Against Conventional Wisdom

Forthcoming in *Philosophers' Imprint*
May 2020 Preprint

Alexander W. Kocurek
Ethan Jerzak
Rachel Etta Rudolph

Abstract. Conventional wisdom has it that truth is always evaluated using our actual linguistic conventions, even when considering counterfactual scenarios in which different conventions are adopted. This principle has been invoked in a number of philosophical arguments, including Kripke's defense of the necessity of identity and Lewy's objection to modal conventionalism. But it is false. It fails in the presence of what Einheuser (2006) calls *c-monsters*, or convention-shifting expressions (on analogy with Kaplan's *monsters*, or context-shifting expressions). We show that c-monsters naturally arise in contexts, such as metalinguistic negotiations, where speakers entertain alternative conventions. We develop an expressivist theory—inspired by Barker (2002) and MacFarlane (2016) on vague predications and Einheuser (2006) on counterconventionals—to model these shifts in convention. Using this framework, we reassess the philosophical arguments that invoked the conventional wisdom.

1 The Conventional Wisdom

There is a famous riddle: If you called a dog's tail a leg, how many legs would a dog have? The answer is four, of course! Calling a tail a leg doesn't make it one.

This answer reflects a pervasive view about how to evaluate truth at hypothetical scenarios: truth is always assessed using our actual linguistic conventions, not those of any hypothesized speakers. The following conditionals illustrate the idea:

(1) If 'water' referred to gasoline, water would fuel fire.

(2) If 'water' referred to gasoline, 'water fuels fire' would be true.

(2) is uncontroversial; we could have spoken differently, and if we had, the sentence 'water fuels fire' would express a true proposition. (1), on the other hand, seems clearly false; speaking differently would not change the laws of chemistry.

The reason for this difference is that we describe counterfactual scenarios with the language we *actually* speak, not the language we would have spoken in that scenario. To invoke a familiar distinction, we are *using* the term ‘water’ in the consequent of (1), whereas we are *mentioning* it in that of (2). As such, ‘water’ in the consequent of (1) refers to what we actually use ‘water’ to refer to, i.e., the liquidy substance that tends to put out fires, whereas in the quotational context of (2), it refers to a word, which could have been interpreted in any number of ways.

This observation is captured by the following general principle:[1]

Conventional Wisdom. Truth at a scenario (counterfactual or otherwise) is evaluated relative to our (or the speaker’s) actual linguistic conventions, even if those conventions diverge from the ones adopted in that scenario.

So when evaluating the truth of a counterfactual of the form $A \mathrel{\Box\!\!\rightarrow} C$, we hold fixed our actual interpretation of A and C and assess whether C, as we actually interpret it, holds on the counterfactual supposition that A, as we actually interpret it.

Conventional Wisdom is not a mere matter of pedantry: it has been wielded in a number of serious philosophical arguments. Kripke (1980) uses Conventional Wisdom to diffuse an argument against the necessity of identity. That argument goes like this. Recall that ‘Hesperus’ is a name for the evening star, i.e., the first celestial body visible after sunset, and ‘Phosphorus’ is a name for the morning star, i.e., the last celestial body visible before sunrise. As a matter of fact, Venus is both the evening star and the morning star. Thus, Hesperus is actually identical to Phosphorus. Imagine a world, however, where Mars is the first celestial body visible after sunset, while Venus is still the last celestial body visible before sunrise. It does not seem true to say that Hesperus is identical to Phosphorus in this scenario. So, one might think, there is a possible world where Hesperus is not identical to Phosphorus—thereby showing that identity is contingent.

Not so, Kripke objects: this is only a world where we use language differently, i.e., where *the names* ‘Hesperus’ and ‘Phosphorus’ are used to refer to different things. It is not a world where *the objects* Hesperus and Phosphorus are distinct—that is, a world where Venus is distinct from itself. *Even if speakers had used the names ‘Hesperus’ and ‘Phosphorus’ to refer to different objects, Hesperus would still be Phosphorus.* Kripke defends this by appeal to Conventional Wisdom:

[1] This principle is what Wright (1985, p. 190) calls “Convention C”.

> [W]hen we speak of a counterfactual situation, we speak of it in English, even if it is part of the description of that counterfactual situation that we were all speaking German in that counterfactual situation. We say, 'suppose we had all been speaking German' or 'suppose we had been using English in a nonstandard way'. Then we are describing a possible world or counterfactual situation in which people, including ourselves, did speak in a certain way different from the way we speak. But still, in describing that world, we use *English* with *our* meanings and *our* references. (p. 77)

The world so-described is not a world where Hesperus *as we use the term* is distinct from Phosphorus *as we use the term*. So thanks to Conventional Wisdom, this thought experiment does not refute the necessity of identity.

Here is a second demonstration of the power of Conventional Wisdom. Certain logical positivists once thought that necessity and contingency were grounded in linguistic conventions. This view is known as **modal conventionalism**. As Ayer (1952, p. 31) puts it, necessary truths "simply record our determination to use symbols in a certain fashion." Modal conventionalism was attractive in part because it promised to explain how we could come to know necessary truths without abandoning a broadly empiricist epistemology: we come to know whether something is necessary or contingent simply by leveraging our linguistic competence and reflecting on how we use words.

But there was a problem. Lewy (1976) argued that modal conventionalism makes false predictions about the modal connection between our conventions and necessary truths. While it is contingent what linguistic conventions we adopt, it is not contingent whether $68 + 57 = 125$—it is not as though we could have made $68+57 = 5$ by resolving to use '+' differently. *No matter how we spoke, it would still be true that* $68 + 57 = 125$. The modal conventionalist seems to deny this, attributing to us awesome mathematical powers. Their mistake is to conflate how we *could* use the term '+' with how we *actually* use it. They use the hypothetical interpretation of '+' in evaluating $68 + 57 = 125$ in the counterfactual scenario—not, as Conventional Wisdom requires, the actual one.[2]

Conventional Wisdom has been sharpened into a principle put forth by Einheuser (2006, pp. 473–475). This principle concerns what she calls **c-monsters**, i.e., expressions that shift the conventions used to interpret an-

[2] See Wright 1985, p. 192 for a response to this objection on behalf of the modal conventionalist.

other (usually embedded) expression. Einheuser argues that Conventional Wisdom supports the following conjecture:

No C-Monsters. There are no c-monsters in English.

In particular, she claims that counterfactual conditionals in English are never interpreted as **counterconventionals**, i.e., as counterfactuals that shift the interpretation of expressions in its constituents. This is why, for instance, (1) seems false; if (1) were interpreted as a counterconventional, it would be true. Conventional Wisdom rules out counterfactual c-monsters.

The notion of a c-monster is related to, but distinct from, Kaplan's notion of a **monster**. Both monsters and c-monsters involve shifting the content of an expression. But these shifts are achieved via different mechanisms: monsters shift the *context* used to determine content (Kaplan, 1977, p. 499), whereas c-monsters shift the *way* context determines content. In Kaplanian terms, monsters are context-shifting expressions, whereas c-monsters are *character*-shifting expressions.

Kaplan famously conjectured that there are no monsters in English.[3] But even if there were monsters, that would not imply the existence of c-monsters. Monsters only affect expressions, such as indexicals or demonstratives, whose content depends on the context of use. Expressions that are not context-sensitive in this sense are not affected by monsters. By contrast, c-monsters can affect the content of non-context-sensitive expressions. Einheuser's conjecture is primarily concerned with c-monsters of this sort, capable of shifting the content even of non-context-sensitive expressions like 'mountain', 'vegetable', and 'planet'. Thus, Einheuser's conjecture is distinct from Kaplan's and deserves independent investigation.[4]

Einheuser defends No C-Monsters and Conventional Wisdom as unsurprising features of natural language:

> The claim that there are no *c*-monsters in English, as well as the claim that [Conventional Wisdom] governs our counterfactual reasoning, is not at all implausible. We judge a situation, actual or counterfactual, against the conceptual background of our actual conceptual practices, simply because it is the deeply entrenched actual conceptual practices (of which we may or may not be aware) which inform our judgements. There is no mystery here. (p. 474)

[3] This claim has undergone intense scrutiny; see for instance Schlenker 2003; Anand and Nevins 2004; Santorio 2012; Rabern 2012; Yli-Vakkuri 2013; Rabern and Ball 2019; Deal 2019.

[4] Conversely, the existence of c-monsters does not obviously imply the existence of monsters. Since we are not primarily concerned with monsters, though, we set this aside.

Einheuser does not deny that one can understand the counterconventional interpretations of counterfactuals. In fact, she argues that conventionalist views in metaphysics (views which hold that truths about metaphysical notions, like existence or modality, are grounded in conventions) ought to be interpreted as merely making counterconventional claims. But in doing so, conventionalists are not speaking ordinary English: they are instead attempting to use language in a novel way.[5]

We have two main goals in this paper. The first is to refute Conventional Wisdom—that is, to show that in many ordinary uses of language, the conventions used to interpret linguistic expressions are not those of the speaker or the assessor. We do so by showing that many c-monsters lurk in the seas of natural language. The second is to develop a systematic theory that accounts for the presence and behavior of c-monsters by incorporating *conventions* into more traditional semantic and pragmatic frameworks. Our broadly expressivist theory captures how we communicate both about the conventions we in fact adopt (e.g., in metalinguistic negotiations) and about possible alternative conventions we might have adopted (with c-monsters).

The plan is as follows. In §2, we present examples of ordinary conversations where Conventional Wisdom is violated. We show that counterfactuals, tenses, attitude verbs, and other expressions can be c-monsters. Our framework for dealing with this behavior involves both pragmatic and semantic innovations. In §3, we develop a generalized Stalnakerian theory of communication that takes the common ground in a conversation to be modeled as a set of world-convention pairs. On this view, assertions don't simply communicate factual matters, but also serve to express speakers' commitments to use language in a certain way. In §4, we develop a semantics for c-monsters, which explains their ability to shift the interpretation of terms, and thus communicate about alternative convention choices, through the introduction of a convention parameter in the index of evaluation. In §5, we return to the philosophical arguments mentioned above that invoked Conventional Wisdom, and reevaluate them in light of its failure. In §6, we conclude with some reflections about what our account means for the nature of language and linguistic competence.

[5] Einheuser's settled view on No C-Monsters is somewhat unclear. In the quoted passage, she seems to be defending No C-Monsters in full, but then in the next paragraph (p. 475), she hedges, saying that we are "ordinarily not aware of the availability of the counterconventional reading" and that counterconventionals are "prima facie less natural" than the ordinary reading of counterfactuals, suggesting that she might be willing to concede that counterconventional readings of counterfactuals exist in natural language but are simply harder to hear.

2 C-Monsters Exist

In this section, we argue against Conventional Wisdom by presenting perfectly natural conversations containing c-monsters (§§2.1–2.2). We also address some potential objections (§§2.3–2.5).

2.1 Bring Pluto Back!

In 2006, the International Astronomical Union (IAU) convened to reconsider the scientific definition of a planet. While Pluto was officially classified as a planet at the time, astronomers were aware of objects in the solar system that had an equally good claim to planethood, including Ceres (a large celestial object in the asteroid belt) and some trans-Neptunian objects (Eris, Haumea, and Makemake). Thus, according to the original definition of the term, the solar system contained more planets than the nine that had been officially recognized. And there was no reason to suspect the list would not continue to grow even more with future astronomical discoveries.

Ultimately, the IAU decided to revise the definition of 'planet' so as to require all planets to "clear their orbital neighborhood", meaning they had to be substantially more massive than anything else in the general vicinity of their orbit. Since Pluto's orbit crosses Neptune's, and Neptune is 10,000 times more massive than Pluto, Pluto failed to meet this condition and so was classified as a dwarf planet rather than as a planet. Other objects (Ceres, Eris, etc.) were classified as dwarf planets for similar reasons.

The public outcry was immense. People took to the streets after the IAU announced that Pluto had been downgraded. They made T-shirts proclaiming, "Bring Pluto back!" and "Pluto is still a planet in my heart".[6] State legislatures in Illinois and California reacted critically, while New Mexico's went as far as declaring Pluto to be a planet whenever it "passes overhead through New Mexico's excellent night skies".[7] Even amongst astronomers, the decision was controversial. Some scientists argued that expressions such as 'general vicinity' and 'significantly more massive' rendered the definition too vague for scientific purposes. Other scientists, albeit a minority, thought the new definition was fine but argued it would have been more convenient for Pluto to have been grandfathered in as a planet.[8]

Consider now the following (not so) hypothetical dialogue:

[6] Still available for purchase online as of this writing.

[7] See State of Illinois 2009; DeVore 2006; State of New Mexico 2007.

[8] For an engaging account of the controversy, see Brown 2010.

(3) **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, there would be dozens of planets in the solar system.

Focus on Beta's last response, which takes the form of a counterfactual:

(4) If Pluto were a planet, there would be dozens of planets in the solar system.

In the given context, (4) is most naturally interpreted as a claim not about what would happen if Pluto cleared its orbital neighborhood but about what would happen if Pluto were classified differently. That is, one hears (4) more along the lines of (5-a) than (5-b):

(5) a. If Pluto were classified as a planet, dozens of other objects in the solar system would be too.

b. If Pluto cleared its orbital neighborhood, dozens of other objects in the solar system would too.

Moreover, what Beta is saying with (4) seems right. The reason for revising the definition was precisely to avoid the proliferation of objects with equal claim to planethood as Pluto. So if Pluto counted as a planet, so would Ceres, Eris, and so on.

For Beta's assertion to be interpreted in this way, the word 'planet' has to be interpreted differently from how Beta interprets it. After all, Beta *agrees* with the IAU on how to use the term 'planet'. And according to the IAU, planets *by definition* must clear their orbital neighborhood. Hence, if we interpreted 'planet' using the linguistic conventions Beta adopts, (4) would have the force of (5-b)—it would suggest that a change in Pluto's orbit would result in significant changes to the orbits of Ceres, Eris, and so on. Not only is that probably false, but it is also uncharitable as an interpretation of what Beta is saying. In order to obtain the correct interpretation of (4), 'planet' must be interpreted relative to conventions other than Beta's own.

Indeed, even if *we*, the assessors, agree with the IAU about how to use the word 'planet', we can still interpret Beta's use of (4) in the way it was intended, i.e., as a claim about Pluto's possible classification rather than about Pluto's possible orbit. So we still have to reinterpret 'planet' in the consequent of (4) to obtain the correct interpretation of the counterfactual.[9] Thus, (4) violates Conventional Wisdom.

Beta's final assertion contains another violation of Conventional Wisdom. Beta ascribes to Alpha a belief that Pluto is a planet:

(6) Alpha thinks that Pluto is a planet.

According to Conventional Wisdom, 'planet' in (6) must be interpreted according to the IAU's definition, not Alpha's conventions, in which case (6) would imply that Alpha thinks Pluto clears its orbital neighborhood. But it is common ground here that Alpha knows that Pluto does not clear its orbital neighborhood. Thus, (6) would be false if Conventional Wisdom were true: it would incorrectly ascribe to Alpha a false empirical belief about Pluto's orbit, rather than a controversial but sensible view about how to define 'planet'.

We think that (4) and (6) have true readings and thus violate Conventional Wisdom.[10] Moreover, (4) and (6) seem to contain c-monsters. In (4), it is the counterfactual context that induces a shift in the conventions used to interpret 'planet'. Without the embedded context, 'planet' must be interpreted using the conventions the speaker actually adopts. For consider Alpha's earlier claim:

(7) Pluto is a planet.

Alpha can felicitously assert this because Pluto counts as a planet on their preferred classification. But Beta, who adopts the IAU's definition excluding

[9] As we use the terms here, to change how one *classifies* things as planets is to change the *conventional interpretation* of 'planet'. We are not committed to the converse, however. We could change the conventions associated with 'planet' so that it refers to vegetables, for instance, without in any interesting sense changing how we classify things as planets (e.g., we would not be revising the IAU's classification but simply co-opting the word for other ends). Indeed, talk of "classifying" in natural language is quite complex and could use more investigation. We thank an anonymous reviewer for pressing us to think more carefully about the relationship between classification and conventions.

[10] We do not deny that they also have false readings. Depending on the context, (4) can be interpreted factually, i.e., holding fixed what Beta means by 'planet'. On this interpretation, (4) sounds more like (5-b), which is clearly false. Similarly, there are contexts in which (6) would be heard as false. One might use (6) precisely to attribute to Alpha the mistaken belief that Pluto fits the IAU's criteria for planethood. (In this case, it would be natural to describe 'planet' as being read *de re*.) We only hold that (4) and (6) are not always used in these ways. It is their true readings, incompatible with Conventional Wisdom, that we aim to account for.

Pluto from planethood, cannot felicitously assert the same sentence (unless, of course, they decide to adopt different conventions for the purposes of conversation). It is only inside the scope of an embedding expression, such as a counterfactual, that ‘planet’ gets (re)interpreted according to conventions that Beta does not adopt. This strongly suggests that it is the counterfactual environment that triggers the shift in the conventional interpretation of ‘planet’—that is, that (4) contains a c-monster.

2.2 Other C-Monsters

Violations of Conventional Wisdom are not isolated to counterfactuals and belief reports. C-monsters occur in plenty of other embedding environments, including tense, circumstantial modality, and other attitude verbs.

First, tense. Suppose it is common knowledge that Pluto’s orbit has not changed and Pluto has never cleared its orbital neighborhood. Even so, it would be appropriate for Beta to say:

(8) a. Pluto used to be a planet, but it isn’t any more.
b. Pluto will no longer be a planet once the IAU’s decision goes into effect.

(8-a) conveys that Pluto’s classification has changed over time: in the past, Pluto was classified as a planet, but no more. Similarly, (8-b) conveys that Pluto’s classification, not its orbit, will change. If we held fixed what Beta currently means by ‘planet’, these interpretations would be unavailable. So just like (4), the examples in (8) violate Conventional Wisdom.

Circumstantial (i.e., “metaphysical”) modals also give rise to c-monstrous interpretations. Suppose it was the vote of a single stubborn scientist that determined how Pluto would be classified. Beta can then say:

(9) a. Pluto could have easily been a planet. But that one stubborn scientist voted for the current definition, so it is not.
b. That one stubborn scientist can still make Pluto a planet again. They just need to rescind their vote.

Other metaphysically loaded words, such as ‘depends’, ‘explains’, and ‘because’, give rise to similar effects:[11]

(10) a. Whether or not Pluto is a planet depends on what definition the members of the IAU agree on.

[11] Thanks to Zoltán Szabó for pointing this out.

b. Part of what explains why Pluto is not a planet is the IAU's decision in 2006 to redefine 'planet'.
c. Because of the IAU's decision in 2006, Pluto is not a planet.

Each of the sentences in (9) and (10) seems true on its intended interpretation. For (9-a) says that Pluto could have easily been *classified* as a planet, (9-b) that Pluto can still be classified as a planet, (10-a) that how Pluto is classified depends on the IAU's decision, and so on. But again, if we held fixed what Beta means by 'planet', none of these sentences would receive plausible readings.

Not all metaphysically loaded words can be c-monsters. A possible exception is 'cause'. It is hard to hear the following as true:

(11) ??The IAU caused Pluto to be a dwarf planet.

Why is 'cause' different from the expressions in (10)? We are not entirely sure; certainly changes in classification can cause other things. But we need not belabor this question. Our main aim here is just to show that a wide variety of expressions can act as c-monsters, not that all do.

Returning to attitude verbs, 'thinks' is not alone in its potential to behave c-monstrously. Recall (6):

(6) Alpha thinks that Pluto is a planet.

The following can also be asserted felicitously, even by someone who adopts the IAU's definition:

(12) a. Beta wishes Pluto were a planet (but they couldn't care less about whether it clears its orbital neighborhood).
b. Gamma fears Pluto is a planet (though Gamma is hardly bothered by the possibility that it clears its orbital neighborhood).

In each case, the content of the attitude ascribed concerns whether Pluto counts as a planet, not Pluto's orbit. (6) says that Alpha counts Pluto as a planet, (12-a) says Beta wishes Pluto counted as a planet, and (12-b) says Gamma fears Pluto counts as a planet. To obtain these readings, we cannot interpret 'planet' according to the IAU's definition. (12-a) does not say, for instance, that Alpha wishes Pluto were a planet *as the IAU defines the term*. So each of these sentences violates Conventional Wisdom.

The attitude verb 'consider' is an interesting case with respect to c-monsters, since it can be used to express opinions about matters of convention but not purely factual beliefs. Thus, Beta can say:

(13) Alpha considers Pluto a planet (though they know Pluto does not clear its orbital neighborhood).

Again, we clearly will not have a plausible reading of this report if we hold fixed what we (or others who agree with the IAU) mean by 'planet'. Alpha does not consider Pluto a planet *in the IAU's sense*. They are instead advocating for a different understanding of 'planet'. So again, we have an example of a c-monster, thus violating Conventional Wisdom.

As Kennedy and Willer (2016, p. 915) have pointed out, 'consider' seems to require a certain classificatory contingency that other attitude verbs (such as 'think') do not. Witness the following contrast:

(14) a. #Alpha considers Earth a planet.
b. Alpha considers Pluto a planet.

Intuitively, (14-a) is marked because on every plausible way of classifying things, Earth will count as a planet. By contrast, (14-b) is fine because different plausible ways of classifying things differ in their verdicts about whether Pluto is a planet. (14-b) can be used to convey that Alpha endorses a way of understanding 'planet' on which Pluto counts, but this is only felicitous because there are other salient classifications on which it would not count. With 'think', however, both are felicitous:

(15) a. Alpha thinks Earth is a planet.
b. Alpha thinks Pluto is a planet.

Unlike 'consider', there is no requirement on 'think' that other salient classifications be available. So while (15-a) can be used merely to express a mundane factual belief, (14-a) must do more. (We will return to this observation in §3.)

Not all attitude verbs can be c-monsters. Factive attitude verbs, for instance, cannot be. In the context of (3), Beta cannot truly assert:

(16) #Alpha knows that Pluto is a planet.

This makes sense given that 'knows' is factive: if one accepts (16), one must accept that Pluto is a planet. So if one adopts an interpretation of 'planet' according to which Pluto does not count as a planet, then (16) should sound false.

There are also expressions that specify the conventions one is to use to evaluate another expression. Suppose Alpha specifically endorses the old definition of 'planet', but does not realize that Ceres meets that definition. One can still say:

(17) a. Ceres is a planet in Alpha's sense (though they do not realize it).
b. Ceres is a planet as Alpha defines it (though they do not realize it).

These qualification phrases 'in Alpha's sense' or 'as Alpha defines it' directly shift the interpretation of 'planet' to the conventions Alpha adopts, even if the speaker does not adopt those conventions.

We thus find that a wide variety of expressions can be c-monsters, thereby violating Conventional Wisdom.[12] In the next sections, we develop a broad framework that accommodates these readings by including conventions in the linguistic theory. The semantic machinery needed to interpret c-monstrous constructions will come in §4. Our account there will build on a pragmatic framework incorporating conventions, which we develop in §3. Before turning to that, however, we wish to forestall three potential responses to the data we have presented. (Readers who are already convinced may skip ahead to §3.)

2.3 Gricean Explanation

One could try to maintain Conventional Wisdom while explaining the conversation in (3) along the following lines. In the relevant context, it is common ground that (4) and (6) express literal falsehoods (or trivialities). Pragmatic repair along familiar Gricean lines then kicks into gear (Grice, 1975). Knowing that Beta did not intend to communicate something absurd or trivial, we understand them as trying to communicate something explicitly metalinguistic (modeled after (2)). For example, we might understand Beta's utterance of (4) as a non-standard attempt to say:

(18) If 'planet' were defined so as to include Pluto, then dozens of objects in the solar system would be called 'planets'.

The defender of this line owes us a theory of how these utterances are transformed into explicitly metalinguistic ones. We think that the prospects

[12] We do not claim our list here to be exhaustive. We suspect some of the following expressions can also be c-monsters, but we are not all in agreement about which: the attitude verbs 'doubt' and 'wish'; agentive 'can'; deontic modals, like 'ought'; and epistemic modals, like 'might' and 'probably'. We leave investigation of these to future work.

for such a theory are not good because the exact nature of the transformation into an explicitly metalinguistic sentence is highly unsystematic.

Take, for example, the belief report from (6). This would not be a good gloss on that sentence:

(19) Alpha thinks that ‘planet’ is defined so as to include Pluto.

For Alpha knows that this is not how the IAU defines it. A better gloss would be:

(20) Alpha thinks that ‘planet’ *should* be defined so as to include Pluto.

But we cannot similarly paraphrase (21-a) with (21-b):

(21) a. If Pluto were a planet, Alpha would be happy.
b. If ‘planet’ should have been defined so as to include Pluto, Alpha would be happy.

Alternatively, we could try to paraphrase (6) with:

(22) Alpha classifies Pluto as a ‘planet’.

But this leaves us unable to paraphrase more complex sentences. For instance, suppose Alpha is unsure how to define ‘planet’, but thinks that it should be defined so as to include Pluto if Pluto is at least 10^{26} kg. Then it sounds fine to say:

(23) Alpha thinks that either Pluto is a planet or Pluto is less than 10^{26} kg.

But we cannot simply paraphrase this as:

(24) #Alpha classifies Pluto as either a ‘planet’ or ‘less than 10^{26} kg’.

While we could keep exploring further epicycles of revisions to this strategy,[13] we think the difficulties here motivate pursuing a different approach—one that takes the data at face value.

[13] One possibility is to assimilate apparent c-monsters to the phenomenon of mixed quotation, or perhaps scare quotes. See, e.g., Shan 2010 and Maier 2014. Then again, we suspect our analysis of c-monsters might help in the analysis of these linguistic devices, though we leave this for future investigation. Thanks to Chris Barker, Dan López de Sa, and John MacFarlane for suggesting this.

2.4 Sameness of Meaning

One might think that Alpha and Beta actually agree on the meaning of 'planet' in (3) in that both are referring to the property of planethood, and instead just disagree over what makes something a planet, i.e., what planethood is.[14] (As an analogy, a utilitarian and a Kantian can disagree over what is right or wrong without meaning different things by the words 'right' and 'wrong'.) In this case, 'thinks' would not need to shift the interpretation of 'planet' in order for (6) to come out true.

While we find much to agree with here, we maintain this strategy will not save Conventional Wisdom, for two reasons. First, it already departs from Conventional Wisdom in an important way. Second, we do not think that, in disagreeing over what makes something a planet, Alpha and Beta are guaranteed to mean the same thing by 'planet'—it depends on what we mean by "meaning".

First: On the conventional view, the interpretation of 'planet' is an *intension*, i.e., a function from worlds to extensions. According to Conventional Wisdom, then, the intension associated with 'planet' in (6) is the intension the speaker (or perhaps the assessor) actually associates with it. Thus, the sentence 'Pluto is a planet', when Beta uses it (even in (6)), picks out the set of worlds where Pluto has the property that Beta actually associates with 'planet', viz., the set of worlds where Pluto has the right size and shape and also clears its orbital neighborhood.

But when Alpha and Beta disagree over what makes something a planet, they disagree over what intension to associate with 'planet', i.e., the rule for determining which objects at which worlds fall in the extension of 'planet'. They do not, therefore, pick out the same set of worlds with 'Pluto is a planet': when Alpha uses that sentence, it picks out a set that includes worlds where Pluto does not clear its orbital neighborhood. So if we want to maintain that Alpha and Beta "mean the same thing" by their use of 'planet'—if we want to preserve Conventional Wisdom or No C-Monsters in face of such examples—then we must adopt a notion of "meaning" or "conventional interpretation" that is not just a simple intension. This already strays from the original spirit of Conventional Wisdom and No C-Monsters.

This leads us to our second point: In what sense do Alpha and Beta "mean the same thing" by their use of 'planet' if they do not associate the same intension with it? We will provide one answer in §3 (see also footnote 29), but we do not claim that our answer is the only reasonable one. There are many things one could mean by "meaning", and it is best to just

[14] We thank an anonymous reviewer for pressing us to discuss this issue.

be clear about what notion one has in mind. The question is whether there is a notion of meaning that saves Conventional Wisdom without trivializing it. Everyone agrees that, in a trivial disquotational sense, both Alpha and Beta refer to planethood with ‘planet’, but this is compatible with different views about whether that’s sufficient for Alpha and Beta to truly mean the same thing.

For example, one could understand the “meaning” of ‘planet’ in terms of the role it plays in our linguistic community. Alpha and Beta might, for instance, agree that the conventions governing ‘planet’ should conform to the IAU’s definition, whatever that is, but Alpha mistakenly believes that this definition does not include the orbital neighborhood condition. Alpha and Beta then associate different intensions with ‘planet’, while nonetheless agreeing to define ‘planet’ as the IAU does.

We are happy to describe this as a case where speakers agree on what ‘planet’ means (in a sense), but not on what makes something a planet. Crucially, however, we do not think that *all* cases where speakers disagree over intension are cases where speakers agree on meaning in this sense. In the original set-up for (3), for instance, Alpha and Beta do not even agree on whether to follow the IAU’s definition of ‘planet’. Even if Alpha and Beta fully agree about which intensions play which roles, they may nevertheless disagree over what role to assign the term ‘planet’ in the first place. Thus, while disagreement over intension can result from a factual disagreement over what intension fills a certain role, it need not do so.[15]

2.5 Externalism

Perhaps, however, the intension associated with ‘planet’ is determined by external factors that are the same for both Alpha and Beta. For example, some externalists (following Putnam (1973)) hold that the meaning of ‘planet’ depends on causal or metaphysical factors. Others (following Burge (1979)) hold that the meaning of ‘planet’ depends on the wider linguistic community. In either case, Alpha and Beta would use ‘planet’ with the same intension, perhaps without realizing it. If so, then Alpha and Beta would pick out the same set of worlds with ‘Pluto is a planet’ after all.

[15] Similarly, a utilitarian and Kantian may agree on the role that the term ‘right’ is to play while disagreeing about what intension fills that role. Then, they could be said to agree about the meaning of ‘right’, but not about what makes something right. Alternatively, they may have a more fundamental disagreement over what role the term is to play. Then, they would not even agree on the meaning of ‘right’ in the sense under consideration. We do not take a stand here on what kind of dispute ethicists tend to be engaged in.

Externalism does not on its own save Conventional Wisdom. Even if ‘planet’ has a meaning that depends on external factors, that doesn’t explain its embedded behavior. Whatever interpretation is associated with ‘planet’, by whatever combination of metaphysical and social facts one likes, Conventional Wisdom stipulates that this interpretation be held fixed in all embedded contexts. So, according to Conventional Wisdom, if external factors determine that ‘planet’ is interpreted so that planets must clear their orbital neighborhood, it will retain this interpretation in counterfactuals such as (4). Thus, either the externally determined interpretation of ‘planet’ is shifted in (4), in which case Conventional Wisdom fails, or it is not, in which case we cannot explain the available true reading of (4).

Still, an externalist might take issue with our characterization of the dispute between Alpha and Beta. A Putnam-style externalist, for instance, might argue that the interpretation of ‘planet’ is a natural kind, so that the IAU did not change the definition of ‘planet’ but rather discovered something about planets. We do not incline towards this kind of view in the case of ‘planet’. But this does not affect our main line of argument. This version of externalism is implausible for conventional kinds such as martinis, SUVs, or smartphones. So there remains motivation for our overall picture even if one is skeptical of the ‘planet’ example. For ease of exposition, we will assume that this kind of externalism is false for ‘planet’, though one should feel free to replace this example with another if desired.

Alternatively, a social externalist might argue that a single individual cannot change the interpretation of ‘planet’ by themselves. However, our view is not incompatible with this claim. We do not require that Alpha can change the conventions associated with ‘planet’ in their community just by asserting ‘Pluto is a planet’. Their assertion would have to be accepted by some relevant portion of their linguistic community for that to happen. Instead, we may view their assertion as a (possibly unsuccessful) *attempt* to get their linguistic community to change its linguistic conventions.[16] Thus, our view is compatible with social externalism like that of Burge (1979), according to which the wider linguistic community, and not just individual speakers, determines the meanings of words.

3 Conventions as Plans for How to Talk

In the previous section, we presented examples where c-monsters naturally emerge and Conventional Wisdom is violated. We now develop a picture

[16] On meaning change and externalism, see, e.g., Lassiter 2008; Koch 2018.

of communication that will provide the foundation for our semantics of c-monsters, in §4.

3.1 Metalinguistic Negotiations

Let us focus for now just on the beginning of the exchange between Alpha and Beta from (3)—before Beta asserts the c-monstrous counterfactual:

(25) **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.

In the literature, exchanges like (25) have been called **metalinguistic negotiations**.[17] A metalinguistic negotiation is a dispute over how to use terms (here, 'planet'), rather than about factual matters. These disputes concern *what counts as what* rather than *what things are like*. So with (7), Alpha does not intend to communicate anything factual about Pluto:

(7) Pluto is a planet.

It is common ground between Alpha and Beta that Pluto is significantly smaller than Neptune, fails to clear its orbital neighborhood, and so on. Rather, Alpha is advocating for a certain way of classifying things as planets, on which Pluto would count.

This distinction between two ways one might use a predicate like 'planet' is analogous to a distinction made by Barker (2002) between two ways of using gradable predicates like 'tall'. Consider the following sentence:

(26) Feynman is tall.

One can use (26) to convey something about Feynman's height. For instance, someone who has never seen Feynman could learn that his height is relatively great. Call this the **descriptive** use of (26). But one can also use (26) differently: to convey something not about Feynman's height but

[17] See Plunkett and Sundell 2013, p. 15 and Plunkett 2015, p. 832. The term was partly inspired by the work of Kyburg and Morreau (2000) and Barker (2002) on vague predications. Metalinguistic negotiations are also what Haslanger (2000, p. 33) calls "analytical projects"; later, Haslanger (2005, p. 11) calls them "ameliorative projects". Related work goes back at least to Gallie 1955. See also McConnell-Ginet 2006, 2008; Burgess and Plunkett 2013a,b; Ludlow 2014; Cappelen 2018; Sterken 2019; Hansen 2019.

instead about *what counts as* tall in that context. For instance, someone who has seen Feynman could learn that his height counts as tall in this country. Call this the **classificatory** use of (26).

One distinguishing feature of classificatory uses of gradable adjectives is the way they interact with phrases such as 'counts as', 'consider', and 'to me'. In general, classificatory uses of sentences containing gradable adjectives can be reported using phrases of the form 'I count x as F' or 'I consider x [to be] F' or 'x is F to me'. The same is not true for descriptive uses, however.

For instance, unless there is some unclarity over how to measure height, the extension of 'taller than' will not depend on what counts as tall in an ordinary context. Thus, 'taller than' generally has only a descriptive use. Suppose we know exactly how tall Feynman and Fermi are. In that case, each of the following is marked:

(27) a. #I count Feynman as taller than Fermi.
b. #I consider Feynman [to be] taller than Fermi.
c. #To me, Feynman is taller than Fermi.

Each of these seems to suggest, incorrectly, that one can simply *decide* that Feynman is taller than Fermi.

By contrast, if someone asks, 'What is tall in your country?', each of the following is felicitous:

(28) a. I count Feynman as tall.
b. I consider Feynman [to be] tall.
c. To me, Feynman is tall.

But they will sound marked again in contexts where the standards for 'tall' are clear. Thus, we can use phrases like 'count as', 'consider', and 'to me' as tests for when a sentence is being used in a classificatory or descriptive way.[18]

The descriptive use of (26) is easy to accommodate on a traditional

[18] Note there is an important difference between 'I count x as F', which only conveys a classificatory use, and 'x counts as F', which can be used descriptively. Suppose there is a sign for a roller coaster ride that says how tall one must be to ride. In that case, 'I count Feynman as tall enough' still seems to suggest the matter can be settled by stipulation. By contrast, it would be perfectly fine to say 'Feynman counts as tall enough' to convey something about Feynman's height. Thus, 'x counts as F' can be used to convey either the descriptive or classificatory use (though perhaps the latter is more common). Similarly for the difference between 'I consider x [to be] F' and 'x is considered [to be] F'.

Stalnakerian theory of assertion,[19] on which both assertoric content and common ground are modeled as sets of worlds—so long as we assume that each context comes equipped with a threshold for gradable adjectives like 'tall'. When a speaker asserts (26) in a context, they are proposing to remove from the common ground worlds in which Feynman is not above the contextually-determined threshold.

The classificatory use is harder to accommodate, however. The problem is that when a speaker uses (26) classificatorily, they are conveying something about the threshold, not just Feynman's height. But if it is commonly accepted in the conversation that Feynman is 6′1″, then an assertion of (26) does nothing to the common ground. So such a classificatory use of (26) is predicted to be semantically uninformative on Stalnaker's theory.

To account for classificatory uses of vague predications, Barker (2002) proposes to model the common ground of a conversation as a set of *world-delineation pairs*, where the delineation specifies the threshold for vague predicates such as 'tall'. If it is common knowledge what counts as tall in a given context, then the delineation parameter will be the same in every pair in this set. But if not, then different pairs may contain different delineations. An assertion of (26) rules out every world-delineation pair $\langle w, d \rangle$ where Feynman at w falls below the threshold picked out by d.

We propose to extend this model of vague predication to predication more broadly.[20] Indeed, just as there are two uses of (26), there are also two uses of (7):

(7) Pluto is a planet.

First, one can use (7) to convey something about Pluto's physical characteristics, such as its orbit. As with vague predications, we might call this the **descriptive** use of (7). Second, one can use (7) to convey something about *what counts as* a planet. We saw such a use in the metalinguistic negotiation in (25). Again, as with vague predications, we might call this the **classificatory** use of (7).

Building on the observation from Kennedy and Willer (2016) that we saw above, the same tests for distinguishing descriptive and classificatory uses of gradable adjectives can be used to distinguish descriptive and classificatory uses of other predicates. For instance, it generally sounds bad to say:

(29) a. #I count Earth as a planet.

[19] See Stalnaker 1970, 1978, 1999, 2002.

[20] Barker (2013, p. 252) and MacFarlane (2018, §11) consider, but do not pursue this possibility.

b. #I consider Earth [to be] a planet.
c. #To me, Earth is a planet.

By contrast, it sounds fine to say:[21]

(30) a. I count Pluto as a planet.
b. I consider Pluto [to be] a planet.
c. To me, Pluto is a planet.

The analogy with vague predications suggests that classificatory uses of (7) can be modeled similarly to how Barker models classificatory uses of (26). Instead of talking about delineations for vague predicates, we talk more generally about *conventions* for interpreting all predicates. Whereas a delineation determines the threshold for gradable adjectives, a convention determines the intension assigned to predicates. On this model, we take the common ground of a conversation not to be a set of worlds, nor a set of world-delineation pairs, but a set of *world-convention* pairs. An assertion of (7) rules out every world-convention pair $\langle w, c\rangle$ where Pluto is not a planet at w as interpreted according to c, i.e., where Pluto at w is not in the extension assigned to 'planet' by c at w.[22]

3.2 What is a Convention?

We think of a linguistic convention as a type of plan: a plan for how to use words.[23] In any given conversational context, each speaker adopts certain

[21] The conditions for classificatory uses are related to the conditions that Kennedy and Willer identify for licensing 'consider' attitude reports, but they do not perfectly line up. Kennedy and Willer hold that 'consider' is felicitous when the embedded claim is "counterstance contingent", meaning that its truth varies across different salient choices for how to define terms. Any use of a sentence that is classificatory in our sense will be counterstance contingent, assuming, as seems reasonable, that the conventions left open in the common ground are all salient. However, the reverse does not hold. A sentence may be counterstance contingent even though all conversational participants are in agreement in their rejection of the, nonetheless salient, linguistic choice that would make it true. For instance, an astronomer talking only with like-minded colleagues, all known to be in support of the IAU's new definition, can still felicitously say that lots of people *consider* Pluto a planet. But they cannot use (7) classificatorily, to rule out conventions left open in the common ground.

[22] This is not to say that conventions can simply replace delineations in an adequate model of vague predication. Delineations may require different structure than conventions to accommodate the ways in which gradable adjectives can be modified by measure phrases (as in '2 feet tall'). We set this aside here.

[23] One might of course, following Lewis (1969), hold that a convention literally just *is* a special kind of plan, i.e., a plan to act in certain ways in special kinds of coordination problems. We do not need to go that far, however. Our use of the word 'convention' should

conventions for how to use language. Most of the time, speakers endorse a common set of conventions, but not always. They may disagree about edge cases for ‘planet’, for example. Speakers might also be undecided about which conventions to adopt—their plans for how to talk may be (and often are) incomplete. Thus, there may be multiple conventions for how to use a word like ‘planet’ in a given conversation left open by what the speakers commonly accept. In that case, an assertion of a sentence like (7) will rule out from the common ground not only certain ways the world could be but also certain ways of speaking. If a speaker’s assertion is accepted by the other conversational participants, then they will come to adopt the speaker’s proposal for how to talk. In effect, accepting (7) will be partly a matter of deciding how to use ‘planet’.

Talk of conventions as plans is meant to invoke a comparison with the expressivist theory of normative discourse developed by Gibbard (2003). For Gibbard, an agent’s mental state is characterized not just by what they believe the world is like, but also by their intentions, i.e., what plans they adopt. A plan can be thought of as a set of what Gibbard calls **hyperplans**, i.e., maximally specific plans for how to act in any situation. According to Gibbard, when a speaker asserts a sentence with purely normative content, such as ‘Sherlock ought to pack’, they express a planning state—that is, they express something about what actions their adopted plan permits in Sherlock’s situation. We can model this in a broadly Stalnakerian framework by taking the content of a sentence, as well as the common ground of a conversation, to be modeled not as a set of worlds, but a set of *world-hyperplan pairs*. Accepting an assertion of ‘Sherlock ought to pack’ rules out of the common ground world-hyperplan pairs $\langle w, h \rangle$ where h permits Sherlock not to pack at w.

It is also analogous to the expressivist theory of vagueness developed by MacFarlane (2016). MacFarlane starts with a Barker-style semantics for vague terms, but proposes a Gibbardian interpretation of it, on which delineations are thought of as hyperplans for resolving vagueness. Thus, when a speaker asserts, say, ‘Apple A is large’, they express a planning state that commits them to counting apple A as large. Accepting such an assertion rules out of the common ground world-delineation pairs $\langle w, d \rangle$ where apple A does not meet the threshold for ‘large’ set by d at w.

In the same way, we propose that a speaker’s mental state is characterized not just by their factual commitments, i.e., by what possibilities they accept, but also by their conventional commitments, i.e., by what conven-

not be taken to provide an analysis of the concept as opposed to a technical stipulation.

tions they adopt. A convention can be modeled as a set of what we'll call **hyperconventions**, i.e., maximally specific interpretations of every single word in the language. Formally, a hyperconvention can be represented as an interpretation function in the usual model-theoretic sense, i.e., a function that assigns an ordinary possible-worlds intension to every non-logical expression (see §4 for details).

So a speaker's conventional commitments can be represented by a set of world-hyperconvention pairs—intuitively, those pairs that are not ruled out by what the speaker thinks the world is like and what conventions they adopt. We represent a speaker's mental state as *a set of pairs* of worlds and hyperconventions, rather than *a pair of sets* of worlds and of hyperconventions, in order to capture the way in which a speaker's conventional commitments may be contingent on what the world is like. For instance, a speaker may be ignorant of the definition of 'planet' the IAU adopts but nevertheless adopt a plan to use the word however the IAU uses it. In that case, the speaker rules out certain combinations of worlds and hyperconventions even though each individual world or hyperconvention is live.

These conventional commitments are very much like commitments to adopt certain plans in Gibbard's sense. Indeed, many of the key ideas Gibbard uses to develop an expressivist theory of normative discourse, and that MacFarlane applies in his expressivist theory of vagueness, can be used to develop a parallel theory of metalinguistic negotiations in a broadly Stalnakerian spirit.

3.3 Communication

Given this picture of a speaker's mental state, we can revise Stalnaker's theory of assertion by simply replacing worlds with world-hyperconvention pairs at every level. The common ground, i.e., what is commonly accepted by the conversational participants both in terms of their factual commitments and their conventional commitments, can be modeled formally as a set of pairs of possible worlds and hyperconventions. Assertion is then just a kind of proposal to remove world-hyperconvention pairs from the common ground.

Thus, an assertion of a simple sentence such as (7) will not just be a proposal to rule out certain worlds from the common ground, but also certain ways of speaking. The assertoric content of (7) is the set of world-hyperconvention pairs $\langle w, c\rangle$ according to which Pluto at w is in the extension assigned to 'planet' by c at w. Accepting a sentence like (7), then, is not merely a matter of committing oneself to the world being a certain way.

It is also, at least in part, a matter of committing oneself to using the word 'planet' in a certain way.

To be clear, we are not saying that an assertion of (7) is equivalent to an assertion of (31):

(31) According to how I define 'planet', Pluto is a planet.

When a speaker says (7), they do not *assert that* they use 'planet' in a certain fashion. It is not part of the content of what is asserted that the speaker adopts a certain convention. Rather, in asserting (7), the speaker merely *expresses* their conventional commitments. The difference between (7) and (31) is similar to the difference between (32-a) and (32-b):

(32) a. It is raining.
b. I think it is raining.

While (32-a) expresses one's belief that it is raining, only (32-b) asserts that one believes it is raining. So too, while (7) expresses the conventions one adopts, only (31) asserts that one adopts these conventions.[24]

Both in Barker's framework for modeling gradable adjectives and our framework for modeling predication more generally, what distinguishes descriptive uses from classificatory uses is the common ground, not the content of the assertions. The *truth conditions* of descriptive and classificatory uses do not differ. Instead, a descriptive use arises when speakers accept a common plan for how to interpret the sentence asserted, whereas a classificatory use arises when speakers do not.

To illustrate, suppose speakers in a conversation are completely coordinated over what conventions to adopt—that is, they all agree on how to use words. In that case, every world-hyperconvention pair in the common ground will have the same hyperconvention component, i.e., if $\langle w_1, c_1 \rangle$ and $\langle w_2, c_2 \rangle$ are in the common ground, then $c_1 = c_2$. Thus, an assertion of (7) in such a context will effectively constitute a proposal to remove possible worlds from the common ground. Hence, the old Stalnakerian picture of assertion can be viewed as a special case of the current proposal where speakers are coordinated on a shared convention.

On the flip side, suppose speakers in a conversation are completely coordinated regarding what the world is like. In that case, every world-hyperconvention pair in the common ground will have the same world component, i.e., if $\langle w_1, c_1 \rangle$ and $\langle w_2, c_2 \rangle$ are in the common ground, then

[24] On the distinction between asserting and expressing, see, e.g., Gibbard 2003, Yalcin 2012.

$w_1 = w_2$. Thus, an assertion of (7) in such a context will effectively constitute a proposal to remove hyperconventions from the common ground. This, in effect, is what is going on in examples like (25): though Alpha and Beta are in complete agreement about the facts concerning Pluto, they disagree over how to classify it.

These are both extreme cases. Usually, conversational participants will not be completely coordinated either on what the world is like or on how to use words. In general, a use of (7) can be mixed, being neither completely descriptive nor completely classificatory. This is a feature of our view, not a bug. When one learns that some sentence *S* is true, one often learns something both about the world and about conventions. When you teach a child how to use the word 'plant' by, for instance, pointing to a rose and saying 'This is a plant', you teach them not only something about how to use the word 'plant' but also something about roses. For clarity, we will use "descriptive" to refer to purely descriptive uses of sentences, and "classificatory" for uses that are at least partially classificatory.

4 Semantics for C-Monsters

In §2, we saw that counterfactuals such as (4) can violate Conventional Wisdom.

(4) If Pluto were a planet, there would be dozens of planets in the solar system.

That is, counterfactuals can be **c-monsters**, in the sense that they trigger a shift in the conventional interpretation of their constituents. In this section, we build on the theory of assertion from §3 to develop a semantic framework for counterfactual c-monsters. We leave for future work the semantics of other c-monstrous vocabulary.

4.1 The Conventionalist Semantics

Our proposal, which is largely inspired by the work of Einheuser (2006), is to take an index of evaluation to be a pair of a world and a hyperconvention and to allow expressions such as counterfactuals to shift the hyperconvention of an index.[25,26]

To demonstrate, we present a simple variant of the Lewis-Stalnaker semantics for counterfactuals (Stalnaker, 1968; Lewis, 1973) that takes into account their c-monstrous behavior. We are not ultimately committed to using this particular semantic framework for modeling counterfactuals; others could be adapted in a similar way.

We start by introducing a simple formal language with names ($a_1, a_2, a_3, \ldots$), predicates of any arity ($P_1^n, P_2^n, P_3^n, \ldots$), boolean connectives ($\neg$, $\wedge$, $\vee$, $\supset$), and a counterfactual conditional ($\Box\!\!\rightarrow$). The well-formed formulas of our language can be summarized in Backus-Naur form as follows:

$$A ::= P(a_1, \ldots, a_n) \mid \neg A \mid (A \wedge A) \mid (A \vee A) \mid (A \supset A) \mid (A \mathrel{\Box\!\!\rightarrow} A).$$

A model for this language is a tuple of the form $\mathcal{M} = \langle W, D, f \rangle$. Here, W is a nonempty set of worlds and D is a nonempty set of objects (we assume a constant domain semantics for technical simplicity). To define the selection function f, we first need some further definitions. A **hyperconvention** over $\mathcal{M}$ is a function c that maps each name to an element of D and each n-ary predicate to an n-ary intension over W:

(i) $c(a) \in D$ for each name a

(ii) $c(P^n)\colon W \longrightarrow \wp(D^n)$ for each n-ary predicate P^n.

An **index** over $\mathcal{M}$ is a world-hyperconvention pair $i = \langle w_i, c_i \rangle$. Where $I^{\mathcal{M}}$ is the set of all indices over $\mathcal{M}$, the selection function is a map $f\colon (\wp(I^{\mathcal{M}}) \times I^{\mathcal{M}}) \longrightarrow \wp(I^{\mathcal{M}})$. The selection function of course needs to satisfy the usual constraints to be realistic (which may be dropped or modified as desired):

[25] Other recent work in a similar spirit include Armstrong 2013, Ch. 3 (on loose talk and imprecision), Muñoz 2019b, Ch. 5 (focusing more on attitude reports), and Gaus 2020 (on counterfactuals).

[26] An alternative approach, advocated by Coppock (2018), does not *add* something like our hyperconventions to points of evaluation, but instead *replaces* worlds in points of evaluation with "outlooks", or refinements on worlds that settle both factual matters and matters of opinion. Outlooks do similar work for Coppock as world-hyperconvention pairs do for us (though she does not discuss counterfactuals). However, where Coppock uses the machinery of worlds and outlooks to define "discretionary" and "objective" predicates, we take all predicates discussed above to have precisely the same status. Whether a predicate is used in a discretionary (classificatory) way simply depends on what convention choices are common ground among speakers in a conversation. (Note, however, that we do not take a stand on how to deal with more traditionally "subjective" predicates, like predicates of personal taste, which is Coppock's main motivation.) For related discussion, see Muñoz 2019a, p. 24. Separating out conventions in the index will also allow us to clearly define different understandings of the counterfactual conditional below.

(i) $f(X, i) \subseteq X$

(ii) If $i \in X$, then $f(X, i) = \{i\}$

(iii) If $X \subseteq Y$ and $f(X, i) \neq \emptyset$, then $f(Y, i) \neq \emptyset$

(iv) If $X \subseteq Y$ and $f(Y, i) \cap X \neq \emptyset$, then $f(X, i) = f(Y, i) \cap X$.

For each sentence A, the semantic value of A in $\mathcal{M}$, i.e., $[\![A]\!]^{\mathcal{M}}$, is defined recursively as follows:

$$\begin{array}{rcl} [\![P^n(a_1, \ldots, a_n)]\!]^{\mathcal{M}} & = & \{i \in I^{\mathcal{M}} \mid \langle c_i(a_1), \ldots, c_i(a_n)\rangle \in c_i(P^n)(w_i)\} \\ [\![\neg A]\!]^{\mathcal{M}} & = & I^{\mathcal{M}} - [\![A]\!]^{\mathcal{M}} \\ [\![A \wedge B]\!]^{\mathcal{M}} & = & [\![A]\!]^{\mathcal{M}} \cap [\![B]\!]^{\mathcal{M}}, \text{etc.} \\ [\![A \mathrel{\Box\!\!\rightarrow} B]\!]^{\mathcal{M}} & = & \{i \in I^{\mathcal{M}} \mid f([\![A]\!]^{\mathcal{M}}, i) \subseteq [\![B]\!]^{\mathcal{M}}\}. \end{array}$$

Consequence is then defined as preservation of truth over all indices: $A_1, \ldots, A_n \models C$ iff for every model $\mathcal{M}$ and every index $i \in I^{\mathcal{M}}$, if $i \in [\![A_k]\!]^{\mathcal{M}}$ for all $1 \leq k \leq n$, then $i \in [\![C]\!]^{\mathcal{M}}$. Call this the **conventionalist semantics**.

It can be shown that the logic of the conventionalist semantics is the logic of the standard Lewis-Stalnaker semantics.[27] Thus, the differences between the two semantics arise not in their consequence relations but in their assignment of truth conditions. In the standard Lewis-Stalnaker semantics, the model provides a fixed hyperconvention (i.e., interpretation function) that is never altered when evaluating the truth of a counterfactual. On the conventionalist semantics, by contrast, models do not provide a fixed hyperconvention for evaluating truth. Instead, the hyperconvention

[27] Proof: it suffices to show that (i) for any pointed conventionalist model, there is a pointed Lewis-Stalnaker model satisfying the same exact formulas, and (ii) for any pointed Lewis-Stalnaker model, there is a pointed conventionalist model satisfying the same exact formulas. A Lewis-Stalnaker model is just a quadruple of the form $\mathcal{S} = \langle W, D, f, V\rangle$, where W and D are as before, $f\colon \wp(W) \times W \longrightarrow \wp(W)$ satisfies the usual constraints on selection functions, and I is an interpretation function such that $V(a) \in D$ and $V(P^n)\colon W \longrightarrow \wp(D^n)$.

For (i): let $\mathcal{M} = \langle W, D, f\rangle$ be a conventionalist model. Define $\mathcal{M}_S = \left\langle I^{\mathcal{M}}, D^{I^{\mathcal{M}}}, f, V\right\rangle$ where $I^{\mathcal{M}}$ is the set of indices over $\mathcal{M}$ and V is the interpretation function such that $V(a)$ is the function that maps any index $\langle w, c\rangle$ to $c(a)$ and $V(P^n)$ is the function that maps any index $\langle w, c\rangle$ to $c(P^n)(w)$. Then by induction, $[\![A]\!]^{\mathcal{M}} = [\![A]\!]^{\mathcal{M}_S}$, i.e., $i \in [\![A]\!]^{\mathcal{M}}$ iff $i \in [\![A]\!]^{\mathcal{M}_S}$.

For (ii): let $\mathcal{S} = \langle W, D, f, V\rangle$ be a Lewis-Stalnaker model. Where X is a set of indices over some model, define $X\restriction_W = \{w \mid \exists c\colon \langle w, c\rangle \in X\}$. Define $\mathcal{S}_C = \langle W, D, f_C\rangle$ where $f_C(X, \langle w, c\rangle) = \{\langle v, c\rangle \in X \mid v \in f(X\restriction_W, w)\}$ (in other words, f_C never shifts the convention component). It is straightforward to check that f_C is a selection function. Then by induction, $w \in [\![A]\!]^{\mathcal{S}}$ iff $\langle w, V\rangle \in [\![A]\!]^{\mathcal{S}_C}$.

appears as a shiftable component of the index. It is in this sense that counterfactuals can shift the conventions used to interpret their constituents and thus can be c-monsters.[28,29]

4.2 Similarity

Just as with the ordinary Lewis-Stalnaker semantics, we may think of $f(X, i)$ as the set of indices in X that are most similar to i in the relevant respects. What "similar in the relevant respects" amounts to will depend on the context. Similarity between indices is partly a matter of similarity in factual component and partly a matter of similarity in conventional component. Different contexts will assign different weights to similarity with respect to these components. It is difficult to say anything more specific about how to determine when one index is more similar than another, but this isn't a special problem for the conventionalist semantics.

Einheuser (2006) articulates three different ways to interpret counterfactual conditionals in light of the possibility of shifting conventions:

Countersubstratum. $A \mathrel{\Box\!\!\rightarrow_s} C$ is true at $\langle w, c\rangle$ iff C is true at every $\langle w', c\rangle$ where w' differs minimally from w so as to make A true.

Counterconventional. $A \mathrel{\Box\!\!\rightarrow_c} C$ is true at $\langle w, c\rangle$ iff C is true at every $\langle w, c'\rangle$ where c' differs minimally from c so as to make A true.

[28] A similar approach (inspired by Stalnaker 1978) is to evaluate truth relative to a pair of worlds, one considered as actual, the other considered as counterfactual, and to let the world-considered-as-actual determine the convention parameter. This view disallows cases where the convention parameter comes apart from the world-considered-as-actual, as in (i).

(i) If Pluto were a planet, there would be many more planets than the IAU actually recognizes.

For this reason we prefer to isolate the convention parameter.

[29] Note that we have not claimed that the "meaning" of an expression is shifted by the counterfactual conditional. There are two things one can mean by "meaning" in our framework: the *intension* (i.e., a function from worlds to extensions) and the *compositional semantic value* of an expression (i.e., a function from hyperconventions to intensions). Counterfactuals can shift the former, but not the latter. Similarly, though Alpha and Beta do not express the same coarse-grained proposition (the same set of worlds) if they say 'Pluto is a planet', they do express the same fine-grained proposition (the same set of world-hyperconvention pairs). So in that sense, Conventional Wisdom and No C-Monsters hold. But in that same sense, these principles are trivial: one can always reinvent one's notion of meaning to be fine-grained enough for these principles to hold. In its original sense, though, where the "conventional interpretation" of an expression is understood to be its intension, these principles fail according to the conventionalist semantics.

Diagonal Counterfactual. $A \mathrel{\Box\!\!\rightarrow_d} C$ is true at $\langle w, c\rangle$ iff C is true at every $\langle w', c'\rangle$ where c' is the convention grounded in the linguistic practices of speakers at w' and w' differs minimally from w so as to make A true at $\langle w', c'\rangle$.[30]

One benefit of our view is that all three of these readings can be predicted by a single context-sensitive semantic entry for $\Box\!\!\rightarrow$. What determines how a counterfactual is interpreted are the features of the context that determine more broadly the notion of similarity used to fix the selection function. In general, maximizing similarity requires holding fixed as much about our current situation that is not relevant to the topic under discussion as possible. In contexts where the conventions being adopted are not at issue, speakers will tend to hold their conventions fixed when entertaining counterfactual suppositions. In contexts where the conventions are at issue, speakers will tend instead to hold matters of fact fixed. Thus, what distinguishes the different uses of conditionals such as (4) are exactly the features of context that determine the selection function f more broadly.[31]

With that said, there may be constraints on how freely context can influence the selection function. In particular, it is hard to come up with a natural case where a counterfactual is read purely counterconventionally, i.e., where the antecedent only shifts the conventional component of an index. It seems as though shifts in convention are very often, if not always, associated with shifts in world. Consider a variant on (4):

(33) If Pluto were a planet, there would be dozens of planets in the solar system, but we wouldn't call all of them 'planets'.

Some speakers have the intuition that (33) is false, even bearing in mind

[30] Einheuser (2006, pp. 467–468) assumes (in our terminology) that each world grounds exactly one convention. This assumption was introduced for technical simplicity, but it is arguably not very realistic, and dropping it raises complications for the statement of Diagonal Counterfactual. We set aside this concern here, though it does raise interesting questions about what exactly a "diagonal" counterfactual reading is, when this simplifying assumption is not taken for granted. We thank James Walsh for pointing this out.

[31] One might think that all counterconventionals in natural language are diagonal counterfactuals in the sense just defined. Here is a potential counterexample to that thought:

(i) If Pluto were a planet, then there would be dozens of planets in the solar system even if everyone were dead.

This would be true on neither a countersubstratum reading nor a diagonal reading. This is yet another reason why we do not adopt a Stalnakerian diagonalization approach to analyzing counterconventionals (cf. footnote 28).

the true reading of (4). If that's right, it suggests that something like the following principle governs selection functions:

No Pure Counterconventionals. If $\langle w, c' \rangle \in f(X, \langle w, c \rangle)$, then $c' = c$.

According to No Pure Counterconventionals, counterfactuals never shift the conventional component of an index without also shifting the factual component.

Officially, we remain neutral about No Pure Counterconventionals. Perhaps there are no pure counterconventional readings in the wild. Or perhaps they are simply rare. Our semantic framework is flexible enough to model counterfactual c-monsters whether or not this principle holds. For ease of exposition, we will continue to use the term "counterconventional" to include any counterfactual that involves a shift in convention (including "impure" counterconventionals that shift both) so as not to prejudge the question over whether No Pure Counterconventionals holds.

In this framework, Conventional Wisdom corresponds to a much stronger principle governing the influence of context on the selection function:

No Counterconventionals. $f(X, \langle w, c \rangle) \subseteq (W \times \{c\})$.

According to No Counterconventionals, the only counterfactuals in natural language are countersubstratums: counterfactuals *never* shift the convention parameter of an index. Perhaps in some contexts, No Counterconventionals holds. But as we have argued, it is implausible to maintain that No Counterconventionals universally governs all conversational contexts.

4.3 Use & Mention

So far, we have only discussed counterfactuals where language is not mentioned anywhere in the antecedent or consequent. Even when language is mentioned, though, much of the story remains the same. For instance, in a context where speakers are debating how to define 'planet', both of the following counterfactuals can shift the conventions used to interpret 'planet' in the consequent:

(4) If Pluto were a planet, there would be dozens of planets in the solar system.

(34) If 'planet' were defined so as to include Pluto, there would be dozens of planets in the solar system.

Thus, to return to the cases we began with in §1, sentences such as (1) and (35) can be heard either as countersubstratums or as counterconventionals, depending on the context:

(1) If 'water' referred to gasoline, water would fuel fire.

(35) If you called a dog's tail a leg, dogs would have five legs.

On its counterconventional reading, (1) can be heard as true (without changing the chemical nature of water); and likewise for (35) (without changing the anatomy of dogs). This is to reject the received wisdom about such sentences, which holds that they have only one reading, namely, the countersubstratum one; and on this reading both are false. When people interpret these sentences as true, the conventional wisdom goes, they are engaging in a sort of use-mention confusion. By contrast, our view rationalizes these "confused" uses as legitimate interpretations of these counterfactuals that naturally fall out of a systematic theory of c-monsters. (The riddle is a riddle at all because there are two ways to interpret the question.)

Still, there is a difference between counterfactuals where language is explicitly mentioned and counterfactuals where it is not. In particular, while (4) and (34) may coincide on their counterconventional reading, they diverge on their countersubstratum reading. On its countersubstratum reading, (4) involves evaluating counterfactual scenarios where Pluto's orbit is different. On the countersubstratum reading of (34), however, we evaluate scenarios where the word 'planet' is used differently. Thus, the countersubstratum readings of (4) and (34) express different (though in both cases false) propositions.

5 Doing without Conventional Wisdom

In this section, we reassess the two influential philosophical arguments from §1 that relied on the now-rejected Conventional Wisdom. We show that Kripke's defense of the necessity of identity survives largely intact, but Lewy's objection to modal conventionalism does not.

5.1 The Necessity of Identity

Recall Kripke's defense of the necessity of identity. Consider a counterfactual world w where Venus is the morning star while Mars is the evening star. Why is w not a world where Hesperus and Phosphorus are distinct? The reason, according to Kripke, is that what is true at w depends on our linguistic conventions, not those of speakers in w. Speakers in w use the names 'Hesperus' and 'Phosphorus' to refer to different individuals, so the sentence 'Hesperus is Phosphorus' will be false in *their* language. But if we describe w in *our* language, where 'Hesperus' and 'Phosphorus' both refer to Venus, then it is still true at w that Hesperus is Phosphorus. And by Conventional Wisdom, only this latter point is relevant to the modal status of identity. Hence, this is not a counterexample to the necessity of identity.

At first, the failure of Conventional Wisdom seems to undermine Kripke's argument. Counterfactuals can be c-monsters, and, as we saw in §2, it is plausible that many metaphysically loaded expressions, such as those expressing metaphysical necessity, can be as well. So we would expect the following sentences to have c-monstrous readings:[32]

(36) a. Had Mars shone but a bit brighter, Hesperus would not be Phosphorus.
b. Hesperus and Phosphorus could have been distinct.

Kripke's argument seems to require that there is only one (false) interpretation of such sentences. When people say things like (36-b) in ordinary conversation, he claims, they are really just speaking loosely or making a mistake (or are using 'could' in an epistemic sense). Their use should be explained by pragmatics or some kind of error theory, not by a semantic theory that allows for such sentences to be literally true.

On this point, of course, we part ways with Kripke. These uses are not just mistakes or loose talk. Rather, they are predicted by a systematic, well-behaved, and independently motivated semantic theory of c-monsters. Indeed, the availability of a c-monstrous interpretation of (36-b) might explain why there was so much initial confusion over contingent identity: people conflated the reading of (36-b) on which the meanings of 'Hesperus' and 'Phosphorus' can be shifted with the reading on which their meanings are held fixed.

Despite this, however, we think that the spirit of Kripke's argument

[32] Notice that the c-monstrous reading of (36-a) shifts the convention governing 'Hesperus' and 'Phosphorus' even though those terms do not occur in the antecedent. Though the examples discussed in §2 did not have this feature, we think such counterfactuals can naturally be interpreted counterconventionally, and our semantics in §4 allows this. As another example, the following seems true on its counterconventional reading, but false on its countersubstratum reading:

(i) If the IAU had decided differently, Pluto would be a planet.

survives. As Kripke (1980, p. 77) already pointed out, the sense in which (36-b) is true gives rise only to a relatively cheap form of contingent identity. In that sense, 2+2 and 4 could have been distinct. Indeed, practically nothing is necessary in this sense, since we can always reinterpret the meanings of our terms however we like.

The philosophically interesting question, and the one Kripke clearly had in mind, is over the descriptive reading of (36-b), where we hold fixed the actual meanings of 'Hesperus' and 'Phosphorus' and ask whether the *objects* those names actually denote could have been distinct. This question is not trivial, as some have argued that identity is contingent even in this sense.[33] Thus, even though Kripke appeals to Conventional Wisdom in defending the necessity of identity, he need not have. He does not have to deny that (36-b) has a true c-monstrous reading. He only has to say that (36-b) is false on its descriptive reading.

One might question, furthermore, whether the sentences in (36) even admit of c-monstrous readings. While *predicates* can undeniably be reinterpreted in the scope of counterfactuals, attitude verbs, and the like, it is less clear that *names* can undergo the same reinterpretation. For instance, the oddness of the sentences in (37) suggests that the names in (36) do not pass the usual tests for classificatory uses (§3.1):

(37) a. ?I count Hesperus as Phosphorus.
b. ?I consider Hesperus [to be] Phosphorus.
c. ?To me, Hesperus is Phosphorus.

Even though a linguistic community may attach names to objects however it pleases, it seems to be odd, once names are attached to objects, to take there still to be a choice about which things are identical. Thus, though Conventional Wisdom is not generally true, there is some reason to suspect that it might be a true principle governing names, in which case Kripke's appeal to it would be perfectly legitimate. Kripke's defense of the necessity of identity along these lines would not, then, be undermined by the failure of Conventional Wisdom. One could not simply appeal to c-monsters to argue for the contingency of identity.[34]

However, the judgments in (37) seem to be affected by the fact that the usage of 'Hesperus' and 'Phosphorus' is fully settled in our linguistic community. In a context where the designations of 'Hesperus' and 'Phos-

[33] See Lewis 1971, 1986; Gibbard 1975; Karmo 1983; Schwarz 2014; Kocurek 2018.

[34] Of course, this would be a purely defensive maneuver on Kripke's part. Other arguments against the necessity of identity exist; see Schwarz 2013 for discussion.

phorus' are not settled, the judgments in (37) improve. Thus, it is likely that the defectiveness of the sentences in (37) only arises because of the lack of classificatory contingency (in the sense of Kennedy and Willer 2016, as discussed above) of the names in question. And it may generally be rarer for the meaning of names to be under dispute, compared with the meaning of predicates.[35] Still, we think it best not to appeal to any distinctive linguistic behavior of names as compared with predicates in rehabilitating Kripke's argument against the necessity of identity.

5.2 Modal Conventionalism

Lewy's objection to modal conventionalism, recall, is that it seems to falsely predict that whether 68 + 57 = 125 depends on our linguistic conventions. One way to put the objection is in terms of counterfactuals. Modal conventionalism wrongly predicts that (38) is true:

(38) If '+' had meant quus, 68 + 57 would be 5.

Another way to get at the objection is in terms of the necessity of necessity: modal conventionalism seems to entail that what is necessary or contingent is itself a contingent matter. This violates a plausible principle of modal reasoning, viz., the 4 axiom, which states that if A is necessary, then A is necessarily necessary ($\Box A \supset \Box\Box A$).

But is modal conventionalism really committed to (38)? As Einheuser (2006) points out, it depends on whether it is interpreted as a countersubstratum or as a counterconventional.[36] Interpreted as a countersubstratum, (38) is false. Holding fixed what *we* mean by '+', 68 + 57 could not be 5. Interpreted as a counterconventional, (38) is true. But, as with contingent identity, it is true in an unproblematic and uncontroversial way. It just says that, had we adopted different conventions, different claims would count as necessary under those conventions. This is something that everybody should accept and does not present a particularly damning problem for the conventionalist.

Einheuser recommends exactly this line on behalf of the conventionalist.

[35] The interesting general question of what types of expressions do and do not allow for convention-shifting is one we have only begun to answer here. An anonymous reviewer notes that it seems harder to get a convention-shifting reading of indexicals and demonstratives than it is for names and predicates. More generally, we suspect it is harder to shift functional or closed class terms than lexical or open class terms. But for now, we leave this as a conjecture. On the shiftability of logical vocabulary, see Kocurek and Jerzak 2020.

[36] Wright (1985) makes a similar point on behalf of the conventionalist.

Though her main objective was to defend **ontological conventionalism**, according to which what exists depends on what conventions we adopt, her defense applies equally to modal conventionalism. Here is how she puts the defense:

> Conventionalists about abstract objects do not claim that since [the sentence $\Box$(there are numbers)] depends on contingent conventions, [the sentence $\Diamond \neg \Box$(there are numbers)] is true. Rather, they claim that, against the conceptual background of our *actual* practices, the existence of numbers is necessary. Had these practices been suitably different, they would have generated a different set of metaphysically possible worlds relative to which the existence of numbers would not be necessary. . . (p. 477)

Thus, according to Einheuser, Lewy's objection gets off the ground only by uncharitably interpreting (38) as a countersubstratum instead of as a counterconventional.

However, there is still something unsatisfying about this way of responding to Lewy's objection. After all, on the counterconventional interpretation, (38) says something that nobody should deny: that if we had spoken differently, different claims would *count* as necessary on that way of speaking. The same holds for any truth whatsoever. Whether 'Grass is green' *counts* as true also depends on our linguistic conventions in this sense. How does this mundane observation reveal anything deep or interesting about the nature of necessity? And how does it provide us with the empiricist-friendly modal epistemology that modal conventionalists seek?

The conventionalist has to provide more than the mere existence of counterconventional readings of (38) in order to make good on the promises of a deflationary, empiricist-friendly epistemology of modality. Towards this end, we propose to supplement (or perhaps, replace) conventionalism with the following claim: the *sole* function of modal claims is to express aspects of our conventional commitments. We call this view **modal expressivism**.

When we look back at Ayer's statement of modal conventionalism, we find that his view is actually closer to modal expressivism than critics have recognized. In the preface of *Language, Truth and Logic*, he outlines his view in this way:

> Like Hume, I divide all genuine propositions into two classes: those which, in his terminology, concern "relations of ideas," and those which concern "matters of fact." The former class comprises the *a priori* propositions of logic and pure mathematics, and these I allow to be necessary and certain only because

> they are analytic. That is, I maintain that the reason why these propositions cannot be confuted in experience is that they do not make any assertion about the empirical world, but **simply record our determination to use symbols in a certain fashion**. (p. 31, emphasis added)

The vocabulary Ayer uses when describing his view is telling. He says necessary truths 'record', 'illustrate', 'call attention to', or 'reveal' the way we use language. He does not say they *assert* anything about language. In fact, he states explicitly that "they do not make any assertion about the empirical world", and so *a fortiori* do not make any claim about linguistic conventions (or how they're connected to necessity). This suggests that Ayer's view comes closer to what we are calling modal expressivism than is generally recognized.

Modal expressivism can accomplish much of what modal conventionalists want their view to do. For instance, it provides us with an empiricist-friendly modal epistemology, on which we can come to know what is necessary or contingent by reflecting on the way we use language—that is, reflecting on what linguistic conventions we adopt. We know that necessarily, all bachelors are unmarried because we know we adopt conventions according to which the sentence 'All bachelors are unmarried' is true regardless of what the world is like.

It also avoids a classic worry for modal conventionalism stemming from Kripke's observation that some necessary truths, such as 'Water is H_2O' are only knowable *a posteriori*. If necessary truths are necessary solely in virtue of our linguistic conventions, then shouldn't all necessary truths be knowable *a priori*? While Ayer did indeed conflate necessity with the *a priori*, and so does seem committed to this claim, modal expressivism need not accept this inference. From the fact that modal claims are mere expressions of our linguistic conventions, it does not follow that we must have infallible epistemic access to every feature of the linguistic conventions we endorse.

To see why, recall an observation we made in §3.2: conventional commitments of a speaker may be conditional on other facts obtaining. For example, someone could adopt the following plan for how to use 'planet': use the word 'planet' in whatever way the IAU uses it. To determine which conventions such a speaker is committed to, they then need to find out how the IAU uses the term. Thus, modal expressivism allows for the existence of necessary *a posteriori* truths while explaining how in general we have epistemic access to the modal truths.

The view we are calling modal expressivism has precursors in the literature. It is closely connected to the modal normativism defended by Thomasson (2013).[37] On this view, modal claims "provide a particularly useful way of expressing constitutive semantic and conceptual rules in the object language" (p. 145). Thus, when one asserts 'Necessarily, all bachelors are men', one is expressing a semantic rule to the effect that 'bachelor' only correctly applies to objects to which 'man' correctly applies. Though they have their differences, modal expressivism and Thomasson's modal normativism are allies in the debate over the nature of modality.[38] Both maintain that modal statements are simply expressions of linguistic conventions in some form or other.

[37] See also Thomasson 2017, 2018. For a defense of a similar view, see Sidelle 1989. It is also in the spirit of ontological expressivism as defended by Flocke (2019), as well as the more global form of expressivism defended by Price (2013).

[38] There are two main differences between the views. First, Thomasson denies the existence of counterconventional readings of sentences such as (38) (pp. 153–154). Thus, on her view, conveying semantic norms is particular to modal expressions, whereas on modal expressivism, it is not—what distinguishes modal expressions from others is that this is their *sole* function. Second, the two views differ over compositional semantics. According to Thomasson, the meaning of 'necessarily' is constituted by two rules:

Necessity Intro. If p is an object-language expression of a constitutive semantic rule, then you are entitled to conclude: 'necessarily p', regardless of any subjunctive suppositions.

Necessity Elim. If you have 'necessarily p' as a premise, you may use p as a premise in your reasoning anywhere, under any subjunctive suppositions.

However, while these rules dictate how to *reason* about necessity, they do not fix the *compositional semantic value* of 'necessarily'. This matters when we consider embeddings such as:

(i) Alpha does not believe that it is necessary that all planets clear their orbital neighborhood.

(ii) Something is necessarily human.

By contrast, our proposal offers a clear answer to the compositional semantic value question. Here is the semantic value of the necessity operator '$\Box$':

$$[\![\Box A]\!]^{\mathcal{M}} = \{\langle w, c\rangle \in I^{\mathcal{M}} \mid \forall v \in W^{\mathcal{M}}\colon\ \langle v, c\rangle \in [\![A]\!]^{\mathcal{M}}\}$$

That is, $\Box A$ is true at i just in case A is true no matter how we change the world parameter of i. Using the standard semantics for 'Alpha believes that', we can state the truth conditions for (i) as follows: (i) is true iff there is a world-hyperconvention pair $\langle w, c\rangle$ compatible with Alpha's doxastic state (including their conventional commitments) such that for some world v, the sentence 'all planets clear their orbital neighborhood' is false at v as interpreted by c. A similar story could be told about quantifiers if we enrich indices with a variable assignment in the standard way.

Unlike Kripke's argument in defense of the necessity of identity, then, Lewy's argument against modal conventionalism seems to lose its force once Conventional Wisdom is dropped. Modal conventionalists can vindicate their ideas by rebranding themselves as modal expressivists. In doing so, they need not accept any bizarre consequences to the effect that we as language-users have supernatural control over mathematics, or that all necessary truths must be scrutable to us. The fact that modal claims are merely expressions of our conventional commitments does not entail that we are omniscient about those conventional commitments.

6 Conclusion

According to the conventional wisdom, there is an important distinction between the language we use to describe a scenario and the language that is used in that scenario. When *we* describe a scenario, we must use *our* language rather than the language that speakers in that scenario would use. As Kripke puts it, even when we describe a scenario where we are all speaking German, we describe it in English. Thus, the linguistic conventions used to calculate the compositional semantic value of an expression always remain anchored to actuality. Embedding expressions never induce a revision to those conventions. Language is safe from c-monsters.

On the contrary: the seas of language are crawling with c-monsters. Counterfactuals, along with many other expressions, shift the conventions used to evaluate their constituents in a wide variety of ordinary conversational contexts, including metalinguistic negotiations. These cases expose the flaw in the conventional wisdom. Fortunately, we can generalize standard semantic and pragmatic frameworks in a way that holds onto their advantages, while removing their commitment to Conventional Wisdom. On the pragmatic side, we generalize Stalnaker's framework to allow for communication not only about worldly facts, but also about how to use words—a move that mirrors how classificatory uses of gradable adjectives have been modeled. On the semantic side, we introduce (hyper)conventions as a shiftable parameter of the index of evaluation, thus accounting for c-monsters within a modified Lewis-Stalnaker semantics for counterfactual conditionals. Our best semantic and pragmatic theoretical tools, it turns out, do not crucially rely on holding fixed our actual conventions. There may be c-monsters, but they can be tamed.

The existence of c-monsters in natural language points to a more nuanced picture of linguistic competence than is standardly employed. On the standard view, understanding a language consists in simply understanding the rules governing how to use particular expressions. It is often assumed that a language can be *identified* with a collection of such rules, which are settled in advance by the linguistic community. It is as if, once we settle the conventional interpretation of a word, it becomes impossible for speakers to use that word in a way that is incompatible with that interpretation while still speaking the same language. Purported examples of c-monsters, on this view, are just cases where speakers are speaking a different language.

From our perspective, this relies on an overly rigid conception of what a language is. One could, of course, use the term "language" so narrowly that any two speakers who did not associate every word in the lexicon with the same exact meanings would count as speaking "different languages". But this would be a highly specialized notion, of limited interest to the study of natural language.[39] No one actually has this conception of a language in mind when talking about natural languages such as English. The boundaries of natural languages are vague. The rules and conventions governing how to use words are neither fully determinate nor static. On this more realistic conception of a language, speakers need only agree on how to talk to a sufficient extent to count as speaking the same language. Whether two people are speaking the same language is best viewed as a matter of degree.

Given the vague and ever-changing boundaries of language, speakers need to be flexible enough to know how to use language according to alternative conventions. A true chess master knows how to play chess well not only in normal circumstances but also under a variety of abnormal circumstances, e.g., when they are handicapped a piece, or when the pieces on the back rank are shuffled randomly. Thus, it is not surprising that natural languages like English include mechanisms for interpreting words under alternative conventions. C-monsters are manifestations of a broad linguistic competence—not only to use a finite stock of words with fixed meanings, but to communicate and negotiate with a flexible and evolving language.

[39] Compare Davidson 1986, p. 174: "There is no such a thing as a language, not if a language is anything like what many philosophers and linguists have supposed. [. . .] We must give up the idea of a clearly defined shared structure which language users acquire and then apply to cases. And we should try again to say how convention in any important sense is involved in language; or, as I think, we should give up the attempt to illuminate how we communicate by appeal to conventions." As should be clear, we agree with the negative point about language as philosophers and linguists have traditionally conceived of it. However, our conventionalist semantics offers a way to reject the overly rigid traditional view while still giving conventions a prominent and systematic—if unconventional—place in our theory of meaning and communication. (See Armstrong 2016 for related discussion.)

Acknowledgements

For their comments and suggestions on drafts, we wish to thank Melissa Fusco, Cameron Domenico Kirk-Giannini, Ted Locke, John MacFarlane, Sally McConnell-Ginet, Patrick Muñoz, David Plunkett, Daniel Proske, Matthew Shields, James Walsh, and two anonymous reviewers for this journal. For helpful discussion, thanks also to Chris Barker, Mathias Boehm, David Boylan, David Braun, Kelly Gaus, Jeremy Goodman, Nat Hansen, Chris Kennedy, Harvey Lederman, Adam Murray, Sarah Murray, Sven Neth, Deniz Rudin, Will Starr, and Zoltán Szabó. Earlier versions of this paper were presented at PhLiP 5 (2018), the 4th Taiwan Philosophical Logic Colloquium at National Taiwan University (2018), the Stanford/CSLI Philosophy of Language & Linguistics Workshop (2019), the Richard Wollheim society at UC Berkeley (2019), the Alabama Philosophical Society (2019), the Cornell Semantics Reading Group (2019), and the Eastern APA (2020). The paper was also workshopped in John MacFarlane's seminar at UC Berkeley in Spring 2020. We are grateful for the feedback from audiences at all of these venues. Special thanks to Emily Perry for her alliterative twist to (36-a) and to Sven Neth for allowing us to include his apt description of our project: "It's like jettisoning all of the progress we've made in philosophy."

References

Anand, Pranav and Nevins, Andrew. 2004. "Shifty Operators in Changing Contexts." *Semantics and Linguistic Theory* 14:20–37.

Armstrong, Joshua. 2013. *Language Change in Context*. Ph.D. thesis, Rutgers University.

—. 2016. "The problem of lexical innovation." *Linguistics and Philosophy* 39:87–118.

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.

—. 2013. "Negotiating Taste." *Inquiry* 56:240–257.

Brown, Mike. 2010. *How I Killed Pluto and Why It Had It Coming*. Spiegel & Grau.

Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4:73–122.

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

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

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Coppock, Elizabeth. 2018. “Outlook-based semantics.” *Linguistics and Philosophy* 41:125–164.

Davidson, Donald. 1986. “A Nice Derangement of Epitaphs.” In Richard Grandy and Richard Warner (eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, 157–174. Oxford: Clarendon Press.

Deal, Amy Rose. 2019. *A Theory of Indexical Shift: Meaning, Grammar, and Crosslinguistic Variation*. MIT Press.

DeVore, Edna. 2006. “Planetary Politics: Protecting Pluto.” https://www.space.com/2855-planetary-politics-protecting-pluto.html. Accessed: 2020-04-27.

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

Flocke, Vera. 2019. “Ontological Expressivism.” In James Miller (ed.), *The Language of Ontology*. Oxford University Press. Forthcoming.

Gallie, Walter Bryce. 1955. “Essentially Contested Concepts.” *Proceedings of the Aristotelian Society* 56:167–198.

Gaus, Kelly. 2020. “Counterfactuals and Counterconventionals.” Manuscript.

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

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

Grice, Paul H. 1975. “Logic and Conversation.” In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, 3: Speech Acts*, 41–58. New York: Academic Press.

Hansen, Nat. 2019. “Metalinguistic Proposals.” *Inquiry* 1–19. https://doi.org/10.1080/0020174X.2019.1658628.

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.

Kaplan, David. 1977. “Demonstratives.” In *Themes from Kaplan*, 481–563. Oxford University Press.

Karmo, Toomas. 1983. “Contingent Non-identity.” *Australasian Journal of Philosophy* 61:185–187.

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

Koch, Steffen. 2018. “The externalist challenge to conceptual engineering.” *Synthese* 1–22. https://doi.org/10.1007/s11229-018-02007-6.

Kocurek, Alexander W. 2018. “Counteridenticals.” *Philosophical Review* 127:323–369.

Kocurek, Alexander W. and Jerzak, Ethan. 2020. “Counterlogicals as Counterconventionals.” Manuscript.

Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kyburg, Alice and Morreau, Michael. 2000. “Fitting Words: Vague Language in Context.” *Linguistics and Philosophy* 23:577–597.

Lassiter, Daniel. 2008. “Semantic Externalism, Language Variation, and Sociolinguistic Accommodation.” *Mind & Language* 23:607–633.

Lewis, David K. 1969. *Convention*. Cambridge: Harvard University Press.

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

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

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

Lewy, Casimir. 1976. *Meaning and Modality*. Cambridge: Cambridge University Press.

Ludlow, Peter. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford: Oxford University Press.

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

—. 2018. "Constraint Semantics." Manuscript.

Maier, Emar. 2014. "Mixed Quotation." *Semantics & Pragmatics* 7:1–64.

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.

Muñoz, Patrick. 2019a. "Behavioral attitudes in hyperintensional grammar." Manuscript.

—. 2019b. *On Tongues: The Grammar of Experiential Evaluation*. Ph.D. thesis, University of Chicago.

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." *Philosophers' Imprint* 13:1–37.

Price, Huw. 2013. *Expressivism, Pragmatism and Representationalism*. Cambridge University Press.

Putnam, Hilary. 1973. "Meaning and Reference." *Journal of Philosophy* 70:699–711.

Rabern, Brian. 2012. "Against the identification of assertoric content with compositional value." *Synthese* 189:75–96.

Rabern, Brian and Ball, Derek. 2019. "Monsters and the Theoretical Role of Context." *Philosophy and Phenomenological Research* 98:392–416.

Santorio, Paolo. 2012. "Reference and Monstrosity." *Philosophical Review* 121:359–406.

Schlenker, Philippe. 2003. "A Plea for Monsters." *Linguistics and Philosophy* 26:29–120.

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

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

Shan, Chung-chien. 2010. "The Character of Quotation." *Linguistics and Philosophy* 33:417–443.

Sidelle, Alan. 1989. *Necessity, Essence, and Individuation: A Defense of Conventionalism*. Cornell University Press.

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In Nicholas Rescher (ed.), *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*, 98–112. Oxford: Basil Blackwell Publishers.

—. 1970. "Pragmatics." *Synthese* 22:272–289.

—. 1978. "Assertion." In Peter Cole (ed.), *Pragmatics (Syntax and Semantics, vol. 9)*, 315–332. New York: Academic Press. (Reprinted in Stalnaker 1999, 78–95).

—. 1999. *Context and Content*. Essays on Intentionality in Speech and Thought. Oxford University Press.

—. 2002. "Common Ground." *Linguistics and Philosophy* 25:701–721.

State of Illinois, 96th General Assembly. 2009. "Bill Status of SR0046." http://www.ilga.gov/legislation/BillStatus.asp?DocNum=46&GAID=10&DocTypeID=SR&LegId=40752&SessionID=76&GA=96. Accessed: 2020-04-27.

State of New Mexico, 48th Legislature. 2007. "House Joint Memorial 54." https://www.nmlegis.gov/Sessions/07%20Regular/memorials/house/HJM054.html. Accessed: 2020-04-27.

Sterken, Rachel Katharine. 2019. "Linguistic Intervention and Transformative Communicative Disruptions." In Herman Cappelen and David

Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Thomasson, Amie L. 2013. "Norms and Necessity." *The Southern Journal of Philosophy* 51:143–160.

—. 2017. "Metaphysical Disputes and Metalinguistic Negotiation." *Analytic Philosophy* 58:1–28.

—. 2018. "How can we come to know metaphysical modal truths?" *Synthese* 1–30. https://doi.org/10.1007/s11229-018-1841-5.

Wright, Crispin. 1985. "In Defense of Conventional Wisdom." In Ian Hacking (ed.), *Exercises in Analysis: Essays by Students of Casimir Lewy*, 171–198. Cambridge University Press.

Yalcin, Seth. 2012. "Bayesian Expressivism." *Proceedings of the Aristotelian Society* 112:123–160.

Yli-Vakkuri, Juhani. 2013. "Propositions and Compositionality." *Philosophical Perspectives* 27:526–563.