $f(X, w) = \{\langle \alpha, w\rangle\}$.

It is easy to establish (i) by induction. As for (ii), note that $[\![\alpha]\!]^{\mathcal{I}} \cap \mathcal{L} = \{\alpha\}$, so $f([\![\alpha]\!]^{\mathcal{I}}, w) = \{\langle \alpha, w\rangle\}$. Hence, $\mathcal{I}, w \Vdash_{\mathsf{i}} \alpha \boxright \beta$ iff $\mathcal{I}, \langle \alpha, w\rangle \Vdash_{\mathsf{i}} \beta$, i.e.,

$V(\beta, \langle \alpha, w \rangle) = 1$, which holds iff $\alpha \boxrightarrow \beta \in \Phi_w$. ■

Corollary A.5. Let θ be any consistent **S5**-formula, and let θ^* be the result of simultaneously uniformly substituting one or more atomic sentences in θ for distinct counterfactuals. Then θ^* is i-satisfiable.

Proof: Let $q_1, \ldots, q_n$ be the atomics in θ that are substituted for distinct counterfactuals $\psi_1, \ldots, \psi_n$ resulting in θ^*. Since θ is consistent, it is **S5**-satisfiable. Let $\mathcal{M}, w \Vdash_{\mathbf{S5}} \theta$. For each $v \in W^{\mathcal{M}}$, define:

$$\Phi_v := \{\psi_i \mid \mathcal{M}, v \Vdash_{\mathbf{S5}} q_i\}$$

By Proposition A.4, this guarantees us an **S5**-equivalent impossible worlds model $\mathcal{I}$ such that $\mathcal{I}, v \Vdash_{\mathsf{i}} \psi$ iff $\psi \in \Phi_v$ where ψ is a counterfactual. Moreover, in this model, $\mathcal{I}, v \Vdash_{\mathsf{i}} \Box(q_i \equiv \psi_i)$. And since $\mathcal{I}, w \Vdash_{\mathsf{i}} \theta$, it follows that $\mathcal{I}, w \Vdash_{\mathsf{i}} \theta^*$. ■

Corollary A.5 effectively says that there are no non-trivial valid inferences governing counterfactuals in the impossible worlds semantics: any inference with counterfactuals that's i-valid is already **S5**-valid.

Theorem A.1(b) is harder to establish. The main issue is that while hyperconventions are allowed to redefine the semantic value of the boolean connectives, they cannot touch the semantics of $\boxrightarrow$. But in the impossible worlds semantics, any set of $\mathcal{L}$-formulas is satisfied at some (perhaps impossible) world in some model, including those containing counterfactuals. Thus, if we are to establish Theorem A.1(b), we need to establish the expressivist analogue of Proposition A.4. Indeed, this can be done, though the proof is more involved.

Proposition A.6. Let $\mathcal{M} = \langle W, i \rangle$ be an **S5**-model and let $\Phi\colon W \to \wp(\mathcal{L})$ map every $w \in W$ to a set Φ_w of counterfactuals. Then there is a expressivist model $\mathcal{E} = \langle W, f \rangle$ and a classical hyperconvention c such that for any $w \in W$:

(i) if ϕ is an **S5**-formula, then $\mathcal{E}, w, c \Vdash_{\mathsf{e}} \phi$ iff $\mathcal{M}, w \Vdash_{\mathbf{S5}} \phi$

(ii) if ψ is a counterfactual, then $\mathcal{E}, w, c \Vdash_{\mathsf{e}} \psi$ iff $\psi \in \Phi_w$.

Proof: Since **S5** is invariant under bisimulation contraction (and so, invariant under duplication of worlds), we may assume WLOG that W is infinite. We define c simply as the classical hyperconvention over W where $c(p) = i(p)$ for all $p \in \mathtt{At}$.

We now set out to define f. Fix an arbitrary $w_0 \in W$. Let $h\colon \mathcal{L} \to$

$W - \{w_0, w\}$ be a bijection. We'll write w_ϕ in place of $h(\phi)$ throughout. Now, let $\Gamma \subseteq \mathcal{L}$. Define the hyperconvention c_Γ as follows (where $\star \in \{\neg, \Box, \Diamond\}$ and $\circ \in \{\wedge, \vee, \rightarrow\}$):

$$c_\Gamma(p) = \begin{cases} \{w_p, w_0\} & \text{if } p \in \Gamma \\ \{w_p\} & \text{otherwise} \end{cases}$$

$$c_\Gamma(\star)(X) = \begin{cases} \{w_{\star\phi} \mid w_\phi \in X\} \cup \{w_0\} & \text{if } \star\phi \in \Gamma \text{ whenever } w_\phi \in X \\ \{w_{\star\phi} \mid w_\phi \in X\} & \text{otherwise} \end{cases}$$

$$c_\Gamma(\circ)(X, Y) = \begin{cases} \{w_{\phi\circ\psi} \mid w_\phi \in X \text{ and } w_\psi \in Y\} \cup \{w_0\} & \text{if } \phi \circ \psi \in \Gamma \text{ whenever } w_\phi \in X \text{ and } w_\psi \in Y \\ \{w_{\phi\circ\psi} \mid w_\phi \in X \text{ and } w_\psi \in Y\} & \text{otherwise} \end{cases}$$

Let $\Gamma^\alpha = \{\beta \mid \alpha \,\Box\!\!\rightarrow \beta \in \Gamma\}$. Define f as follows:

$$f(X, w_0, c_\Gamma) = \{\langle w_0, c_{\Gamma^\alpha}\rangle \mid \langle w_\alpha, c_\Gamma\rangle \in X\}$$

$$f(X, w_\phi, c_\Gamma) = \begin{cases} \{\langle w_\beta, c_\Gamma\rangle\} & \text{if } \phi = \alpha \,\Box\!\!\rightarrow \beta \text{ and } \langle w_\alpha, c_\Gamma\rangle \in X \\ I_W & \text{otherwise} \end{cases}$$

$$f(X, w, c) = \{\langle w_0, c_{\Phi_w^\alpha}\rangle \mid \langle w_\alpha, c_{\Phi_w}\rangle \in X\}$$

$$f(X, w, d) = X \text{ for any other } d.$$

Let $\mathcal{E} = \langle W, f\rangle$. It is easy to check that (i) holds by induction. So we just need to establish (ii). First, some intermediate claims:

Claim (*1*): For any Γ and any ϕ, ψ: $\mathcal{E}, w_\phi, c_\Gamma \Vdash \psi$ iff $\phi = \psi$.

Proof: By induction. The atomic case holds by definition of c_Γ. The cases for the connectives is straightforward. For the counterfactual, $\mathcal{E}, w_\phi, c_\Gamma \Vdash \alpha \,\Box\!\!\rightarrow \beta$ iff $f([\![\alpha]\!]^{\mathcal{E}}, w_\phi, c_\Gamma) \subseteq [\![\beta]\!]^{\mathcal{E}}$. By induction hypothesis, $\langle w_\gamma, c_\Gamma\rangle \in [\![\beta]\!]^{\mathcal{E}}$ iff $\gamma = \beta$. Hence, $[\![\beta]\!]^{\mathcal{E}} \neq I_W$, which means $f([\![\alpha]\!]^{\mathcal{E}}, w_\phi, c_\Gamma) \subseteq [\![\beta]\!]^{\mathcal{E}}$ iff $f([\![\alpha]\!]^{\mathcal{E}}, w_\phi, c_\Gamma) = \{\langle w_\beta, c_\Gamma\rangle\}$, which holds iff $\phi = \alpha \,\Box\!\!\rightarrow \beta$ and $\langle w_\alpha, c_\Gamma\rangle \in [\![\alpha]\!]^{\mathcal{E}}$. But again by induction hypothesis, $\langle w_\alpha, c_\Gamma\rangle \in [\![\alpha]\!]^{\mathcal{E}}$. Thus, $\mathcal{E}, w_\phi, c_\Gamma \Vdash \alpha \,\Box\!\!\rightarrow \beta$ iff $\phi = \alpha \,\Box\!\!\rightarrow \beta$. ■

Claim (2): For any Γ and any ϕ: $\mathcal{E}, w_0, c_\Gamma \Vdash \phi$ iff $\phi \in \Gamma$.

Proof: By induction. The atomic case holds by definition of c_Γ. The cases for the connectives is straightforward using Claim (1). For the counterfactual, $\mathcal{E}, w_0, c_\Gamma \Vdash \alpha \boxdot\!\!\rightarrow \beta$ iff $f([\![\alpha]\!]^{\mathcal{E}}, w_0, c_\Gamma) \subseteq [\![\beta]\!]^{\mathcal{E}}$. By Claim (1), $\langle w_\gamma, c_\Gamma\rangle \in [\![\alpha]\!]^{\mathcal{E}}$ iff $\gamma = \alpha$. So $f([\![\alpha]\!]^{\mathcal{E}}, w_0, c_\Gamma) = \{\langle w_0, c_{\Gamma^\alpha}\rangle\}$. Hence, $\mathcal{E}, w_0, c_\Gamma \Vdash \alpha \boxdot\!\!\rightarrow \beta$ iff $\mathcal{E}, w_0, c_{\Gamma^\alpha} \Vdash \beta$. But again by induction hypothesis, this holds iff $\beta \in \Gamma^\alpha$, i.e., $\alpha \boxdot\!\!\rightarrow \beta \in \Gamma$. ■

We are now ready to prove (ii). $\mathcal{E}, w, c \Vdash \alpha \boxdot\!\!\rightarrow \beta$ iff $f([\![\alpha]\!]^{\mathcal{E}}, w, c) \subseteq [\![\beta]\!]^{\mathcal{E}}$. By Claim (1), $\langle w_\gamma, c_{\Phi_w}\rangle \in [\![\alpha]\!]^{\mathcal{E}}$ iff $\gamma = \alpha$. Hence, $f([\![\alpha]\!]^{\mathcal{E}}, w, c) = \{\langle w_0, c_{\Phi_w^\alpha}\rangle\}$. So $\mathcal{E}, w, c \Vdash \alpha \boxdot\!\!\rightarrow \beta$ iff $\mathcal{E}, w_0, c_{\Phi_w^\alpha} \Vdash \beta$, which by Claim (2) holds iff $\beta \in \Phi_w^\alpha$, i.e., $\alpha \boxdot\!\!\rightarrow \beta \in \Phi_w$. ■

Corollary A.7. Let θ be any consistent **S5**-formula, and let θ^* be the result of uniformly substituting one or more atomic sentences in θ for distinct counterfactuals. Then θ^* is e-satisfiable.

Now we can establish Theorem A.1(b):

Proof (*Theorem A.1(b)*): Let $\mathcal{I} = \langle W, P, f, V\rangle$ and first let $w \in P$. Let:

$$\begin{aligned} \Phi &= \{\phi \mid \phi \text{ is an } \mathbf{S5}\text{-formula and } \mathcal{I}, w \Vdash_{\mathsf{i}} \phi\} \\ \Psi &= \{\phi \mid \phi \text{ is a counterfactual and } \mathcal{I}, w \Vdash_{\mathsf{i}} \phi\}. \end{aligned}$$

By Proposition A.6, there is a expressivist model $\mathcal{I}^{\mathsf{e}} = \langle W, f^{\mathsf{e}}\rangle$ and a classical hyperconvention c such that $\mathcal{I}^{\mathsf{e}}, w, c \Vdash_{\mathsf{e}} \Phi \cup \Psi$ and if ϕ is a counterfactual not in Ψ, $\mathcal{I}^{\mathsf{e}}, w, c \nVdash_{\mathsf{e}} \phi$. Hence, by a simple induction, $\mathcal{I}, w \Vdash_{\mathsf{i}} \phi$ iff $\mathcal{I}^{\mathsf{e}}, w, c \Vdash_{\mathsf{e}} \phi$.

Now let $w \notin P$. Let $\Gamma = \{\phi \mid V(\phi, w) = 1\}$ and let $\mathcal{I}^{\mathsf{e}}$ be $\langle W, f\rangle$ where f is constructed as in Proposition A.6. Then by Claim (2), $\mathcal{I}^{\mathsf{e}}, w_0, c_\Gamma \Vdash_{\mathsf{e}} \phi$ iff $\phi \in \Gamma$. Hence, we can take $s = \langle w_0, c_\Gamma\rangle$. ■

References

Anderson, Alan R. and Belnap, Nuel D. 1975. *Entailment: The Logic of Relevance and Neccessity*, volume 1. Princeton University Press.

Areces, Carlos and ten Cate, Balder. 2007. "Hybrid Logics." In Patrick Blackburn, Frank Wolter, and Johan van Benthem (eds.), *Handbook of Modal Logic*, 821–868. Elsevier.

Ayer, Alfred Jules. 1952. *Language, Truth and Logic*. New York: Dover Publications, Inc.

Barker, Chris. 2002. "The Dynamics of Vagueness." *Linguistics and Philosophy* 25:1–36.

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

Bernstein, Sara. 2016. "Omission impossible." *Philosophical Studies* 1–15.

Berto, Francesco. 2017. "Impossible Worlds and the Logic of Imagination." *Erkenntnis* 82:1277–1297.

Berto, Francesco, French, Rohan, Priest, Graham, and Ripley, David. 2018. "Williamson on Counterpossibles." *Journal of Philosophical Logic* 47:693–713.

Briggs, R. A. 2012. "Interventionist Counterfactuals." *Philosophical Studies* 160:139–166.

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

Burgess, Alexis and Plunkett, David. 2013a. "Conceptual Ethics I." *Philosophy Compass* 8:1091–1101.

—. 2013b. "Conceptual Ethics II." *Philosophy Compass* 8:1102–1110.

Carnap, Rudolph. 1937. *The Logical Syntax of Language*. New York: Harcourt, Brace and Company.

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

—. 1990. "On What Cannot Be." In *Truth or Consequences*, 123–132. Dordrecht: Springer Netherlands.

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

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

Downing, P B. 1959. "Subjunctive Conditionals, Time Order, and Causation." *Proceedings of the Aristotelian Society* 59:125–140.

Einheuser, Iris. 2006. "Counterconventional Conditionals." *Philosophical Studies* 127:459–482.

Emery, Nina and Hill, Christopher S. 2016. "Impossible Worlds and Metaphysical Explanation: Comments on Kment's Modality and Explanatory Reasoning." *Analysis* 1–15.

Field, Hartry. 2009a. "Pluralism in Logic." *Review of Symbolic Logic* 2:342–359.

—. 2009b. "What is the Normative Role of Logic?" *Aristotelian Society Supplementary Volume* 83:251–268.

Fine, Kit. 1970. "Propositional Quantifiers in Modal Logic." *Theoria* 36:336–346.

Frege, Gottlob. 1892. "On Sense and Reference." *Zeitschrift für Philosophie und philosophische Kritik* 100:25–50. Translated in 1948, *The Philosophical Review*, 57(3): 209–230.

French, Rohan, Girard, Patrick, and Ripley, David. 2020. "Classical Counterpossibles." *Review of Symbolic Logic* 1–20.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

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

Gregory, Dominic. 2007. "Imagining possibilities." *Philosophy and Phenomenological Research* 69:327–348.

Hansen, Nat. 2019. "Metalinguistic Proposals." *Inquiry* 1–19. doi:10.1080/0020174X.2019.1658628.

Harman, Gilbert. 1984. "Logic and Reasoning." *Synthese* 60:107–127.

Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them To Be?" *Noûs* 34:31–55.

—. 2005. "What are we talking about? The semantics and politics of social kinds." *Hypatia* 20:10–26.

Jenny, Matthias. 2016. "Counterpossibles in Science: The Case of Relative Computability." *Noûs* 1–31.

Kaplan, David. 1977. "Demonstratives." In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, 481–563. Oxford: Oxford University Press.

Keefe, Rosanna. 2014. "What Logical Pluralism Cannot Be." *Synthese* 191:1375–1390.

Kennedy, Christopher and Willer, Malte. 2016. "Subjective Attitudes and Counterstance Contingency." In *Proceedings of SALT 26*, 913–933.

Kim, Seahwa and Maslen, Cei. 2006. "Counterfactuals as Short Stories." *Philosophical Studies* 129:81–117.

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

Kocurek, Alexander W. 2020. "Logic Talk." Manuscript.

Kocurek, Alexander W., Jerzak, Ethan J., and Rudolph, Rachel E. 2020. "Against Conventional Wisdom." *Philosophers' Imprint* 20:1–27.

Kouri Kissel, Teresa. 2018. "Logical Pluralism from a Pragmatic Perspective." *Australasian Journal of Philosophy* 96:578–591.

—. 2019. "Metalinguistic Negotiation and Logical Pluralism." *Synthese* 1–12.

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

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*. Cambridge, MA: Harvard University Press.

Kung, Peter. 2010. "Imagining as a Guide to Possibility." *Philosophy and Phenomenological Research* LXXXI:620–663.

—. 2016. "You Really Do Imagine It: Against Error Theories of Imagination." *Noûs* 50:90–120.

Lewis, David K. 1973. *Counterfactuals*. Blackwell Publishing.

Locke, Theodore. 2019. "Counterpossibles for Modal Normativists." *Synthese* 1–23.

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

MacFarlane, John. 2004. "In What Sense (If Any) Is Logic Normative for Thought?" Unpublished.

—. 2016. "Vagueness as Indecision." *Aristotelian Society Supplementary Volume* 90:255–283.

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

—. 2012. "Relevance Logic." *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/logic-relevance/.

McConnell-Ginet, Sally. 2006. "Why defining is seldom 'just semantics': Marriage and *marriage*." In Betty Birner and Gregory Ward (eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, 223–246. Amsterdam: John Benjamins.

—. 2008. "Words in the world: How and why meanings can matter." *Language* 83:497–527.

Muskens, Reinhard A. 1991. "Hyperfine-grained meanings in classical logic." *Logique et Analyse* 34:159–176.

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

Plunkett, David. 2015. "Which Concepts Should We Use?: Metalinguistic Negotiations and The Methodology of Philosophy." *Inquiry* 58:828–874.

Plunkett, David and Sundell, Tim. 2013. "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosopher's Imprint* 13:1–37.

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

—. 2005. *The Logic of Scientific Discovery*. Routledge.

Priest, Graham. 2008. "Logical Pluralism Hollandaise." *Australasian Journal of Logic* 6:210–214.

Putnam, Hilary. 1969. "Is logic empirical?" In *Boston Studies in the Philosophy of Science*, 216–241. Springer, Dordrecht.

Quine, Willard van Orman. 1970. "Deviant Logics." In *Philosophy of Logic*, 80–94.

Read, Stephen. 2006. "Monism: The One True Logic." In D. de Vidi and T. Kenyon (eds.), *A Logical Approach to Philosophy: Essays in Memory of Graham Solomon*. Springer.

Ripley, David. 2016. "Experimental Philosophical Logic." In Justin Sytsma (ed.), *A Companion to Experimental Philosophy*, chapter 36, 523–534. Wesley Buckwalter.

Russell, Gillian. 2008. "One True Logic?" *Journal of Philosophical Logic* 37:593–611.

Sandgren, Alexander and Tanaka, Koji. 2019. "Two Kinds of Logical Impossibility." *Noûs* 1–12. doi:10.1111/nous.12281.

Stalnaker, Robert C. 1978. "Assertion." In *Syntax and Semantics*, 315–332. Oxford University Press.

Steinberger, Florian. 2019a. "Logical Pluralism and Logical Normativity." *Philosophers' Imprint* 19:1–19.

—. 2019b. "Three Ways in Which Logic Might Be Normative." *Journal of Philosophy* 116:5–31.

van Inwagen, Peter. 1998. "Modal Epistemology." *Philosophical Studies* 92:67–84.

Vander Laan, David A. 2004. "Counterpossibles and Similarity." In Frank Jackson and Graham Priest (eds.), *Lewisian Themes*.

Vetter, Barbara. 2016. "Counterpossibles (not only) for dispositionalists." *Philosophical Studies* 173:2681–2700.

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

—. 2009. "Probability and Danger." *The Amherst Lecture in Philosophy* 4:1–35.

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

Yablo, Stephen. 1993. "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53:1–42.