

An interactive demonstration of counterfactual truth conditions*

Proposal for a Bachelor Thesis (v0.5, 3 August 2022)

Andreas Paul Bruno Lönne
loenne@campus.tu-berlin.de

Technische Universität Berlin
discourse Degree program: Bachelor Informatik / Computer Science

Definitions

0.1 Counterfactual formulas

$Atoms = \{x, y, \dots\}$
 $\Phi = \{\varphi, \psi, \dots\}$
 $\varphi, \psi ::= \perp \mid x \mid \neg\varphi \mid \Box\varphi \mid \Diamond\varphi \mid \varphi \vee \psi \mid \varphi \wedge \psi \mid \varphi \Box\rightarrow \psi \mid \varphi \Diamond\rightarrow \psi$

0.2 Worlds

$W = \{w, v, \dots\}$

0.3 Facts

$F: W \rightarrow 2^{Atoms}$

0.4 Similarity relation

$\rightsquigarrow: W \times \mathbb{R} \times W$

0.5 Accessible worlds

$W_w = \{w' \mid w \rightsquigarrow^r w'\}$

*Further title proposals: (B) Creating an educational computer game about counterfactuals in terms of a centered system of spheres (C) Implementing a computer game illustrating the truth conditions of counterfactuals as variably strict conditionals

0.6 Truth conditions of counterfactual logic

$w \models \perp$ is always false.

$w \models \top$ is always true.

$w \models x$ iff $x \in F(w)$.

$w \models \neg\varphi$ iff $w \not\models \varphi$.

$w \models \varphi \vee \psi$ iff ($w \models \varphi$ or $w \models \psi$)

$w \models \varphi \wedge \psi$ iff ($w \models \varphi$ and $w \models \psi$)

$w \models \Box\varphi$ iff for every world w' , for which an r with $w \xrightarrow{r} w'$ exists, $w' \models \varphi$ holds true.

$w \models \Diamond\varphi$ iff a world w' and an r exist, such that $w \xrightarrow{r} w'$ and $w' \models \varphi$ hold true.

$w \models \varphi \Box\rightarrow \psi$, if no world w' and r exist, such that $w' \models \varphi$ and $w \xrightarrow{r} w'$.

$w \models \varphi \Box\rightarrow \psi$, if a world w' and an r exist, such that $w' \models \varphi$ and $w \xrightarrow{r} w'$ and for each world w^* , for which a $r^* \leq r$ exists, such that $w \xrightarrow{r^*} w^*$, $w^* \models \psi \vee \neg\varphi$ holds true.

$w \models \varphi \Diamond\rightarrow \psi$, iff a world w' and an r exist, such that $w \xrightarrow{r} w'$ and $w' \models \varphi$ hold and for each world w'' , for which an r'' exists, such that $w \xrightarrow{r''} w''$ and $w'' \models \varphi$ hold true, a world w^* and an r^* exist, such that $r^* \leq r''$ and $w \xrightarrow{r''} w''$ and $w'' \models \varphi \wedge \psi$.

0.7 Similarity graph

$$G = (V, E, F), \text{ such that } V \subseteq W \text{ and } E \subseteq \rightsquigarrow \quad (1)$$

Rules of the semantic game

$$(\top, w)_a \quad \text{Attacker wins} \quad (2)$$

$$(\top, w)_d \quad \text{Defender wins} \quad (3)$$

$$(\perp, w)_a \quad \text{Attacker loses} \quad (4)$$

$$(\perp, w)_d \quad \text{Defender loses} \quad (5)$$

Fig. 1: Win conditions

The win conditions for attacker and defender are identical. A player who reaches a top-symbol wins and a player who reaches a bottom symbol loses. Since attacker and defender are treated equally in this game formulation i will introduce the shorthands $e \in \{a, d\}$ and $o \in \{a, d\} \setminus \{e\}$ to avoid the duplication of every rule.

$$(x, w)_e \xrightarrow{x \in F(w)} (\top, w)_e \quad (6)$$

$$(x, w)_e \xrightarrow{x \notin F(w)} (\perp, w)_e \quad (7)$$

$$(\neg\varphi, w)_e \rightarrow (\varphi, w)_o \quad (8)$$

Fig. 2: Atom resolution & Negation

The negation is resolved by switching the active player.

$$(\varphi \vee \psi, w)_e \rightarrow (\varphi, w)_e \quad (9)$$

$$(\varphi \vee \psi, w)_e \rightarrow (\psi, w)_e \quad (10)$$

$$(\varphi \wedge \psi, w)_e \rightarrow (And, \varphi \wedge \psi, w)_o \quad (11)$$

$$(And, \varphi \wedge \psi, w)_e \rightarrow (\varphi, w)_o \quad (12)$$

$$(And, \varphi \wedge \psi, w)_e \rightarrow (\psi, w)_o \quad (13)$$

Fig. 3: Disjunction & Conjunction

According to the disjunctions truth conditions the active-("proving")-player may choose which subformula to evaluate further. Conversely the conjunctions truth conditions are modelled by allowing the nonactive-("disproving")-player to make that choice instead.

$$(\Diamond\varphi, w)_e \xrightarrow{[w \xrightarrow{r} w']} (\varphi, w')_e \quad (14)$$

$$(\Diamond\varphi, w)_e \rightarrow (\perp, w)_e, \quad (15)$$

if no world w' and r exist, such that $w \xrightarrow{r} w'$.

$$(\Box\varphi, w)_e \rightarrow (Nec, \Box\varphi, w')_o \quad (16)$$

$$(\Box\varphi, w)_e \rightarrow (\top, w)_e \quad (17)$$

if no world w' and r exist, such that $w \xrightarrow{r} w'$.

$$(Nec, \Box\varphi, w)_e \xrightarrow{[w \xrightarrow{r} w']} (\varphi, w')_o \quad (18)$$

Fig. 4: Possibility & Necessity

Regarding the case that no world is accessible the modal possibility operator evaluates to false, since it stipulates the existence of a world. The necessity operator on the other hand only stipulates its subformula to hold at each accessible world and would thus be vacuously true.

$$(\varphi \Diamond \rightarrow \psi, w)_e \xrightarrow{[w \rightsquigarrow^r w']} (Cf, \varphi \Diamond \rightarrow \psi, w, w', r)_o \quad (19)$$

$$(\varphi \Diamond \rightarrow \psi, w)_e \rightarrow (\perp, w)_e \quad (20)$$

if no world w' and r exist, such that $w \rightsquigarrow^r w'$.

$$(Cf, \varphi \Diamond \rightarrow \psi, w, w', r)_e \xrightarrow{[w \rightsquigarrow^{r*} w^*, r^* < r]} (\varphi, w^*)_o \quad (21)$$

$$(Cf, \varphi \Diamond \rightarrow \psi, w, w', r)_e \rightarrow (\varphi \wedge \psi, w')_o \quad (22)$$

Fig. 5: Counterfactual might

The truth conditions of the counterfactual might operator are as described in figure 7.

I have reduced Lewis' truth conditions to the existence of a φ & ψ -world among the closest φ -worlds. This is derived from Lewis' truth conditions as follows.

Suppose no φ -world is accessible from w , that is, no r and w' exist such that $w \rightsquigarrow^r w'$ and $w' \models \varphi$. Then it follows that no r and w' exist such that $w \rightsquigarrow^r w'$ and $w' \models \varphi \wedge \psi$.

Now suppose that a closest φ -world w' to w exists, that is, some r and w' exist such that $w \rightsquigarrow^r w'$ and $w' \models \varphi$ and no r^* and w^* exist such that $w \rightsquigarrow^{r*} w^*$, $w^* \models \varphi$ and $r^* < r$.

Then if a φ & ψ -world w'' exists such that $w \rightsquigarrow^r w''$, $w'' \models \varphi \wedge \psi$, w'' is included in every sphere w' is included in. Since we supposed that w' is the closest φ -world to w , we can conclude by the nesting property of centered systems of spheres, that the set of spheres centered on w , that contain w' , is the same set as the set of spheres centered on w , that contain at least one φ -world. Thus every sphere centered on w , that contains a φ -world also contains a $\varphi \wedge \psi$ -world.

On the other hand if no φ & ψ -world w'' exists such that $w \rightsquigarrow^r w''$, $w'' \models \varphi \wedge \psi$, then there simply exists the sphere delimited by w' and centered on w , that contains at least a φ -world we named w' and no $\varphi \wedge \psi$ -world.

Here just a few quick comments about the Rules:

Rule (20) makes the active-("proving")-player lose in case no worlds are accessible at all.

Rule (19) is the active players choice of a sphere of accessibility, by choosing a delimiting world. The delimiting world has to be a $\varphi \wedge \psi$ -world and a closest φ -world to win the game.

Rule (21) gives the previous non-active-player the opportunity to disprove the chosen world is a closest φ -world.

Rule (22) lets the previous non-active-player evaluate the chosen world on whether it is a $\varphi \wedge \psi$ -world.

$$(\varphi \Box \rightarrow \psi, w)_e \rightarrow (\text{Would}, \varphi \Box \rightarrow \psi, w)_o \quad (23)$$

$$(\text{Would}, \varphi \Box \rightarrow \psi, w)_e \rightarrow (\perp, w)_e \quad (24)$$

if no world w' and r exist, such that $w \overset{r}{\rightsquigarrow} w'$.

$$(\text{Would}, \varphi \Box \rightarrow \psi, w)_e \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (Cf, \varphi \Box \rightarrow \psi, w, w', r)_o \quad (25)$$

$$(Cf, \varphi \Box \rightarrow \psi, w, w', r)_e \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* < r]} (\varphi, w^*)_e \quad (26)$$

$$(Cf, \varphi \Box \rightarrow \psi, w, w', r)_e \rightarrow (\neg \varphi \vee \psi, w')_e \quad (27)$$

Fig. 6: Counterfactual would

The rules for the counterfactual would are defined analogously to the counterfactual might rules and stated simplest by the question "Is no $\varphi \ \& \ \neg\psi$ -world among the closest φ -worlds?".

The non-active-("disproving")-player becomes active and has to choose a $\varphi \ \& \ \neg\psi$ -world, thats also a closest φ -world to win. Then the formerly active-("proving")-player becomes active once more and may choose to either contend that the chosen world is a closest φ -world (Rule (26)) or claim that the chosen world is not a $\varphi \ \& \ \neg\psi$ -world.

Note that the game is defined in a way to have the defender i.e. player always start as the active player and prove a formula. Should the defender not start as the active player their goal simply changes to disprove the formula instead of proving it.

$\phi \Diamond \rightarrow \psi$ is true at a world i (according to a system of spheres $\$$) if and only if both

- (1) some ϕ -world belongs to some sphere S in $\$_i$, and
- (2) every sphere S in $\$_i$ that contains at least one ϕ -world contains at least one world where $\phi \ \& \ \psi$ holds.

Fig. 7: Lewis' counterfactual might truth conditions

$\phi \Box \rightarrow \psi$ is true at a world i (according to a system of spheres $\$$) if and only if either

- (1) no ϕ -world belongs to any sphere S in $\$_i$, or
- (2) some sphere S in $\$_i$ does contain at least one ϕ -world, and $\phi \supset \psi$ holds at every world in S .

Fig. 8: Lewis' counterfactual would truth conditions

Let $\$$ be an assignment to each possible world i of a set $\$_i$ of sets of possible worlds. Then $\$$ is called a (centered) system of spheres, and the members of each $\$_i$ are called spheres around i , if and only if, for each world i , the following conditions hold.

- (C) $\$_i$ is centered on i ; that is, the set $\{i\}$ having i as its only member belongs to $\$_i$.
- (1) $\$_i$ is nested; that is, whenever S and T belong to $\$_i$, either S is included in T or T is included in S .
- (2) $\$_i$ is closed under unions; that is, whenever \mathcal{S} is a subset of $\$_i$ and $\bigcup \mathcal{S}$ is the set of all worlds j such that j belongs to some member of \mathcal{S} , $\bigcup \mathcal{S}$ belongs to $\$_i$.
- (3) $\$_i$ is closed under (nonempty) intersections; that is, whenever \mathcal{S} is a nonempty subset of $\$_i$ and $\bigcap \mathcal{S}$ is the set of all worlds j such that j belongs to every member of \mathcal{S} , $\bigcap \mathcal{S}$ belongs to $\$_i$.

Fig. 9: Lewis' (centered) system of spheres

Correctness and Termination Proofs

0.8 Strategy

A strategy is a partial function $s : S \rightarrow P$ from the set of game states S to the set of plays \rightarrow . A strategy is called *winning*, if no losing sequence of moves exists, where every move the defender made was part of the winning strategy.

0.9 Correctness

This proof will show that whenever a counterfactual formula is true, its proving player has a winning strategy.

For the formula \top , that is satisfied by every world w ($w \models \top$), and is consequently always true, the proving player either wins through strategy s with $((\top, w)_e, \text{rule2}) \in s$ or $((\top, w)_e, \text{rule3}) \in s$.

For the formula \perp , that is satisfied by no world and is consequently never true, the proving player loses as per rules 4 and 5.

0.10 stuff

This proof will show that (a) whenever a counterfactual formula is true, the defender has a winning strategy and that (b) the semantic game always terminates. I will show, that for each formula, the corresponding semantic game has these properties.

For the formula \top , that is satisfied by every world w ($w \models \top$), and is consequently always true, the corresponding starting state is $(\top, w)_d$. As per rule 2 the defender wins.

The formula $\neg\varphi$ is satisfied by any world w ($w \models \neg\varphi$), that doesn't satisfy φ ($w \models \varphi$ is false). If the disproving player of $\neg\varphi$ can prove $w \models \varphi$, he also disproves $w \models \neg\varphi$.

The formula $\varphi \vee \psi$ is satisfied by a world w ($w \models \varphi \vee \psi$) iff either φ is satisfied by w ($w \models \varphi$) or ψ is satisfied by w ($w \models \psi$). If $w \models \varphi$ or $w \models \psi$, the defender can choose the according case and thus, win.