Modal Logic and Counterfactuals in First-Order Structures

August 7, 2019

Figure out definition/theorem/etc numbering.

A lot of math mode stuff wraps in ways that aren't great.

Maybe use display mode in more places?

Thought to self: failure of initiality when the antecedent is a disjunction can occur if there are two irreconcilable choices of model. This seems related to the Lewis vs Stalnaker "multiple closest worlds" thing! Is there a concept of "joint initiality"?

Need more of a natural progression to the paper—why look at each topic? why move on to the next one?

1 Introduction

Revise this section to better reflect the ideas of the rest of the paper.

In natural language, *counterfactuals* are "what if?" questions: conditional statements with potentially false antecedents which nevertheless have nontrivial semantic content even in cases where the standard mathematical account of implication would render them vacuously true. A number of common constructions, particularly in algebra, are reminiscent of counterfactuals. For example, the field $\mathbb{C}$ is intuitively the answer to the counterfactual "what if $\mathbb{R}$ had a root for the polynomial $q(i) = i^2 + 1$?", and a claim that a formula ϕ holds in $\mathbb{C}$ can be intuitively viewed as a claim that the formula $\exists i(i^2 = -1) > \phi$ holds in $\mathbb{R}$, where $>$ is the *counterfactual conditional* connective. Similarly, the integers mod p are like the answer to the question "what if $p = 0$ in $\mathbb{Z}$?", and a claim ϕ in $\mathbb{Z}/p\mathbb{Z}$ can be seen as a claim $p = 0 > \phi$ in $\mathbb{Z}$.

This concept can be clarified by a few observations. First, counterfactual conditionals may be modeled in a system of possible worlds by introducing an appropriate notion of "nearness" between worlds, so that the consequent of a conditional can be interpreted in the closest world[s] satisfying the antecedent, if any exist. By regarding structures as possible worlds, we can obtain the above

Maybe cite something talking about this?

correspondences as natural consequences of a notion of nearness or "minimal difference" that correctly picks out $\mathbb{C}$ as the nearest world to $\mathbb{R}$ with a root for q and $\mathbb{Z}/p\mathbb{Z}$ as the nearest world to $\mathbb{Z}$ with $p = 0$. Second, the shared characteristic between these cases that identifies $\mathbb{C}$ or $\mathbb{Z}/p\mathbb{Z}$ as being "nearest" is the possession of a relevant universal property. In the case of $\mathbb{C}$, it is that field homomorphisms $\mathbb{R} \to K$ where K has a root for q factor uniquely through the inclusion $\mathbb{R} \subseteq \mathbb{C}$ once $\mathbb{C}$ and K are pointed with a particular root of q. In the case of $\mathbb{Z}/p\mathbb{Z}$, it is that ring homomorphisms $\mathbb{Z} \to R$ where $p = 0$ in R factor uniquely through the quotient $\mathbb{Z} \twoheadrightarrow \mathbb{Z}/p\mathbb{Z}$.

The universal property of $\mathbb{C}$ hinges on there being a canonical choices of root of q in the relevant fields; without this, it fails because of the existence of nontrivial $\mathbb{R}$-fixing automorphisms of $\mathbb{C}$ (e.g., conjugation). A definition of $>$ that works in this case would require a method of deriving the appropriate extra conditions from the form $\exists i(i^2 = -1)$ of the antecedent. In this case, the connection seems fairly obvious—the necessary extra equipment is a choice of witness for the existential—but it becomes rather subtle when trying to generalize. For that reason, this paper will particularly investigate one significant class of well-behaved antecedents: universal Horn formulas.

2 Preliminaries

Definition 2.1. A *signature* is a tuple $\sigma = (\mathrm{Func}, \mathrm{Rel}, \mathrm{Const}, \mathrm{ar})$, where Func, Rel, and Const are disjoint sets of *function symbols*, *relation symbols*, and *constant symbols*, respectively, and $\mathrm{ar} : \mathrm{Func} \cup \mathrm{Rel} \to \mathbb{N}^+$ is an assignment of a positive *arity* to each function and relation symbol. The variables σ and τ will generally range over signatures.

Definition 2.2. For a signature $\sigma = (\mathrm{Func}, \mathrm{Rel}, \mathrm{Const}, \mathrm{ar})$, a *$\sigma$-structure* or *structure over σ* is a tuple $\mathbf{A} = (A, I)$, where A is a non-empty set called the *universe* or *domain* of $\mathbf{A}$, and I is an *interpretation* of σ in A; i.e., a function $I(f) : A^{\mathrm{ar}(f)} \to A$ for each $f \in \mathrm{Func}$, a relation $I(R) \subseteq A^{\mathrm{ar}(r)}$ for each $R \in \mathrm{Rel}$, and an element $I(c) \in A$ for each $c \in \mathrm{Const}$. We write $f^{\mathbf{A}}$ for $I(f)$, $R^{\mathbf{A}}$ for $I(R)$, and $c^{\mathbf{A}}$ for $I(c)$. The variables $\mathbf{A}$ and $\mathbf{B}$ will generally range over structures.

Definition 2.3. For a signature σ and σ-structures $\mathbf{A}, \mathbf{B}$, a function $\rho : A \to B$ is *homomorphism* $\mathbf{A} \to \mathbf{B}$ if

1. $\rho(f^{\mathbf{A}}(x_1, \ldots, x_n)) = f^{\mathbf{B}}(\rho(x_1), \ldots, \rho(x_n))$ for each n-ary function symbol f of σ and all $x_1, \ldots, x_n \in A$;
2. $R^{\mathbf{A}}(x_1, \ldots, x_n) \implies R^{\mathbf{B}}(\rho(x_1), \ldots, \rho(x_n))$ for each n-ary relation symbol R of σ and all $x_1, \ldots, x_n \in A$;
3. $\rho(c^{\mathbf{A}}) = c^{\mathbf{B}}$ for all constant symbols c of σ.

It is a *strong* homomorphism if the implication in 2 is replaced with an equivalence. An *embedding* is an injective strong homomorphism. The *full category*

of σ-structures, denoted by $\mathrm{Struct}(\sigma)$, is the category whose objects are the σ-structures, whose morphisms $\mathbf{A} \to \mathbf{B}$ are the homomorphisms $\mathbf{A} \to \mathbf{B}$, and whose composition is just function composition. This easily does form a category.

Definition 2.4. For a signature σ, the set of *terms over* σ is inductively generated by a countably infinite set of variables $\mathbb{V}$, the constant symbols of σ, and n-ary application of the n-ary function symbols of σ. The variable t will generally range over terms. The set of *atomic formulas over* σ is inductively generated by equalities of terms and n-ary applications of n-ary relation symbols to terms. The set of *first-order formulas over* σ, denoted $\mathbb{FOL}(\sigma)$, is inductively generated by the atomic formulas, negation ($\neg$), conjunction ($\wedge$), and universal quantification ($\forall$). Disjunction ($\vee$), implication ($\to$), true ($\top$), false ($\bot$), and existential quantification ($\exists$) can be defined as shorthand in the usual ways. The variables ϕ, ψ, and χ will generally range over formulas in $\mathbb{FOL}(\sigma)$ or extensions of it. A term or formula with no free variables is *closed*. A *sentence* is a closed formula.

Should we pick some other logically complete set of connectives? Does it matter?

Definition 2.5. A formula is a *literal* if it is either an atomic formula or a negated atomic formula. A formula is in *prenex normal form* if it consists of some quantifiers, the *prenex*, followed by a quantifier-free portion, the *matrix*. A quantifier-free formula is in *disjunctive normal form* if it is a disjunction of conjunctions of literals. A prenex-normal formula is *existential* if all of its quantifiers are existential, and it is *universal* if all of its quantifiers are universal. A prenex-normal formula with a disjunctive normal matrix is *positive* if all of the literals are non-negated.

Definition 2.6. For a structure $\mathbf{A}$, a *variable assignment into* $\mathbf{A}$, or just an *assignment into* $\mathbf{A}$ is a function $\mathbb{V} \to A$. The variable π will generally range over variable assignments. Given a variable assignment π into a structure $\mathbf{A}$, we extend π to arbitrary terms by setting $\pi(c) = c^{\mathbf{A}}$ and $\pi(f(t_1, \ldots, t_n)) = f^{\mathbf{A}}(\pi(t_1), \ldots, \pi(t_n))$. If a term t is closed, then $\pi(t) = \pi'(t)$ for any assignments π, π'; in this event, write $t^{\mathbf{A}}$ for the value of t under any assignment. The notation $\pi\{x := a\}$ denotes the assignment which sends x to a and otherwise agrees with π.

Definition 2.7. The trinary *satisfaction relation* $(\mathbf{A}, \pi) \models \phi$ between structures, variable assignments into them, and formulas, is defined inductively on formulas by

$$
\begin{aligned}
(\mathbf{A}, \pi) &\models t_1 = t_2 && \iff \pi(t_1) = \pi(t_2) \\
(\mathbf{A}, \pi) &\models R(t_1, \ldots, t_n) && \iff R^{\mathbf{A}}(t_1, \ldots, t_n) \\
(\mathbf{A}, \pi) &\models \neg\phi && \iff (\mathbf{A}, \pi) \not\models \phi \\
(\mathbf{A}, \pi) &\models \phi \wedge \psi && \iff (\mathbf{A}, \pi) \models \phi \text{ and } (\mathbf{A}, \pi) \models \psi \\
(\mathbf{A}, \pi) &\models \forall x \phi && \iff \text{for all } a \in A, \text{ we have } (\mathbf{A}, \pi\{x := a\}) \models \phi.
\end{aligned}
$$

Write $\mathbf{A} \models \phi$ to mean that $(\mathbf{A}, \pi) \models \phi$ for all π, and write $\models \phi$ to mean that $\mathbf{A} \models \phi$ for all $\mathbf{A}$. Two formulas ϕ, ψ are *logically equivalent*, written $\phi \equiv \psi$, if $\models \phi \leftrightarrow \psi$.

Definition 2.8. A *theory* over a signature σ is a set of sentences over σ. The variable T will generally range over theories. A structure $\mathbf{A}$ is a *model* of a theory T if it satisfies every sentence in T. The *full category of models of* T, denoted by $\mathrm{Mod}(T)$, is the full subcategory of $\mathrm{Struct}(\sigma)$ whose objects are the models of T.

3 Modal First-Order Logic

In order to apply the conceptual framework of possible worlds to categories of structures, we first fix a broad notion of what kind of "categories of structures" we will be considering.

Definition 3.1. Given a signature σ, a *category of σ-structures* is a subcategory of $\mathrm{Struct}(\sigma)$. The variable $\mathcal{C}$ will generally range over categories of structures, and the variables f, g, and h will generally range over their morphisms. When some structures $\mathbf{A}$ and $\mathbf{B}$ are being considered as objects of some category of structures $\mathcal{C}$, the notation $f : \mathbf{A} \to \mathbf{B}$ will mean that f, beyond being a homomorphism, specifically belongs to $\mathcal{C}$.

Let $\mathcal{C}$ be a category of structures. If we view the objects of $\mathcal{C}$ as possible worlds, then we can understand a morphism $f : \mathbf{A} \to \mathbf{B}$ as a way of picking out a *counterpart* in $\mathbf{B}$ for each individual of $\mathbf{A}$, or as an exhibition of $\mathbf{B}$ as "analogous" in some way to $\mathbf{A}$. The fact that there may be other, distinct morphisms $\mathbf{A} \to \mathbf{B}$ reflects the fact that there may be other acceptable but "inconsistent-with-f" ways of interpreting $\mathbf{B}$ as a hypothetical variation on $\mathbf{A}$. This is a restricted instance of a more general family of notions which are described in Kracht and Kutz [3, §7]. Of the types of structures discussed there, the approach we will take is most similar in intent to the modal structures in Kracht and Kutz [4], although it is closer in consequences to the presheaf models in Ghilardi [1].

We analyze systems of possible worlds by interpreting the language of modal logic into them; later, we will also consider strict conditional and counterfactual conditional connectives. For now, we consider the necessitation and possibility modalities, $\Box$ and $\Diamond$.

Definition 3.2. The set of *modal first-order formulas over* σ, denoted $\mathbb{MFOL}(\sigma)$, is inductively generated by the first-order connectives and a new unary modal operator $\Box$. We define $\Diamond\phi$ to be shorthand for $\neg\Box\neg\phi$. Both of these connectives have the same precedence as negation.

In standard Kripke semantics, these connectives quantify over possible worlds, but here this is insufficient. If a statement makes reference to individuals, then before we can ask whether it holds in another possible world, we must know how to translate it into a statement about some corresponding individuals in

that world. Thus, rather than quantifying over *reachable worlds*, we quantify over *ways of reaching those worlds*; i.e., morphisms.

Definition 3.3. The 4-ary *modal satisfaction relation* $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi$, where $\mathcal{C}$ is a category of σ-structures, $\mathbf{A}$ is an object of $\mathcal{C}$, π is an assignment into $\mathbf{A}$, and ϕ is a formula of $\mathbb{MFOL}(\sigma)$, is defined by adding $\mathcal{C}$ as a parameter to the prior inductive definition of $\models$, with the existing clauses leaving it untouched, and then adding the rule

$$(\mathbf{A}, \pi) \models_{\mathcal{C}} \Box\phi \iff \text{for all } \mathbf{B} \in \mathcal{C} \text{ and } f : \mathbf{A} \to \mathbf{B}, \text{ we have } (\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi.$$

Write $\mathbf{A} \models_{\mathcal{C}} \phi$ to mean that $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi$ for all π, write $\models_{\mathcal{C}} \phi$ to mean that $\mathbf{A} \models_{\mathcal{C}} \phi$ for all $\mathbf{A} \in \mathcal{C}$, and write $\models \phi$ to mean that $\models_{\mathcal{C}} \phi$ for all $\mathcal{C}$. We will say that ϕ and ψ are logically equivalent *in* $\mathcal{C}$, written $\phi \equiv_{\mathcal{C}} \psi$, if $\models_{\mathcal{C}} \phi \leftrightarrow \psi$.

Many common validities of modal logic hold in these semantics. Notably, axioms **T** and **4** follow from the existence of compositions and identities, analogously to how they follow from transitivity and reflexivity in propositional Kripke frames.[1]

Should this even be a footnote?

Theorem 1. *The basic rules of modus ponens and generalization are admissible. I.e., for all $\phi, \psi \in \mathbb{MFOL}(\sigma)$, all $\mathcal{C}$, and all $\mathbf{A}$:*

***MP**: If $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi \to \psi$ and $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi$, then $(\mathbf{A}, \pi) \models_{\mathcal{C}} \psi$, and this also holds when "$(\mathbf{A}, \pi) \models_{\mathcal{C}}$" is replaced by any of $\mathbf{A} \models_{\mathcal{C}}$, $\models_{\mathcal{C}}$, or $\models$.*

***Gen**: If $\mathbf{A} \models_{\mathcal{C}} \phi$, then $\mathbf{A} \models_{\mathcal{C}} \forall x\phi$, and this also holds when "$\mathbf{A} \models_{\mathcal{C}}$" is replaced by $\models_{\mathcal{C}}$ or $\models$.*

The rules of the modal logic S4 are additionally admissible. I.e., for all $\phi, \psi \in \mathbb{MFOL}(\sigma)$:

***N**: For all $\mathcal{C}$, if $\models_{\mathcal{C}} \phi$, then $\models_{\mathcal{C}} \Box\phi$; if $\models \phi$, then $\models \Box\phi$.*

***K**: $\models \Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$*

***T**: $\models \Box\phi \to \phi$*

***4**: $\models \Box\phi \to \Box\Box\phi$*

Proof. **MP** and **Gen** hold by the same arguments as in ordinary first-order logic. Each of the modal rules reduces to a complicated but easy-to-prove statement once the definition of $\models$ is expanded. To make things more readable, I will elide some of the quantification.

[1] This analogy does not hold up perfectly—in propositional Kripke frames, one considers frames prior to the association of specific truth assignments to them, and a converse holds where **T** and **4** *imply* transitivity and reflexivity when they hold in all possible truth assignments. In this paper, we require the objects of our category to be equipped with interpretations from the start in order to be able to demand that morphisms are homomorphisms; and even if we generalized to allow non-homomorphisms and the absence of composition and identities (this is essentially the semantics of [4]), **T** and **4** would only imply something significantly weaker than being a category—see [4] for details.

N: The expanded statement of the first claim is: If $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi$ for all $(\mathbf{A}, \pi)$, then for all $\mathbf{A}', \pi', f : \mathbf{A}' \to \mathbf{B}$, we have $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$. This follows by instantiating the premise with $(\mathbf{B}, f \circ \pi)$. The second claim follows quickly from the first.

K: The expanded statement is: If for all $f : \mathbf{A} \to \mathbf{B}$ with $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$, we have $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \psi$; and for all $f' : \mathbf{A} \to \mathbf{B}'$ we have $(\mathbf{B}', f' \circ \pi) \models_{\mathcal{C}} \phi$; then for all $f'' : \mathbf{A} \to \mathbf{B}''$, we have $(\mathbf{B}'', f'' \circ \pi) \models_{\mathcal{C}} \psi$. This follows by instantiating the premises with f''.

T: The expanded statement is: If $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$ for all $f : \mathbf{A} \to \mathbf{B}$, then we have $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi$. This follows by instantiating the premise with f as the identity at $\mathbf{A}$.

4: The expanded statement is: If $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$ for all $f : \mathbf{A} \to \mathbf{B}$, then for all $f' : \mathbf{A} \to \mathbf{B}'$ and $f'' : \mathbf{B}' \to \mathbf{B}''$ we have $(\mathbf{B}'', f'' \circ f' \circ \pi) \models_{\mathcal{C}} \phi$. This follows by instantiating the premise with f as $f'' \circ f'$.

□

Theorem 2. *For any category of structures $\mathcal{C}$, logical equivalence in $\mathcal{C}$ is a congruence; i.e., if $\phi \equiv_{\mathcal{C}} \psi$, and ϕ occurs as a subformula of χ, then replacing this subformula with ψ produces a result logically equivalent to χ in $\mathcal{C}$. Since this holds for all $\mathcal{C}$, it is also true for full logical equivalence.*

Do we want a version of this for $\models_{\mathcal{C}} t = s$ too?

Proof. This follows by induction on χ, showing that each connective enclosing ϕ respects logical equivalence in $\mathcal{C}$. This holds for each of the first-order connectives by standard arguments. For $\Box$, use **N** and **K** to find that

$$\models_{\mathcal{C}} \phi \to \psi \implies \models_{\mathcal{C}} \Box(\phi \to \psi) \implies \models_{\mathcal{C}} \Box\phi \to \Box\psi,$$

and similarly for $\psi \to \phi$. Therefore, $\phi \equiv_{\mathcal{C}} \psi \implies \Box\phi \equiv_{\mathcal{C}} \Box\psi$. □

Corollary 1. *For all $\phi, \psi \in \mathbb{MFOL}(\sigma)$:*

- $\Box(\phi \wedge \psi) \equiv \Box\phi \wedge \Box\psi$
- $\Diamond(\phi \vee \psi) \equiv \Diamond\phi \vee \Diamond\psi$
- $\models \Box(\phi \leftrightarrow \psi) \to (\Box\phi \leftrightarrow \Box\psi)$

Proof.

- From left to right: we have $\models \phi \wedge \psi \to \phi$, so by **N** and **K**, we have $\models \Box(\phi \wedge \psi) \to \Box\phi$. By similar reasoning we have $\models \Box(\phi \wedge \psi) \to \Box\psi$. Combining these gives $\models \Box(\phi \wedge \psi) \to \Box\phi \wedge \Box\psi$. From right to left: we have $\models \phi \to \psi \to \phi \wedge \psi$, so by **N** and **K**, we have $\models \Box\phi \to \Box\psi \to \Box(\phi \wedge \psi)$, and then $\models \Box\phi \wedge \Box\psi \to \Box(\phi \wedge \psi)$.
- From the previous equivalence, we have $\Box(\neg\phi \wedge \neg\psi) \equiv \Box\neg\phi \wedge \Box\neg\psi$. Negating both sides and distributing the negation gives $\Diamond(\neg\neg\phi \vee \neg\neg\psi) \equiv \Diamond\neg\neg\phi \vee \Diamond\neg\neg\psi$. Cancelling the negations gives the result.

- Since $\leftrightarrow$ is just a conjunction of implications, by the first equivalence we have $\Box(\phi \leftrightarrow \psi) \equiv \Box(\phi \to \psi) \wedge \Box(\psi \to \phi)$. Then applying **K** to both conjuncts gives the result.

□

If a formula ϕ is preserved by a morphism $f : \mathbf{A} \to \mathbf{B}$, we can think of this as meaning that the truth of ϕ remains unchanged when considered in another possible world by means of f, or that the analogy represented by f does not distort the significance of ϕ's truth. Thus, if ϕ is preserved by *all* morphisms of a category $\mathcal{C}$, we can think of this as meaning that, in the event that ϕ is known, it can be safely assumed to hold in any hypothetical as well—i.e., $\models_{\mathcal{C}} \phi \to \Box\phi$. Then the basic preservation results give rise to modal validities.

Say something about what it means to be a homomorphism in this light.

Theorem 3. *For $\phi \in \mathbb{FOL}(\sigma)$ and $\mathcal{C}$ a category of σ-structures, we have $\models_{\mathcal{C}} \phi \to \Box\phi$ under any of the following conditions:*

- *ϕ is logically equivalent to an existential formula and every morphism of $\mathcal{C}$ is an embedding.*
- *ϕ is logically equivalent to a positive formula and every morphism of $\mathcal{C}$ is surjective.*
- *ϕ is logically equivalent to an existential positive formula.*

This is kind of an awkward shift between intuitive handwaving and unjustified formal claims.

In particular. we have $\models \phi \to \Box\phi$ for existential positive ϕ. This includes all atomic formulas.

Proof. For specificity, define "$f : \mathbf{A} \to \mathbf{B}$ preserves ϕ" to mean that for all assignments π into $\mathbf{A}$, if $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi$, then $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$. Then, by expanding definitions and shuffling around some quantifiers, we can rephrase $\models_{\mathcal{C}} \phi \to \Box\phi$ as "for all $\mathbf{A} \in \mathcal{C}$ and $f : \mathbf{A} \to \mathbf{B}$, f preserves ϕ". Then the theorem follows from the following standard results, which can be found in, e.g., [2, §2.4]:

kinda awkward, nonspecific citation

Fact 1.

- *Any formula logically equivalent to an existential formula is preserved by embeddings.*
- *Any formula logically equivalent to a positive formula is preserved by surjections.*
- *Any formula logically equivalent to an existential positive formula is preserved by all homomorphisms.*

□

Maybe also bring up the Barcan and converse Barcan schemata?

4 Conditionals

Definition 4.1. The *strict conditional* $\phi \dashv\!\!\exists\, \psi$ is defined as notation for $\Box(\phi \to \psi)$. It will be right-associative in our notation and will have the same precedence as $\to$. If we expand the semantics for $\Box$ and $\to$, we see that

$$(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi \dashv\!\!\exists\, \psi \iff \text{for all } \mathbf{B} \in \mathcal{C} \text{ and } f : \mathbf{A} \to \mathbf{B} \text{ such that } (\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi, \text{ we have } (\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \psi.$$

This conditional expresses necessary, as opposed to contingent, consequence. To say $\phi \dashv\!\!\exists\, \psi$ is to say that any hypothetical which affirms ϕ must also affirm ψ, or, contrapositively, that to reject ψ for the purpose of a hypothetical forces one to also lose ϕ.

The basic properties of the strict conditional drop out of the basic properties of $\Box$.

Theorem 4. *The following rules are admissible for all* $\phi, \psi, \chi \in \mathbb{MFOL}(\sigma)$*:*

- *For all* $\mathcal{C}$*, if* $\models_{\mathcal{C}} \phi \to \psi$*, then* $\models_{\mathcal{C}} \phi \dashv\!\!\exists\, \psi$*; if* $\models \phi \to \psi$*, then* $\models \phi \dashv\!\!\exists\, \psi$*.*
- $\models \Box\psi \dashv\!\!\exists\, \phi \dashv\!\!\exists\, \psi$
- $\models \neg\Diamond\phi \dashv\!\!\exists\, \phi \dashv\!\!\exists\, \psi$
- $(\phi \dashv\!\!\exists\, \psi) \equiv (\neg\psi \dashv\!\!\exists\, \neg\phi)$
- $\models (\phi \dashv\!\!\exists\, \psi) \to (\psi \dashv\!\!\exists\, \chi) \to (\phi \dashv\!\!\exists\, \chi)$

Figure out which rules and non-rules are actually important and then write them down!!!

However, we have the following non-rules (for some *choices of* ϕ, ψ, χ*, in each case):*

- $\not\models \psi \dashv\!\!\exists\, \phi \dashv\!\!\exists\, \psi$*.*
- $\not\models \neg\Diamond\phi \dashv\!\!\exists\, \phi \dashv\!\!\exists\, \phi$*.*

Write a proof.

The strict conditional can be used to express claims like "if -1 were to have at least one square root, then it would have to have exactly two square roots": if F is the theory of fields, then it is true and non-vacuous that

$$\mathbb{R} \models_{\mathrm{Mod}(F)} \exists i(i^2 = -1) \dashv\!\!\exists\, \exists_{=2} i(i^2 = -1),$$

where $\exists_{=n}$ is the usual first-order-definable "exists exactly n" quantifier. But for some purposes, the strict conditional is too strong. In particular, it has been critiqued as an inadequate interpretation of natural language counterfactuals on the basis that such conditionals are *non-monotonic*; that is, they can fail if their antecedent is strengthened. One major approach to explaining their meaning is given by Stalnaker [6] as follows:

Elaborate and/or cite something!!

> Consider a possible world in which A is true, and which otherwise differs minimally from the actual world. "If A, then B" is true (false) just in case B is true (false) in that possible world.

This view can be naturally applied to statements within the framework set up thus far. Consider the claim of the integers that "if 2 were equal to 0, then every number would have to be equal to either 0 or 1". If we interpret this as a strict conditional—i.e., as $2 = 0 \dashrightarrow \forall n(n = 0 \vee n = 1)$—then the claim is false: letting R be the theory of rings with 1,

$$\mathbb{Z} \not\models_{\mathrm{Mod}(R)} 2 = 0 \dashrightarrow \forall n(n = 0 \vee n = 1).$$

This is because there are ring homomorphisms out of $\mathbb{Z}$ whose codomains satisfy the antecedent but not the consequent—for example, there is a homomorphism to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ (the quotient followed by the diagonal), but

$$\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \not\models_{\mathrm{Mod}(R)} 2 = 0 \to \forall n(n = 0 \vee n = 1).$$

If we think of the informal claim as a counterfactual, however, we shouldn't care about $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, because it differs from $\mathbb{Z}$ in ways other than the one we are interested in. What we should really be interested in is $\mathbb{Z}/2\mathbb{Z}$, the obvious contender for "a ring in which $2 = 0$, and which otherwise differs minimally from $\mathbb{Z}$"; and in $\mathbb{Z}/2\mathbb{Z}$, the consequent is of course true, rendering the overall conditional true when thought of as a counterfactual.

The reason why $\mathbb{Z}/2\mathbb{Z}$ is the obvious contender for this position is that it is the quotient of $\mathbb{Z}$ by the ideal generated by 2, quotients are fairly intuitively the meaning of "identifying together some elements", and "generated by" is a kind of minimality. But the reason why quotients are the correct notion of "identification of elements" is, from the perspective of category theory, because they satisfy a universal property: in this case, any homomorphism from $\mathbb{Z}$ to a ring of characteristic 2 factors uniquely through the quotient map to $\mathbb{Z}/2\mathbb{Z}$; or in other words, the quotient map is initial in the category of rings under $\mathbb{Z}$ of characteristic 2.

Connect "minimal difference" to "freeness" to "initiality".

As another example, suppose a counterfactual in $\mathbb{R}$ (considered as a field) has antecedent $\exists i(i^2 + 1 = 0)$. It is intuitive that the field in which this is true, and which otherwise differs minimally from $\mathbb{R}$—i.e., the field in which the consequent of the counterfactual should be considered—is $\mathbb{C}$. This is because $\mathbb{C}$ is the field extension of $\mathbb{R}$ given by the polynomial $p(i) = i^2 + 1 = 0$. Once again, this can be justified as the correct notion of "adding a square root of -1" because there is a universal property: given a field homomorphism $f : \mathbb{R} \to F$ such that F has a root for p, and in particular a fixed choice of root i_F, there is a unique extension $\tilde{f} : \mathbb{C} \to F$ of f such that $\tilde{f}(i) = i_F$. In other words, the inclusion $\mathbb{R} \subseteq \mathbb{C}$ is initial in the category of fields under $\mathbb{R}$ equipped with a root of p, once $\mathbb{C}$ is equipped with i.

Finally, suppose a counterfactual in $\mathbb{N}$, considered as a poset, has antecedent $\forall n \exists n'(n' < n)$. While not as obvious as the prior two cases, it is at least

plausible that the poset in which the consequent should be considered is $\mathbb{Z}$. For a third time, this can be justified—albeit a bit more tenuously—by suggesting a universal property: for any monotone function $f : \mathbb{N} \to P$, and choice of function $p : P \to P$ such that $p(x) < x$ for all x, there is a unique monotone extension $\tilde{f} : \mathbb{Z} \to P$ such that $\tilde{f}(n-1) = p(f(n))$. In other words, the inclusion $\mathbb{N} \subseteq \mathbb{Z}$ is initial in the category of posets under $\mathbb{N}$ equipped with a "choice of smaller element" function, once $\mathbb{Z}$ is equipped with the predecessor operation.

These examples suggest that, to give semantics to counterfactuals in a vein similar to Stalnaker [6], a reasonable approach to finding a "minimally different" world in which to evaluate the consequent is to look for a morphism out of the current world which is initial in some appropriate category of structures under the current world. Unfortunately, the latter two examples require a choice of extra equipment for these structures. In these cases, the type of equipment offered has a clear resemblance to the form of the antecedent, but finding a well-behaved general criterion is subtle; for example, the antecedent of the $\mathbb{N}$ example is logically equivalent to $\neg\exists l \forall n (l \leq n)$, which is no longer obviously identifiable with the structure of a "choice of smaller element" function. One appealing approach would would be to Skolemize any antecedent of a counterfactual and use the newly-introduced function and constant symbols as the needed extra structure, but there could perfectly well be more than one Skolem normal form, and the truth of the counterfactual might not be the same between the options. For this paper, we resign ourself to a semantics for counterfactuals which considers initiality only in an under category with no extra equipment on the structures. This cannot account for the field and poset examples, but it does give rise to the ring example, and it has at least one significant class of very well-behaved cases which is investigated in §5.

Definition 4.2. The set of *counterfactual first-order formulas over* σ, denoted $\mathbb{CFOL}(\sigma)$, is inductively generated by the first-order connectives, the unary modal operator $\Box$, and the *counterfactual conditional* connective $>$. Like $\boxright$, this connective will be right-associative in our notation and will have the same precedence as $\to$.

Which results from above might fail if the formulas in them are allowed to include $>$? I don't think any, at the moment. Probably worth mentioning in the text.

The semantics for $>$ will be a refinement of those for $\boxright$. As with $(\mathbf{A}, \pi) \models_{\mathcal{C}} \phi \boxright \psi$, we will want to consider morphisms $f : \mathbf{A} \to \mathbf{B}$ such that $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$, but instead of evaluating ψ in all cases, we check it only in the f (if any) which is *initial*, as discussed above. We first define precisely which category we mean for f to be initial in, making use of the soon-to-be-defined satisfaction relation for $\mathbb{CFOL}(\sigma)$.[2]

Definition 4.3. Let $\mathcal{C}$ be a category of σ-structures, $\mathbf{A}$ an object of $\mathcal{C}$, π a variable assignment $\mathbb{V} \to \mathbf{A}$, and ϕ a formula of $\mathbb{CFOL}(\sigma)$. Let $\mathbf{A} \downarrow \mathcal{C}$ denote the under category of morphisms out of $\mathbf{A}$. Then define $(\mathbf{A}, \pi) \downarrow_{\phi} \mathcal{C}$ to be the full subcategory of $\mathbf{A} \downarrow \mathcal{C}$ whose objects $(\mathbf{B} \in \mathcal{C}, f : \mathbf{A} \to \mathbf{B})$ are those that satisfy $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$.

[2] Even though the satisfaction relation refers back to this definition, there is not a true circularity, just a diversion in the self-reference used by the inductive definition of $\models$.

Note that

Fact 2.

$$(\mathbf{A},\pi) \models_{\mathcal{C}} \phi \dashv\!\!\models \psi \iff \textit{for all objects } (\mathbf{B},f) \textit{ of } (\mathbf{A},\pi)\downarrow_{\phi} \mathcal{C}, \textit{ we have } (\mathbf{B}, f\circ\pi) \models_{\mathcal{C}} \psi.$$

We can now give semantics to $>$.

Definition 4.4. We extend the modal satisfaction relation to $\mathbb{CFOL}(\sigma)$ by extending its definition with

$$(\mathbf{A},\pi) \models_{\mathcal{C}} \phi > \psi \iff \text{for all } \textit{initial} \text{ objects } (\mathbf{B},f) \text{ of } (\mathbf{A},\pi)\downarrow_{\phi} \mathcal{C}, \text{ we have } (\mathbf{B}, f\circ\pi) \models_{\mathcal{C}} \psi.$$

Since any two initial objects are isomorphic, this is equivalent to

$$(\mathbf{A},\pi) \models_{\mathcal{C}} \phi > \psi \iff \text{either } (\mathbf{A},\pi)\downarrow_{\phi} \mathcal{C} \text{ has no initial objects, or there is an initial } (\mathbf{B},f) \text{ with } (\mathbf{B}, f\circ\pi) \models_{\mathcal{C}} \psi.$$

This can be checked to validate the above example in rings:

$$\mathbb{Z} \models_{\mathrm{Mod}(R)} 2 = 0 > \forall n(n = 0 \vee n = 1).$$

As mentioned, however, it fails to pick out $\mathbb{C}$ in the example in fields: $\mathbb{C}$ is not initial in $(\mathbb{R},\pi)\downarrow_{\phi} \mathcal{C}$, because it has nontrivial automorphisms fixing $\mathbb{R}$.

Theorem 5. *The strict conditional is weaker than the counterfactual conditional; i.e., for all $\phi,\psi \in \mathbb{CFOL}(\sigma)$,*

$$\models (\phi \dashv\!\!\models \psi) \to (\phi > \psi).$$

The converse does not hold in general: there exist ϕ,ψ such that

$$\not\models (\phi > \psi) \to (\phi \dashv\!\!\models \psi).$$

Proof. The first claim follows immediately from Fact 2 and the definition of $>$. The second claim follows from the counterexample in $\mathbb{Z}$. □

When reasoning counterfactually, we get some tools unsound for the strict conditional that arise from the fact that the consequent of a counterfactual is evaluated in at most one world (up to isomorphism).

Theorem 6. *The counterfactual conditional satisfies a kind of law of excluded middle: for all ϕ,ψ,*

$$\models (\phi > \psi) \vee (\phi > \neg\psi).$$

This immediately implies that

$$\models (\phi \not> \psi) \to (\phi > \neg\psi).$$

Proof. Consider any particular $\mathcal{C}$, $\mathbf{A}$, π; we show that $(\mathbf{A},\pi) \models_{\mathcal{C}} (\phi > \psi) \vee (\phi > \neg\psi)$. If $(\mathbf{A},\pi)\downarrow_{\phi} \mathcal{C}$ has no initial objects, then this holds vacuously. If it does have an initial object $(\mathbf{B},f)$, then either $(\mathbf{B}, f\circ\pi) \models_{\mathcal{C}} \psi$ or $(\mathbf{B}, f\circ\pi) \models_{\mathcal{C}} \neg\psi$; in either case, we have our goal. □

5 Universal Horn Formulas

Probably'd be a good idea to talk at some point about prior art on the application of Horn formulas to counterfactual-y stuff.

We consider the case where $\mathcal{C}$ is the full category of models of some universal Horn theory and the antecedent of $>$ is logically equivalent to a universal Horn formula; this case will turn out to have very good behavior.

Definition 5.1. A *basic Horn formula* is one of the form $\phi_1 \wedge \cdots \wedge \phi_n \to \psi$, where each ϕ_n is atomic and ψ is either atomic or $\bot$. A *Horn formula* is a prenex-normal formula whose matrix is a conjunction of basic Horn formulas. A *universal Horn theory* is a theory whose axioms are all logically equivalent to universal Horn formulas. We will frequently abuse the distinction between Horn formulas and formulas logically equivalent to them.

Universal Horn theories include, for example, the theories of groups, unital rings, preorders, and partial orders; but not, for example, the theories of fields (an existential quantifier is needed for existence of multiplicative inverses) or total orders (a disjunction is necessary to express totality).

In this setting, we will always have the initial models we want, as long as the antecedent is possible.

Theorem 7. *Let T be a universal Horn theory over σ, $\mathbf{A}$ a model of T, π a variable assignment into $\mathbf{A}$, and ϕ a universal Horn formula (not necessarily a sentence) of $\mathbb{FOL}(\sigma)$. Then*

(i) $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$ *has an initial object iff it is nonempty.*

(ii) For any initial $(\mathbf{B}, f)$, *f is surjective.*

Proof. We will need a key fact about universal Horn theories.

Fact 3. *Let T' be a universal Horn theory over a signature containing at least one constant symbol. For any theory Δ of atomic sentences,* $\mathrm{Mod}(T' \cup \Delta)$ *has an initial object iff it is nonempty. Furthermore, every element of the initial object will be the value of some closed term.*

Proof. This is essentially a rephrasing of Theorem 3.8 in [5]. □

With this in hand, we will prove our result by giving a T', Δ satisfying these conditions, and such that $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$ is equivalent to $\mathrm{Mod}(T' \cup \Delta)$. Then (i) follows immediately, and (ii) is easy: if $(\mathbf{B}, f)$ is initial, and $b \in B$, then Fact 3 says that there is some closed term t with $t^{\mathbf{B}} = b$, and then since f is a homomorphism, $f(t^{\mathbf{A}}) = t^{\mathbf{B}} = b$.

Our T', Δ will be over the signature $\sigma(\mathbf{A})$ given by extending σ with fresh constant symbols c_a for $a \in A$. Set T' to be $T \cup \{\phi_\pi\}$, where ϕ_π is the sentence given by replacing each free variable x of ϕ with $c_{\pi(x)}$, and set Δ to be the positive diagram of $\mathbf{A}$—i.e., the $\sigma(\mathbf{A})$-theory comprising every atomic sentence true in $\mathbf{A}$ (but *not* the negations of those that are false). Then by the assump-

Wait that's not well-typed

tions on T and ϕ, and by the definition of the positive diagram, T', Δ satisfy the conditions of Fact 3.

We now define an inverse pair of functors $F : (\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T) \to \mathrm{Mod}(T' \cup \Delta)$ and $G : \mathrm{Mod}(T' \cup \Delta) \to (\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$.

For an object $(\mathbf{B}, f : \mathbf{A} \to \mathbf{B})$ of F's domain, define $F(\mathbf{B}, f)$ to be the $\sigma(\mathbf{A})$-structure which expands $\mathbf{B}$ by interpreting each new constant symbol c_a as $f(a)$. For a morphism $\rho : (\mathbf{B}, f) \to (\mathbf{B}, f')$ of F's domain, just set $F(f) = f$. For an object $\overline{\mathbf{B}}$ of G's domain, define $\mathbf{B}$ to be the σ-structure given by removing the interpretations of the c_as from $\overline{\mathbf{B}}$, define $f : \mathbf{A} \to \mathbf{B}$ by $f(a) = c_a^{\overline{\mathbf{B}}}$, and set $G(\overline{\mathbf{B}}) = (\mathbf{B}, f)$. For a morphism $f : \overline{\mathbf{B}} \to \overline{\mathbf{B}}'$ of G's domain, just set $G(f) = f$. These both trivially preserve identities and composition, and are easily seen to be inverses of each other, so we just need to show that the constructions given actually produce objects and morphisms of the codomains.

For F, we must show that $F(\mathbf{B}, f)$ really is a model of $T' \cup \Delta$, and that $F(f)$ will always be a homomorphism of $\sigma(\mathbf{A})$-structures. $F(\mathbf{B}, f)$ models T because $\mathbf{B}$ was drawn from $\mathrm{Mod}(T)$; it models ϕ_π because $(\mathbf{B}, f \circ \pi) \models_{\mathcal{C}} \phi$; and it models Δ because...

elaborate

Invoke diagram lemma or something?

For G, we must show that... □

Corollary 2. *For universal Horn T and ϕ, and positive ψ,*

$$\models_{\mathrm{Mod}(T)} \psi \to \phi > \psi.$$

Proof. Suppose $(\mathbf{A}, \pi) \models_{\mathrm{Mod}(T)} \psi$. If $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$ has no initial object, then the result is vacuously true. If it does have one, call it $(\mathbf{B}, f)$; we want to show that $(\mathbf{B}, f \circ \pi) \models_{\mathrm{Mod}(T)} \psi$. Then this follows because f is surjective by Theorem 7 and surjections preserve positive formulas by Fact 1. □

If initial models may not exist, then $\phi > \psi$ can be vacuously true even while $\phi \not\Box\!\!\to \neg\psi$ is non-vacuously true. The existence (when possible) of initial models precludes this; a counterfactual statement is only vacuously true when the corresponding strict conditional is.

Corollary 3. *For universal Horn T and ϕ,*

$$(\phi > \bot) \equiv_{\mathrm{Mod}(T)} (\phi \not\Box\!\!\to \bot),$$

and hence also (by negating both sides)

$$(\phi \not> \bot) \equiv_{\mathrm{Mod}(T)} \Diamond\phi.$$

Proof. Consider any $(\mathbf{A}, \pi)$. On the left, $(\mathbf{A}, \pi) \models_{\mathrm{Mod}(T)} \phi \not\Box\!\!\to \bot$ iff $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$ is empty. On the right, $(\mathbf{A}, \pi) \models_{\mathrm{Mod}(T)} \phi > \bot$ iff $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$ has no initial object. Theorem 7 says that these are equivalent. □

Counterfactuals *do* coincide with strict conditionals for a limited class of consequents, meaning that we can derive necessary consequence for certain kinds of facts by considering them only in a minimally different case.

Theorem 8. *For universal Horn T and ϕ, and ψ logically equivalent to an existential positive formula,*

$$(\phi > \psi) \equiv_{\mathrm{Mod}(T)} (\phi \mathrel{-\!\!\!\ni} \psi).$$

Proof. The right-to-left direction is just Theorem 5. For the left-to-right direction, suppose that $(\mathbf{A}, \pi) \models_{\mathrm{Mod}(T)} \phi > \psi$; we want $(\mathbf{A}, \pi) \models_{\mathrm{Mod}(T)} \phi \mathrel{-\!\!\!\ni} \psi$. If $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$ is empty, then this holds vacuously. Otherwise, it has an initial object $(\mathbf{B}, f)$, and by the assumption, $(\mathbf{B}, f \circ \pi) \models_{\mathrm{Mod}(T)} \psi$. Using Fact 2, suppose that $(\mathbf{B}', f')$ is another object; we want $(\mathbf{B}', f' \circ \pi) \models_{\mathrm{Mod}(T)} \psi$. Since $(\mathbf{B}, f)$ is initial, there is a morphism $\rho : \mathbf{B} \to \mathbf{B}'$ of $(\mathbf{A}, \pi) \downarrow_\phi \mathrm{Mod}(T)$, which by definition of the under category satisfies $f' = \rho \circ f$. Then since ψ is logically equivalent to an existential positive formula, Fact 1 tells us that ρ preserves it, so $(\mathbf{B}', \rho \circ f \circ \pi) \models_{\mathrm{Mod}(T)} \psi$, as desired. □

We can formally state one kind of "minimal difference".

Theorem 9. *For universal Horn T and ϕ, and ψ logically equivalent to an existential positive formula,*

$$\models_{\mathrm{Mod}(T)} (\phi \not\ni \psi) \to (\phi > \neg\psi).$$

Proof. Taking the contrapositive of the left-to-right direction of Theorem 8 gives $\models_{\mathrm{Mod}(T)} (\phi \not\ni \psi) \to (\phi \not> \psi)$. Then we can compose this with Theorem 6. □

Intuitively: if taking ϕ does not force us to take ψ, then ψ is false when only ϕ is taken as a counterfactual.

...

References

[1] Silvio Ghilardi. "Presheaf semantics and independence results for some non-classical first-order logics". In: *Archive for Mathematical Logic* 29 (June 1989), pp. 125–136. DOI: 10.1007/BF01620621.

[2] Wilfrid Hodges. *Model Theory*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1993. DOI: 10.1017/CBO9780511551574.

[3] Marcus Kracht and Oliver Kutz. "Logically Possible Worlds and Counterpart Semantics for Modal Logic". In: *Philosophy of Logic* (Dec. 2007). DOI: 10.1016/B978-044451541-4/50025-7.

[4] Marcus Kracht and Oliver Kutz. "The Semantics of Modal Predicate Logic I: Counterpart-Frames". In: *Advances in Modal Logic, Volume 3*. Jan. 2000, pp. 299–320. DOI: 10.1142/9789812776471_0016.

[5] Johann Makowsky. "Why Horn Formulas Matter in Computer Science: Initial Structures and Generic Examples". In: *Mathematical Foundations of Software Development*. Mar. 1985, pp. 374–387. DOI: 10.1007/3-540-15198-2_24.

[6] Robert C. Stalnaker. “A Theory of Conditionals”. In: *IFS: Conditionals, Belief, Decision, Chance and Time*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce. Dordrecht: Springer Netherlands, 1981, pp. 41–55. ISBN: 978-94-009-9117-0. DOI: 10.1007/978-94-009-9117-0_2.