VIII. Counterfactuals

AS.150.498: Modal Logic and Its Applications
Johns Hopkins University, Spring 2017

Our next application is *counterfactuals*. The possible worlds framework that we have been working with affords a powerful semantic analysis of natural language conditionals such as these:

(1) If kangaroos had no tails, they would topple over.

(2) If Oswald had not killed Kennedy, then someone else would have.

The basic idea is that a counterfactual conditional is true just in case its consequent holds at all of the *closest* worlds in which its antecedent holds.

1 Syntax

We will be working with the following language:

Definition 8.1. The **language of counterfactuals** $\mathcal{L}_{cf}$ extends the basic sentential language with two counterfactual conditional operators:

$$p \mid \bot \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid (\varphi \mathrel{\Box\!\!\rightarrow} \varphi) \mid (\varphi \mathrel{\Diamond\!\!\rightarrow} \varphi)$$

Read $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ as 'If it were the case that φ then it would be the case that ψ' and $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi$ as 'If it were the case that φ then it might be the case that ψ.'

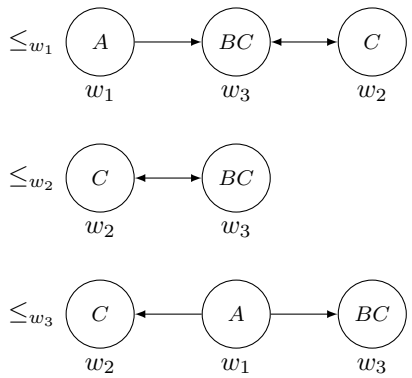
Note that this language is redundant since $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$ are interdefinable: $\varphi \mathrel{\Box\!\!\rightarrow} \psi \equiv \neg(\varphi \mathrel{\Diamond\!\!\rightarrow} \neg\psi)$ and $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi \equiv \neg(\varphi \mathrel{\Box\!\!\rightarrow} \neg\psi)$.

2 Semantics

To evaluate counterfactuals, we will use models that include a relation of comparative similarity between worlds.

Definition 8.2. A **Kripke world-ordered model** $\mathcal{M} = \langle \mathcal{W}, \{\leq_w\}_{w\in\mathcal{W}}, \mathcal{V}\rangle$ for $\mathcal{L}_{cf}$ consists of a nonempty set $\mathcal{W}$ of world states, a valuation function $\mathcal{V} : At_{\mathcal{L}_{cf}} \times \mathcal{W} \to \{T, F\}$, and a set of orders $\{\leq_w\}_{w\in\mathcal{W}}$ (one for each world) where $v \leq_w u$ just in case v is at least as similar to w as u ($v <_w u$ abbreviates $v \leq_w u \wedge u \not\leq_w v$). It is required that each $\leq_w$ is transitive and reflexive.

Here is an example. If an arrow extends from v to u in the row for $\leq_w$, then $v \leq_w u$. The arrows required for transitivity and reflexivity are omitted.



For instance, $w_1 <_{w_1} w_2$, $w_2 \leq_{w_2} w_3$, and $w_3 \leq_{w_2} w_2$.

This world-ordering model has some strange features. First, $\leq_{w_2}$ is not defined over all of $\mathcal{W}$ (letting $\mathcal{W}_w$ designate the set of worlds over which $\leq_w$ is defined, $\mathcal{W}_{w_2} \neq \mathcal{W}$). Second, w_3 is as similar to w_2 as w_2 itself. Third, w_1 is *more* similar to w_3 than w_3 itself. Fourth, w_2 are w_3 are incomparable according to $\leq_{w_3}$.

To block these oddities, we might impose some additional constraints.

Definition 8.3. For any proposition $X \subseteq \mathcal{W}$ and world $w \in \mathcal{W}$, the set of **closest X-worlds** to w is this:

$$\mathrm{Min}_{\leq_w}(X) = \{v \in X \cap \mathcal{W}_w : \neg\exists u \in X(u <_w v)\}$$

Centering. $\mathrm{Min}_{\leq_w}(\mathcal{W}) = \{w\}$

Weak Centering. $w \in \mathrm{Min}_{\leq_w}(\mathcal{W})$

Totality. $\forall u, v \in \mathcal{W}_w(v \leq_w u \vee u \leq_w v)$

Well-Foundedness. $\forall X \subseteq \mathcal{W}_w(X \neq \emptyset \supset \mathrm{Min}_{\leq_w}(X) \neq \emptyset)$

In the above model, $\leq_{w_2}$ and $\leq_{w_3}$ violate Centering, $\leq_{w_3}$ violates Weak Centering and Totality, but all of the orderings satisfy Well-Foundedness.

Assuming Well-Foundedness, the semantics for counterfactuals is fairly straightforward:

Definition 8.4. The following recursive clauses lift $\mathcal{V}$ to the complete interpretation function $[\![\ ]\!]_{\mathcal{M}} : S_{\mathcal{L}_{cf}} \times \mathcal{W} \to \{T, F\}$ for $\mathcal{L}_{cf}$:

$$
\begin{array}{rcl}
[\![p]\!]^w_{\mathcal{M}} = T & \text{iff} & \mathcal{V}(p,w) = T \\
[\![\bot]\!]^w_{\mathcal{M}} = T & \text{iff} & 0 = 1 \\
[\![\neg\varphi]\!]^w_{\mathcal{M}} = T & \text{iff} & [\![\varphi]\!]^w_{\mathcal{M}} = F \\
[\![(\varphi \wedge \psi)]\!]^w_{\mathcal{M}} = T & \text{iff} & [\![\varphi]\!]^w_{\mathcal{M}} = [\![\psi]\!]^w_{\mathcal{M}} = T \\
[\![(\varphi \mathrel{\Box\!\!\rightarrow} \psi)]\!]^w_{\mathcal{M}} = T & \text{iff} & \text{Min}_{\leq_w}([\varphi]_{\mathcal{M}}) \subseteq [\psi]_{\mathcal{M}} \\
[\![(\varphi \mathrel{\Diamond\!\!\rightarrow} \psi)]\!]^w_{\mathcal{M}} = T & \text{iff} & \text{Min}_{\leq_w}([\varphi]_{\mathcal{M}}) \cap [\psi]_{\mathcal{M}} \neq \emptyset
\end{array}
$$

That is, $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ is true at w in $\mathcal{M}$ just in case all of the closest φ-worlds to w are also ψ-worlds, and $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi$ is true just in case some of these closest φ-worlds are ψ-worlds.[1]

For instance, in the above model, $[\![(C \mathrel{\Box\!\!\rightarrow} B)]\!]^{w_1}_{\mathcal{M}} = F$, $[\![(C \mathrel{\Diamond\!\!\rightarrow} B)]\!]^{w_2}_{\mathcal{M}} = T$, and $[\![(B \mathrel{\Box\!\!\rightarrow} C)]\!]^{w_3}_{\mathcal{M}} = T$.

3 Inferences

Assuming Centering, it is easy to show that the following inference forms are valid for the counterfactual conditional:

Modus Ponens/Tollens.

$$\frac{\varphi \mathrel{\Box\!\!\rightarrow} \psi \qquad \varphi}{\psi} \qquad \frac{\varphi \mathrel{\Box\!\!\rightarrow} \psi \qquad \neg\psi}{\neg\varphi}$$

However, other inference forms that are valid for the material conditional are (correctly) invalidated by the semantics in §2:

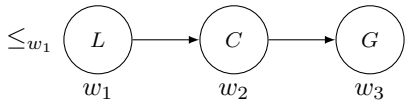
Strengthening the Antecedent.

$$\frac{\varphi \mathrel{\Box\!\!\rightarrow} \chi}{(\varphi \wedge \psi) \mathrel{\Box\!\!\rightarrow} \chi}$$

Counterexample:

(P1) If the Liberals had not won the last election, then the Conservatives would have won it.

(C) If the Liberals had not won the last election and the Greens had gotten ninety percent of the popular vote, then the Conservatives would have won it.



[1] If we drop Well-Foundedness, then the semantic clauses for the counterfactual conditional operators are more complex. See Lewis [1973] for details.

Transitivity.

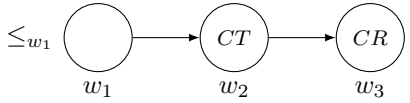
$$\frac{\varphi \mathrel{\Box\!\!\rightarrow} \psi \qquad \psi \mathrel{\Box\!\!\rightarrow} \chi}{\varphi \mathrel{\Box\!\!\rightarrow} \chi}$$

Counterexample:

(P1) If J. Edgar Hoover had been born a Russian, then he would have been a communist.

(P2) If J. Edgar Hoover had been a communist, then he would have been a traitor.

(C) If J. Edgar Hoover had been born a Russian, then he would have been a traitor.



Contraposition.

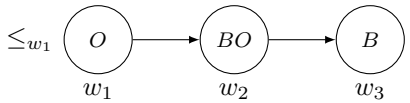
$$\frac{\varphi \mathrel{\Box\!\!\rightarrow} \psi}{\neg\psi \mathrel{\Box\!\!\rightarrow} \neg\varphi}$$

Counterexample:

(P1) If Boris had gone to the party, Olga would still have gone.

(C) If Olga had not gone, then Boris would still not have gone.

(Background context: Boris wanted to attend the party but stayed away to avoid Olga who has been pursuing his heart)



So far so good. But what about the following inference pattern?

Simplification of Disjunctive Antecedents.

$$\frac{(\varphi \vee \psi) \mathrel{\Box\!\!\rightarrow} \chi}{(\varphi \mathrel{\Box\!\!\rightarrow} \chi) \wedge (\psi \mathrel{\Box\!\!\rightarrow} \chi)}$$

SDA is also invalidated by the semantics in §2. But it seems good:

(P1) If either Oswald had not fired or Kennedy had been in a bulletproof car, then Kennedy would still be alive.

(C1) If Oswald had not fired, then Kennedy would still be alive.

(C2) If Kennedy had been in a bulletproof car, then he would still be alive.