Hyperlogic: A System for Talking about Logics*

Alexander W. Kocurek[1]

Cornell University, Ithaca, NY, USA
awk78@cornell.edu

Abstract

Sentences about logic are often used to show that certain embedding expressions, including attitude verbs, conditionals, and epistemic modals, are hyperintensional. Yet it not clear how to regiment "logic talk" in the object language so that it can be compositionally embedded under such expressions. This paper does two things. First, it argues against a standard account of logic talk, viz., the impossible worlds semantics [2]. It is shown that this semantics does not easily extend to a language with propositional quantifiers, which are necessary for regimenting some logic talk. Second, it develops an alternative framework based on logical expressivism, which explains logic talk using shifting conventions [6]. When combined with the standard **S5**π+ semantics for propositional quantifiers, this framework results in a well-behaved system that does not face the problems of the impossible worlds semantics. It can also be naturally extended with hybrid operators [1] to regiment a broader range of logic talk, e.g., claims about what laws hold according to other logics. The resulting system, called *hyperlogic*, is therefore a better framework for modeling logic talk than previous accounts.

1 Introduction

Sentences like (1)–(3) suggest that attitude verbs, conditionals, and epistemic modals are hyperintensional, i.e., they do not validate the replacement of necessary (or even logical) equivalents.

(1) Inej believes intuitionistic logic is the correct logic.

(2) If the Liar were both true and not true, the law of non-contradiction would fail.

(3) Classical logic might not be correct.

In order to develop a semantics for these expressions that captures their hyperintensionality, we first need a way of regimenting sentences like 'intuitionistic logic is the correct logic' and 'the law of non-contradiction fails' in the object language so that they may be meaningfully and non-trivially compositionally embedded. Yet it is not entirely clear how this can be done. We cannot, for instance, simply regiment (2) as $(l \wedge \neg l) \boxdot\!\!\rightarrow \neg \vdash \neg(p \wedge \neg p)$, since this illictly imports notation from the metalanguage into the object language. We might instead try to regiment (2) as $(l \wedge \neg l) \boxdot\!\!\rightarrow \neg lnc$, where lnc is some primitive atomic standing for the law of non-contradiction. But this is unsatisfactory, as atomic formulas are generally assumed to be logically contingent, whereas the law of non-contradiction, intuitively, is not.

This paper does two things. First, it presents and argues against a standard hyperintensional framework for modeling logic talk, viz., the ***impossible worlds semantics***, which introduces logically impossible worlds to accommodate logic talk [2]. The main problem with this approach is that it cannot be easily extended to a language with propositional quantifiers [4]. Propositional quantifiers are useful, and even necessary in some cases, for regimenting laws

*Thanks to Yifeng Ding, Melissa Fusco, Wes Holliday, Ethan Jerzak, and Rachel Rudolph, as well as the participants of the 11th Semantics and Philosophy in Europe conference for helpful discussion.

of logic in the object language. But it turns out that when we try to interpret propositional quantifiers in the presence of logically impossible worlds, we run into serious troubles.

Second, this paper presents an alternative to the impossible worlds semantics. This system, which I call ***hyperlogic***, is inspired by a philosophical view on the nature of logic known as ***logical expressivism***, which explains logic talk by appealing to shifting conventions rather than impossible worlds [6]. It turns out that logical expressivism, when combined with the standard **S5**$\pi+$ semantics for propositional quantifiers, does not face any of the problems that plagued the impossible worlds semantics, and generally results in a nice, well-behaved system. What's more, this system can be naturally extended to a more expressive language that can accommodate a broader range of logic talk. In particular, we can introduce operators borrowed and modified from hybrid logic [1] to regiment claims about what laws hold *according to* other logics and to regiment the distinction between axioms and rules in the object language.

2 The Impossible Worlds Semantics

To begin, let's review the impossible worlds semantics and its account of logic talk. To simplify the discussion, I will only focus on counterfactuals. The points made about counterfactuals easily extend to attitude verbs and modals more generally, but I leave that to future work.

We start with a simple base language $\mathcal{L}_0$ consisting of an infinite stock of propositional variables $\mathsf{Prop} = \{p_1, p_2, p_3, \ldots\}$, all the standard boolean connectives ($\neg$, $\wedge$, $\vee$, $\rightarrow$), a pair of modalities ($\Box$ and $\Diamond$), and a counterfactual operator ($\boxright$). The syntax of $\mathcal{L}_0$ is summarized in Backus-Naur form as follows:

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

On the standard, intensional semantics for counterfactuals, $\phi \boxright \psi$ is true iff all of the closest possible ϕ-worlds are ψ-worlds [8, 10]. The impossible worlds semantics takes this idea and adds a twist: $\phi \boxright \psi$ is true iff all of the closest ϕ-worlds, whether or not those worlds are possible, are ψ-worlds [2]. In other words, the impossible worlds semantics differs from the standard semantics in allowing the closest ϕ-worlds to include impossible worlds.

What is an "impossible" world? First, think of a world as a kind of ersatz entity: a world (possible or not) is just a set of formulas. The members of an ersatz world intuitively represent what is true according to that world—that is, ϕ is true at an ersatz world w iff $\phi \in w$. On this understanding of a world, a possible world is just a special kind of set, viz., one that is maximally compossible, whereas an impossible world is just a set that is *not* maximally compossible. Thus, impossible worlds are not a wholly new or alien kind of entity: we are already committed to them if we accept ersatz possible worlds [9].

Using this conception of impossible worlds, here is the impossible worlds semantics for $\mathcal{L}_0$. First, an ***impossible worlds model*** is a quadruple of the form $\mathcal{I} = \langle W, P, f, V \rangle$, where:

- $W \neq \varnothing$ is the set of *worlds*;
- $\varnothing \neq P \subseteq W$ is the set of *possible worlds*;
- $f \colon \wp W \times W \rightarrow \wp W$ is the *selection function*;
- V is the *valuation function*, where $V \colon \mathsf{Prop} \times P \rightarrow \{0, 1\}$ and $V \colon \mathcal{L}_0 \times \overline{P} \rightarrow \{0, 1\}$.

Intuitively, $f(X, w)$ is the set of worlds "closest" to w in X. Various constraints may be placed on f if desired; e.g., many authors require $f(X, w) \subseteq X$ (corresponding to $\phi \boxright \phi$) or $w \in f(X, w)$ if $w \in X$ (corresponding to $(\phi \boxright \psi) \rightarrow (\phi \rightarrow \psi)$). Even simple constraints like

these, however, are controversial in the context of counterlogicals [3, 9]. For our purposes, it does not particularly matter what constraints we impose so long as $f(X, w)$ can contain some impossible worlds, i.e., so long as we don't require that $f(X, w) \subseteq P$ for all X and w.

Valuation functions determine the truth of *every* formula at impossible worlds. This is because, in order to model logic talk, the impossible worlds semantics needs to appeal to *logically* impossible worlds. In general, though, there is no single rule for determining the truth of complex formulas from the truth of atomics that applies to every logically impossible world. When it comes to the logically impossible, anything goes: there are impossible worlds governed by logics where conjunction is equivalent to disjunction, where negations are redundant, or even where everything is true. Any collection of formulas constitute a world and can be said to conform to some wacky logic or other. Such impossible worlds are *strange*, for sure, but nothing in the intuitive conception of an impossible world rules them out. So for impossible worlds, truth must be determined by fiat via the valuation function.

Given an impossible worlds model $\mathcal{I} = \langle W, P, f, V\rangle$ and a $w \in W$, we define ***satisfaction*** $\Vdash_{\mathbf{I}}$ as follows. First, if $w \in \overline{P}$, then $\mathcal{I}, w \Vdash_{\mathbf{I}} \phi$ iff $V(\phi, w) = 1$. Second, if $w \in P$, then satisfaction is defined recursively (where $[\![\phi]\!]^{\mathcal{I}} = \{u \in W \mid \mathcal{I}, u \Vdash_{\mathbf{I}} \phi\}$):

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

Finally, given a set $\Gamma \subseteq \mathcal{L}_0$ and a $\phi \in \mathcal{L}_0$, we say Γ ***I-entails*** ϕ, or $\Gamma \models_{\mathbf{I}} \phi$, if for every impossible worlds model $\mathcal{I} = \langle W, P, f, V\rangle$ and every $w \in P$, if $\mathcal{I}, w \Vdash_{\mathbf{I}} \Gamma$, then $\mathcal{I}, w \Vdash_{\mathbf{I}} \phi$. In other words, consequence is satisfaction-preservation over *possible* worlds. This ensures that $\models_{\mathbf{I}}$ is an extension of classical logic even though the constituents of counterfactuals may behave nonclassically. Thus, $\models_{\mathbf{I}} p \vee \neg\, p$, even though $\not\models_{\mathbf{I}} \phi \,\Box\!\!\rightarrow (p \vee \neg\, p)$, since the closest ϕ-worlds may include some impossible worlds w where $V(p \vee \neg\, p, w) = 0$.

3 Adding Propositional Quantifiers

Now we will consider extending our base language $\mathcal{L}_0$ with propositional quantifiers. Why? Abstractly, of course, it would be unfortunate for the impossible worlds semantics if it could not be extended with propositional quantifiers. But also, there are several reasons specific to logic talk for introducing propositional quantifiers into the language. I will mention two.

First, propositional quantifiers are useful for regimenting *laws* of logic. As it stands, the best we can do in $\mathcal{L}_0$ is pick an arbitrary propositional variable l for each law and simply stipulate, in the metalanguage, that l stands for that law. This is less than ideal. We would like to capture the sense in which, say, the law of excluded middle is a valid principle by saying not just that it is *true* but that it is *logically necessary*. So where *lem* stands for the law of excluded middle, we would like to say that $\models \mathit{lem}$. Propositional variables are logically contingent, however, in that there are no constraints on what truth values we can assign to them. Unless we impose further *ad hoc* constraints on our class of models, it won't be the case that $\models \mathit{lem}$.

Second, some logic talk seems to require propositional quantifiers. Here are some examples:

(4) No contradiction is true.

(5) Everything that is intuitionistically valid is classically valid.

(6) Some propositions are neither true nor not true.

(7) If the Liar were both true and not true, everything would be true.

What's more, we need propositional quantifiers to make familiar distinctions between *de dicto* and *de re* counterfactuals. Contrast the following:

(8) a. There is a contradiction such that, if it were true, everything would be true and not true.
b. If there were a true contradiction, everything would be true and not true.

These do not seem equivalent. Intuitively, (8-a) seems true: e.g., the conjunction of "Everything would be true and not true" and its negation would be one such contradiction. But (8-b) seems false (or at least could be false): if there were a true contradiction, paraconsistent logic would be correct, so not everything would be true and not true. Regimenting laws as brute atomic formulas leaves us ill-equipped for distinguishing such counterfactuals.

Propositional quantifiers allows us to avoid these problems. Instead of choosing an arbitrary propositional variable to represent the law of excluded middle, we could regiment it as a universally quantified claim, viz., $p \vee \neg p$ necessarily holds *for any proposition* p:

$$\forall p \, \Box(p \vee \neg p).$$

Assuming we define the semantics correctly, we will not need to stipulate that this formula is valid: it will simply fall out of the semantics for propositional quantifiers that $\vDash \forall p \, \Box(p \vee \neg p)$. This can be so even if counterfactuals are hyperintensional, and so non-trivially embed in counterfactuals. So writing the law of non-contradiction as $\forall p \, \Box \neg(p \wedge \neg p)$, (2) becomes:

$$(l \wedge \neg l) \mathrel{\Box\!\!\rightarrow} \neg \forall p \, \Box \neg(p \wedge \neg p).$$

Propositional quantifiers also make it easier to regiment the quantified examples above. For instance, (4) and (7) could be regimented respectively as:

$$\neg \exists p(p \wedge \neg p)$$
$$(l \wedge \neg l) \mathrel{\Box\!\!\rightarrow} \forall pp.$$

Moreover, the distinction between (8-a) and (8-b) can be captured using scope:

$$\exists p((p \wedge \neg p) \mathrel{\Box\!\!\rightarrow} \forall q(q \wedge \neg q))$$
$$(\exists p(p \wedge \neg p) \mathrel{\Box\!\!\rightarrow} \forall q(q \wedge \neg q)).$$

Thus, it makes sense to consider extending $\mathcal{L}_0$ with propositional quantifiers. So let's extend $\mathcal{L}_0$ to a language $\mathcal{L}_{\mathsf{Q}}$ with propositional quantifiers $\forall p$ and $\exists p$ binding into sentence position:

$$\phi ::= p \mid \neg \phi \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid (\phi \rightarrow \phi) \mid \Box \phi \mid \Diamond \phi \mid (\phi \mathrel{\Box\!\!\rightarrow} \phi) \mid \forall p\phi \mid \exists p\phi.$$

The simplest semantics for modal logic with propositional quantifiers (**S5**π+) interprets the quantifiers as ranging over arbitrary *sets of worlds* [4]. The semantics looks something like this:

$$\mathcal{M}, w \Vdash \forall p\phi \quad \Leftrightarrow \quad \text{for all } X \subseteq W\text{: } \mathcal{M}^p_X, w \Vdash \phi$$
$$\mathcal{M}, w \Vdash \exists p\phi \quad \Leftrightarrow \quad \text{for some } X \subseteq W\text{: } \mathcal{M}^p_X, w \Vdash \phi$$

where $\mathcal{M}^p_X = \langle W, R, V^p_X\rangle$ and V^p_X is exactly like V except that $V^p_X(p) = X$. In words, $\forall p\phi$ is true just in case ϕ comes out true on any way of interpreting p, where "a way of interpreting p" is just an assignment of p to a possible worlds proposition (i.e., a set of worlds).

A natural strategy for extending the impossible worlds semantics to $\mathcal{L}_{\mathsf{Q}}$ is to simply import the **S5**π+ semantic entries directly. Thus, where $w \in P$, where $\mathcal{I}^p_X = \langle W, P, f, V^p_X\rangle$, and where V^p_X is exactly like V except that for all $w \in W$, $V^p_X(p, w) = 1$ iff $w \in X$:

$$\begin{aligned} \mathcal{I}, w \Vdash_{\mathbf{I}} \forall p\phi &\quad\Leftrightarrow\quad \text{for all } X \subseteq W\text{: } \mathcal{I}^p_X, v \Vdash_{\mathbf{I}} \phi \\ \mathcal{I}, w \Vdash_{\mathbf{I}} \exists p\phi &\quad\Leftrightarrow\quad \text{for some } X \subseteq W\text{: } \mathcal{I}^p_X, v \Vdash_{\mathbf{I}} \phi \end{aligned}$$

Notice, however, that as it's defined, V^p_X only differs from V in the interpretation of p. V^p_X does not differ from V on *complex* formulas involving p, at least at impossible worlds: for instance, if $w \in \overline{P}$, then $V^p_X(\neg p, w) = V(\neg p, w)$ regardless of X. In the **S5**π+ semantics, this doesn't matter since the interpretation of a complex formula involving p is recursively determined from the interpretation of p. But in the impossible worlds semantics, truth at impossible worlds is not determined in a recursive manner: it's determined *by fiat* by the valuation function. So we cannot, in general, determine how to change the interpretation of, say, $\neg p$ or $p \vee q$ at an arbitrary impossible world when we change the interpretation of p.

Not only is this counterintuitive, it leads to formal difficulties. For instance, the following is valid on the current semantics for $\mathcal{L}_{\mathsf{Q}}$:

$$\exists p((p \wedge \neg p) \boxdot\!\!\rightarrow q) \rightarrow \forall p((p \wedge \neg p) \boxdot\!\!\rightarrow q).$$

Since $[\![p \wedge \neg p]\!]^{\mathcal{I}} \subseteq \overline{P}$ for any $\mathcal{I}$ and since $V^p_X(p \wedge \neg p, w) = V(p \wedge \neg p, w)$ for any $w \in \overline{P}$, it follows that $[\![p \wedge \neg p]\!]^{\mathcal{I}^p_X} = [\![p \wedge \neg p]\!]^{\mathcal{I}}$, and so $f([\![p \wedge \neg p]\!]^{\mathcal{I}^p_X}, w) = f([\![p \wedge \neg p]\!]^{\mathcal{I}}, w)$ for any $X \subseteq W$. Hence, if $f([\![p \wedge \neg p]\!]^{\mathcal{I}^p_X}, w) \subseteq [\![q]\!]^{\mathcal{I}^p_X} = [\![q]\!]^{\mathcal{I}}$ for one $X \subseteq W$, it holds for all X. But this principle is implausible: just because *one* contradiction counterfactually implies q, it does not follow that *all* contradictions do.

Clearly, the solution to this problem will involve defining V^p_X so that it differs from V not just on the interpretation p but also on the interpretation of complex formulas involving p. The trouble is that it is not clear how to do this. Again, when it comes to the logically impossible, anything goes. Even if p is true and $\neg p$ is false at an impossible world, making p false at that world does not automatically mean we must make $\neg p$ true at that world: we might be at a world that allows a sentence and its negation to be false.

Here is one promising line of thought. The problem seems to stem from our conception of impossible worlds as fixed sets of formulas. Perhaps impossible worlds need to also be equipped with a *rule* for how truth is determined, which we could model as a function from an interpretation of the propositional variables to an interpretation of complex formulas. Such rules need not be "natural"; an impossible world governed by a wacky logic will have a wacky rule for determining truth. But however wacky, such a rule will help us interpret propositional quantifiers by making sure we are not at a loss for how to reinterpret complex formulas when we reinterpret propositional variables.

More precisely, define a ***variable assignment*** over W to be a function $g\colon \mathsf{Prop} \rightarrow \wp W$ mapping each propositional variable to a proposition. Where VA_W is the set of variable assignments over W, let's redefine a valuation function to be a map $V\colon \mathcal{L}_{\mathsf{Q}} \times \overline{P} \times \mathsf{VA}_W \rightarrow \{0, 1\}$ such that $V(p, w, g) = 1$ iff $w \in g(p)$. In other words, V is a rule for determining what complex formulas are true at an impossible world *given* an interpretation of the propositional variables. Likewise, let's redefine impossible worlds models as quadruples $\mathcal{I} = \langle W, P, f, V\rangle$ with our new valuation functions. We now relativize truth to a world and a variable assignment. So if $w \in \overline{P}$,

then $\mathcal{I}, w, g \Vdash_{\mathbf{I}} \phi$ iff $V(\phi, w, g) = 1$. Otherwise, the truth conditions are as before (except relativized to variable assignments) with the following amendments:

$$\begin{aligned}
\mathcal{I}, w, g \Vdash_{\mathbf{I}} p &\quad\Leftrightarrow\quad w \in g(p) \\
\mathcal{I}, w, g \Vdash_{\mathbf{I}} \forall p\phi &\quad\Leftrightarrow\quad \text{for all } X \subseteq W\text{: } \mathcal{I}, w, g^p_X \Vdash \phi \\
\mathcal{I}, w, g \Vdash_{\mathbf{I}} \exists p\phi &\quad\Leftrightarrow\quad \text{for some } X \subseteq W\text{: } \mathcal{I}, w, g^p_X \Vdash \phi.
\end{aligned}$$

This semantics does not validate $\exists p((p \wedge \neg p) \boxdot\!\!\rightarrow q) \rightarrow \forall p((p \wedge \neg p) \boxdot\!\!\rightarrow q)$, since $V(p \wedge \neg p, w, g^p_X)$ need not be the same for every $X \subseteq W$.

Unfortunately, this proposal faces further problems. Whereas before the impossible worlds semantics validated too much, now it validates too little. For example, it does not validate existential introduction ($\phi(\chi) \vDash \exists p\phi(p)$, subject to the usual restrictions), variable exchange ($\exists p\phi(p) \vDash \exists q\phi(q)$, subject to the usual restrictions), or vacuous quantification ($\phi \vDash \forall p\phi$ where p does not occur free in ϕ). To illustrate, none of the following are valid:

$$\begin{aligned}
((q \wedge \neg q) \boxdot\!\!\rightarrow r) &\rightarrow \exists p((q \wedge p) \boxdot\!\!\rightarrow r) \\
\exists p(p \boxdot\!\!\rightarrow r) &\rightarrow \exists q(q \boxdot\!\!\rightarrow r) \\
(q \boxdot\!\!\rightarrow r) &\rightarrow \forall p(q \boxdot\!\!\rightarrow r).
\end{aligned}$$

To see why, consider the first principle. $\exists p((q \wedge p) \boxdot\!\!\rightarrow r)$ is true at w iff for some $X \subseteq W$, $f([\![q \wedge p]\!]^{\mathcal{I}, g^p_X}, w) \subseteq [\![r]\!]^{\mathcal{I}, g^p_X}$. But even if $X = [\![\neg q]\!]^{\mathcal{I}, g}$, we cannot be sure that $f([\![q \wedge p]\!]^{\mathcal{I}, g^p_X}, w) = f([\![q \wedge \neg q]\!]^{\mathcal{I}, g}, w)$ since it need not be that $V(q \wedge \neg q, w, g) = V(q \wedge p, w, g^p_X)$. And this gap can be exploited to construct a counterexample to the first principle.

We could try to impose various constraints on V to avoid these problems. But it is not obvious how to do this in a systematic fashion. For each such problem, we would need to impose an additional constraint on V to block that specific problem. Such a gerrymandered approach seems undesirable, to say the least. What's more, it is not clear what would conceptually motivate such constraints apart from the fact that they help avoid these technical problems.

4 Logical Expressivism

The previous section outlined a problem with the impossible worlds semantics: the presence of logically impossible worlds makes it difficult to interpret propositional quantifiers. I will now show how a new approach, viz., logical expressivism, does better.

Logical expressivism is motivated by the thought that counterlogicals seem to involve shifts in the meaning of the logical connectives [6]. For example, consider the following inference:

(9) a. If intuitionistic logic were correct, the continuum hypothesis would not not be true.
b. ∴ If intuitionistic logic were correct, the continuum hypothesis would be true.

Intuitively, this inference seems invalid because in the consequent of (9-a), we are interpreting 'not' according to an intuitionistic interpretation, which does not validate the law of double negation elimination. If we held fixed the actual, classical meaning of 'not', then the inference would be valid. The antecedent 'if intuitionistic logic were correct' seems to be triggering a shift in the meaning of 'not' from a classical to an intuitionistic one.

This sort of phenomenon arises in ordinary, non-logical examples, too [7]. There is an old joke: if a dog's tail were called a "leg", how many legs would a dog have? The answer is supposed to be four—calling a tail a leg doesn't make it one! At the risk of ruining a (bad) joke by explaining it, the reason this is a joke is that there are two natural readings of (10).

(10) If a dog's tail were called a "leg", a dog would have five legs.

On one reading—call it the ***shifty*** reading—(10) is true because we interpret 'leg' in the consequent according to the conventions described in the antecedent. On another reading—call it the ***rigid*** reading—(10) is false because we interpret 'leg' in the consequent according to our actual conventions, on which a tail does not count as a leg. Both readings are available, though the shifty reading seems more salient in the context of the joke; hence why the joke is "funny".

Logical expressivism holds that the same thing is happening in (9). On this view, logic just is a convention governing logical vocabulary. Just as speakers may adopt any number of conventions for how to talk, so too, they may adopt any number of conventions for how to use words like 'not', 'and', 'or', and so on. There is no such thing as the "correct" or "one true" logic, just as there is no such thing as the "correct" or "one true" language. By adopting a non-classical logic, one is not thereby describing things inaccurately. Logic is more a matter of *decision* than *discovery*. But when speakers interpret counterlogicals, they interpret the logical connectives according to nonclassical conventions, even if they adopt classical logic. That is, counterlogicals, on their most natural interpretation, are really ***counterconventionals*** [6].

We can develop a formal semantics for logical expressivism by adapting another well-known expressivist semantics, viz., the hyperplan semantics for normative discourse due to [5]. Informally, a hyperplan can be thought of as a maximally specific plan, specifying what actions to take in every conceivable situation. Formally, a hyperplan is just a total function from worlds to sets of permissible actions. Gibbard's proposal was to think of normative vocabulary ('ought', 'may', etc.) as being sensitive to hyperplans. Thus, $\ulcorner \alpha$ ought to $\phi \urcorner$ is true relative a world-hyperplan pair $\langle w, h \rangle$ iff ϕing is included amongst the actions in $h(w)$.

Since logical expressivism thinks of a logic as a kind of convention, where a convention can be thought of as a kind of plan for how to use words, logics are, effectively, just special kinds of plans. Thus, we can define a *hyper*convention as a maximally specific plan for how to use words. Formally, we can model hyperconventions as interpretation functions: functions mapping propositional variables, connectives, and so on to intensions. Then a sentence of the form $\ulcorner \phi$ is valid$\urcorner$ is true relative to a world-hyperconvention pair $\langle w, c \rangle$ iff ϕ as interpreted according to c holds in every situation.

To make this more precise, define a ***hyperconvention*** over W to be a function c where:

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

In other words, if we think of a proposition as just a set of worlds, hyperconventions map each propositional variable to a proposition and each n-place connective to an n-ary operation on propositions.[1] An ***index*** over W is a pair $\langle w, c \rangle$ where $w \in W$ and c is a hyperconvention over W. The set of indices over W is denoted Ind_W.

Now for the semantics. An ***expressivist model*** is a pair of the form $\mathcal{E} = \langle W, f \rangle$, where:

- $W \neq \varnothing$ is the set of *possible worlds*;
- $f\colon \wp\mathsf{Ind}_W \times \mathsf{Ind}_W \rightarrow \wp\mathsf{Ind}_W$ is the *selection function*.

Notice that expressivist models do not include a valuation function since hyperconventions can play that role. Notice also that W is described as the set of *possible* worlds. This is because the role of impossible worlds is effectively played by possible worlds-under-descriptions.

[1] Note that this set excludes $\Box\!\!\rightarrow$. Since $\Box\!\!\rightarrow$ denotes an operation on sets of indices, which themselves contain hyperconventions, we cannot also have hyperconventions interpret $\Box\!\!\rightarrow$ without circularity.

Satisfaction is defined relative to indices, i.e., world-hyperconvention pairs. Thus, given an expressivist model $\mathcal{E} = \langle W, f\rangle$ and a $\langle w, c\rangle \in \mathsf{Ind}_W$, we define ***satisfaction*** $\Vdash_{\mathbf{E}}$ as follows (where $\triangle \in \{\neg, \wedge, \vee, \rightarrow, \leftrightarrow, \Box, \Diamond\}$):

$$\begin{array}{lcl} \mathcal{E}, w, c \Vdash_{\mathbf{E}} p & \Leftrightarrow & w \in c(p) \\ \mathcal{E}, w, c \Vdash_{\mathbf{E}} \triangle(\phi_1, \ldots, \phi_n) & \Leftrightarrow & w \in c(\triangle)([\![\phi_1]\!]^{\mathcal{E},c}, \ldots, [\![\phi_n]\!]^{\mathcal{E},c}) \\ \mathcal{E}, w, c \Vdash_{\mathbf{E}} \phi \,\Box\!\!\rightarrow \psi & \Leftrightarrow & f([\![\phi]\!]^{\mathcal{E}}, w, c) \subseteq [\![\psi]\!]^{\mathcal{E}}, \end{array}$$

where $[\![\phi]\!]^{\mathcal{E}} \coloneqq \{\langle w, c\rangle \in \mathsf{Ind}_W \mid \mathcal{E}, w, c \Vdash_{\mathbf{E}} \phi\}$ and $[\![\phi]\!]^{\mathcal{E},c} \coloneqq \{w \in W \mid \mathcal{E}, w, c \Vdash_{\mathbf{E}} \phi\}$.

Finally, consequence is defined as satisfaction-preservation over "classical" indices, i.e., indices that interpret the connectives in the ordinary classical way. More precisely, let's say a hyperconvention c is ***classical*** if the following conditions are met for all $X, Y \subseteq W$:

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

An index is ***classical*** if its hyperconvention is classical. We let CInd_W be the set of classical indices over W. Then we say Γ ***E-entails*** ϕ, or $\Gamma \models_{\mathbf{E}} \phi$, iff for every expressivist model $\mathcal{E}$ and every *classical* $\langle w, c\rangle \in \mathsf{CInd}_W$, if $\mathcal{E}, w, c \Vdash_{\mathbf{E}} \Gamma$, then $\mathcal{E}, w, c \Vdash_{\mathbf{E}} \phi$. Restricting consequence to truth-preservation over classical indices ensures that $\models_{\mathbf{E}}$ is classical. In fact, over $\mathcal{L}_0$, the logic of logical expressivism exactly matches the logic of the impossible worlds semantics (see [6]):

Theorem 1. *For all $\Gamma \subseteq \mathcal{L}_0$ and $\phi \in \mathcal{L}_0$, $\Gamma \models_{\boldsymbol{I}} \phi$ iff $\Gamma \models_{\boldsymbol{E}} \phi$.*

In particular, this means that (i) $\models_{\mathbf{E}}$ is an extension of classical logic, and (ii) the logical expressivist semantics is hyperintensional, i.e., $\neg\Diamond\phi \not\models_{\mathbf{E}} \phi \,\Box\!\!\rightarrow \psi$. Thus, when it comes to $\mathcal{L}_0$, the impossible worlds semantics and logical expressivist semantics are on a par: anything one can do, the other can do as well.

But the logical expressivist semantics does significantly better when extended to $\mathcal{L}_{\mathsf{Q}}$. Let us say an ***expressivist variable assignment*** over W is a function $g\colon \mathsf{Prop} \to \wp\mathsf{Ind}_W$. Here, then, are the semantic clauses for propositional quantifiers in the logical expressivist framework:

$$\begin{array}{lcl} \mathcal{E}, w, c, g \Vdash_{\mathbf{E}} p & \Leftrightarrow & \langle w, c\rangle \in g(p) \\ \mathcal{E}, w, c, g \Vdash_{\mathbf{E}} \forall p\phi & \Leftrightarrow & \text{for all } X \subseteq \mathsf{Ind}_W\colon\ \mathcal{E}, w, c, g^p_X \Vdash \phi \\ \mathcal{E}, w, c, g \Vdash_{\mathbf{E}} \exists p\phi & \Leftrightarrow & \text{for some } X \subseteq \mathsf{Ind}_W\colon\ \mathcal{E}, w, c, g^p_X \Vdash \phi. \end{array}$$

Let's say $\langle w, c, g\rangle$ is ***initialized*** if $\langle w, c\rangle \in g(p)$ iff $w \in c(p)$. If we define consequence as satisfaction-preservation over initialized (classical) points, we do not get any of the counterintuitive consequences that plagued the impossible worlds semantics. On the one hand, $\exists p((p \wedge \neg p) \,\Box\!\!\rightarrow q)$ does not entail $\forall p((p \wedge \neg p) \,\Box\!\!\rightarrow q)$, since $[\![p \wedge \neg p]\!]^{\mathcal{E},g^p_X}$ need not be the same for all $X \subseteq \mathsf{Ind}_W$. On the other hand, this semantics does validate the normal principles governing universal and existential propositional quantifiers, such as existential introduction, variable exchange, and vacuous quantification. So we do validate the principles from the end of § 3 (e.g., $\exists p(p \,\Box\!\!\rightarrow r) \rightarrow \exists q(q \,\Box\!\!\rightarrow r)$). Thus, the logical expressivist framework already does significantly better than the impossible worlds semantics at interpreting propositional quantifiers.

5 Hybrid Logic

In the previous section, we saw that logical expressivism fared better than the impossible worlds semantics when we extended $\mathcal{L}_0$ to $\mathcal{L}_{\mathsf{Q}}$. However, there are two expressive limitations of $\mathcal{L}_{\mathsf{Q}}$

that affect its ability to regiment logic talk. First, $\mathcal{L}_{\mathsf{Q}}$ has no way of expressing claims about whole logics, such as 'intuitionistic logic is correct'. Second, $\mathcal{L}_{\mathsf{Q}}$ cannot regiment claims about what holds *according to* a particular logic. For instnace, even though classical and intuitionistic logicians disagree over (11-a), both parties agree that (11-b) is true:

(11) a. The law of excluded middle holds.
b. According to classical logic, the law of excluded middle holds.

Third, while $\mathcal{L}_{\mathsf{Q}}$ has the resources to regiment *axioms*, it does not have the resources to regiment *rules of inference*. For instance, consider the rule of modus ponens:

$$\phi, \phi \to \psi \therefore \psi$$

The only way to represent this rule in $\mathcal{L}_{\mathsf{Q}}$ is as the propositionally quantified sentence:

$$\forall p \forall q \, \Box((p \wedge (p \to q)) \to q).$$

But this treats modus ponens as an axiom, not as a rule. The distinction matters since some logics accept one but not the other. For example, the strong Kleene logic **K3** rejects modus ponens as an axiom ($\nvdash_{\mathbf{K3}} (\phi \wedge (\phi \to \psi)) \to \psi$) but accepts it as a rule ($\phi, \phi \to \psi \vdash_{\mathbf{K3}} \psi$).

To overcome this last problem, we need a way to distinguish in the object language between *necessarily,* $(p \wedge (p \to q) \to q)$ *is true* and *necessarily, if* p *and* $p \to q$ *are true, then* q *is true.* The key difference is that we are using the *if…,then…* construction of the (classical) *meta*language to state the rule, whereas we are using the *object* language $\to$ to state the axiom. Thus, what we would like is access to the classical interpretation of the connectives in the object language—written, say, as $\sim$, $\&$, $\supset$, etc.—and then to regiment the rule of modus ponens as:

$$\forall p \forall q \, \Box((p \;\&\; (p \to q)) \supset q).$$

Intuitively, we want this to say that for any propositions p and q, necessarily, if p and $p \to q$ are true, then q is true—rather than say for any p and q, necessarily, $(p \wedge (p \to q)) \to q$ is true.

Fortunately, all three limitations can be overcome by extending to a language with hybrid operators [1]. In short, we'll extend $\mathcal{L}_{\mathsf{Q}}$ with an infinite stock of ***interpretation variables*** $\mathsf{IVar} = \{i_1, i_2, i_3, \ldots\}$ as new atomic formulas (intuitively standing for particular logics), and with two new hybrid operators @ ("according to") and $\downarrow$ (a variable binding operator). The resulting syntax $\mathcal{L}_{\mathsf{QH}}$ is summarized as follows:

$$\phi ::= p \mid \neg \phi \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid (\phi \to \phi) \mid \Box \phi \mid \Diamond \phi \mid (\phi \,\Box\!\!\to\, \phi) \mid \forall p \phi \mid \exists p \phi \mid i \mid @_i \phi \mid \downarrow i.\phi$$

Very roughly, we can think of i as standing for "i is the correct logic", $@_i\phi$ for "according to i, ϕ", and $\downarrow i.\phi$ for "where i is the current logic, ϕ".

It is easy to extend the expressivist semantics to $\mathcal{L}_{\mathsf{QH}}$. The models are as before. The main difference is that now we will extend variable assignments g to not only assign each $p \in \mathsf{Prop}$ to a set of indices, but also each $i \in \mathsf{IVar}$ to a hyperconvention. The semantics for the new hybrid vocabulary then becomes:

$$\begin{aligned}
\mathcal{E}, w, c, g \Vdash_{\mathbf{E}} i \quad &\Leftrightarrow \quad g(i) = c \\
\mathcal{E}, w, c, g \Vdash_{\mathbf{E}} @_i\phi \quad &\Leftrightarrow \quad \mathcal{E}, w, g(i), g \Vdash_{\mathbf{E}} \phi \\
\mathcal{E}, w, c, g \Vdash_{\mathbf{E}} \downarrow i.\phi \quad &\Leftrightarrow \quad \mathcal{E}, w, c, g^i_c \Vdash_{\mathbf{E}} \phi
\end{aligned}$$

As before, consequence is defined as satisfaction preservation over initialized classical points. Call the resulting semantics ***hyperlogic***.

Hyperlogic has the expressive resources to overcome the limitations mentioned above. For instance, suppose we single out an interpretation variable k to stand for our initial, classical hyperconvention (we can always rewrite formulas with k into equivalent formulas without it). Then we can regiment the difference between (11-a) and (11-b) as:

$$\forall p \, \Box(p \vee \neg p)$$
$$@_k \forall p \, \Box(p \vee \neg p).$$

To account for the difference between axioms and rules, we can use the binder $\downarrow$ to grant us access, in the object language, to the classical interpretation of connectives in the scope of operators (such as @) that might shift the underlying logic. So we can define $\sim$, $\&$, $\supset$, and so on as follows (where i does not occur in ϕ or ψ):

$$\sim \phi \coloneqq \downarrow i.@_k \neg @_i \phi$$
$$(\phi \ \& \ \psi) \coloneqq \downarrow i.@_k(@_i \phi \wedge @_i \psi)$$
$$(\phi \supset \psi) \coloneqq \downarrow i.@_k(@_i \phi \rightarrow @_i \psi).$$

This means we actually can just regiment the rule of modus ponens as $\forall p \forall q \, \Box((p \ \& \ (p \rightarrow q)) \supset q)$. After a little simplifying, this becomes:

$$\forall p \forall q \, \Box \downarrow i.((@_i p \wedge @_i(p \rightarrow q)) \rightarrow @_i q).$$

Thus, hyperlogic offers a promising new framework for regimenting a broad range of logic talk, including claims about axioms, rules, and even entire logics, so that such talk can be compositionally embedded. By contrast, it is unclear how the impossible worlds semantics can be extended to such expressions, given that it simply appeals to impossible worlds. Somewhere, the role of logic has to be made explicit. Hyperlogic is able to achieve its success in large part because it makes the logic governing the connectives play a central role in the semantics.

References

[1] Carlos Areces and Balder ten Cate. Hybrid Logics. In Patrick Blackburn, Frank Wolter, and Johan van Benthem, editors, *Handbook of Modal Logic*, pages 821–868. Elsevier, 2006.

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

[3] Daniel H Cohen. On What Cannot Be. In *Truth or Consequences*, pages 123–132. Springer Netherlands, Dordrecht, 1990.

[4] Kit Fine. Propositional quantifiers in modal logic. *Theoria*, 36(3):336–346, 1970.

[5] Allan Gibbard. *Thinking How to Live*. Harvard University Press, Cambridge, MA, June 2003.

[6] Alexander W. Kocurek and Ethan Jerzak. Counterlogicals as counterconventionals. Manuscript, 2019.

[7] Alexander W. Kocurek, Ethan Jerzak, and Rachel Rudolph. Against conventional wisdom. Manuscript, 2019.

[8] David K Lewis. *Counterfactuals*. Blackwell Publishing, 1973.

[9] Daniel Nolan. Impossible Worlds: A Modest Approach. *Notre Dame Journal of Formal Logic*, 38(4):535–572, 1997.

[10] Robert C Stalnaker. A theory of conditionals. In Nicholas Rescher, editor, *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*, pages 98–112. Basil Blackwell Publishers, Oxford, 1968.