Chapter udf

Minimal Change Semantics

min.1 Introduction

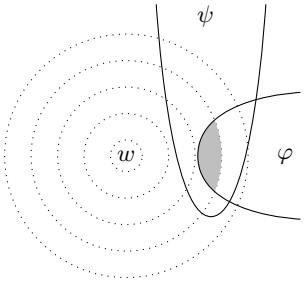
cnt:min:int:sec

Stalnaker and Lewis proposed accounts of counterfactual conditionals such as "If the match were struck, it would light." Their accounts were proposals for how to properly understand the truth conditions for such sentences. The idea behind both proposals is this: to evaluate whether a counterfactual conditional is true, we have to consider those possible worlds which are minimally different from the way the world actually is to make the antecedent true. If the consequent is true in these possible worlds, then the counterfactual is true. For instance, suppose I hold a match and a matchbook in my hand. In the actual world I only look at them and ponder what would happen if I were to strike the match. The minimal change from the actual world where I strike the match is that where I decide to act and strike the match. It is minimal in that nothing else changes: I don't also jump in the air, striking the match doesn't also light my hair on fire, I don't suddenly lose all strength in my fingers, I am not simultaneously doused with water in a SuperSoaker ambush, etc. In that alternative possibility, the match lights. Hence, it's true that if I were to strike the match, it would light.

This intuitive account can be paired with formal semantics for logics of counterfactuals. Lewis introduced the symbol "$\boxright$" for the counterfactual while Stalnaker used the symbol ">". We'll use $\boxright$, and add it as a binary connective to propositional logic. So, we have, in addition to formulas of the form $\varphi \to \psi$ also formulas of the form $\varphi \boxright \psi$. The formal semantics, like the relational semantics for modal logic, is based on models in which formulas are evaluated at worlds, and the satisfaction condition defining $\mathfrak{M}, w \Vdash \varphi \boxright \psi$ is given in terms of $\mathfrak{M}, w' \Vdash \varphi$ and $\mathfrak{M}, w' \Vdash \psi$ for some (other) worlds w'. Which w'? Intuitively, the one(s) closest to w for which it holds that $\mathfrak{M}, w' \Vdash \varphi$. This requires that a relation of "closeness" has to be included in the model as well.

Lewis introduced an instructive way of representing counterfactual situations graphically. Each possible world is at the center of a set of nested spheres containing other worlds—we draw these spheres as concentric circles. The

worlds between two spheres are equally close to the world at the center as each other, those contained in a nested sphere are closer, and those in a surrounding sphere further away.



The closest φ-worlds are those worlds w' where φ is satisfied which lie in the smallest sphere around the center world w (the gray area). Intuitively, $\varphi \boxdot\!\!\rightarrow \psi$ is satisfied at w if ψ is true at all closest φ-worlds.

min.2 Sphere Models

con:min:sph:sec

One way of providing a formal semantics for counterfactuals is to turn Lewis's informal account into a mathematical structure. The spheres around a world w then are sets of worlds. Since the spheres are nested, the sets of worlds around w have to be linearly ordered by the subset relation.

Definition min.1. A *sphere model* is a triple $\mathfrak{M} = \langle W, O, V\rangle$ where W is a non-empty set of worlds, $V\colon \mathrm{At}_0 \to \wp(W)$ is a valuation, and $O\colon W \to \wp(\wp(W))$ assigns to each world w a *system of spheres* O_w. For each w, O_w is a set of sets of worlds, and must satisfy:

1. O_w is *centered* on w: $\{w\} \in O_w$.
2. O_w is *nested*: whenever S_1, $S_2 \in O_w$, $S_1 \subseteq S_2$ or $S_2 \subseteq S_1$, i.e., O_w is linearly ordered by $\subseteq$.
3. O_w is closed under non-empty unions.
4. O_w is closed under non-empty intersections.

The intuition behind O_w is that the worlds "around" w are stratified according to how far away they are from w. The innermost sphere is just w by itself, i.e., the set $\{w\}$: w is closer to w than the worlds in any other sphere. If $S \subsetneq S'$, then the worlds in $S' \setminus S$ are further way from w than the worlds in S: $S' \setminus S$ is the "layer" between the S and the worlds outside of S'. In particular, we have to think of the spheres as containing all the worlds within their outer surface; they are not just the individual layers.

The diagram in Figure min.1 corresponds to the sphere model with $W = \{w, w_1, \ldots, w_7\}$, $V(p) = \{w_5, w_6, w_7\}$. The innermost sphere $S_1 = \{w\}$. The

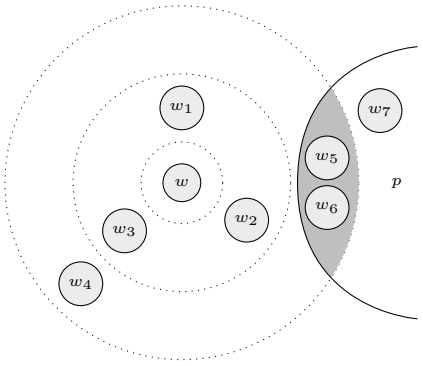


Figure min.1: Diagram of a sphere model

con:min:sph:
fig:sphere-model

closest worlds to w are w_1, w_2, w_3, so the next larger sphere is $S_2 = \{w, w_1, w_2, w_3\}$. The worlds further out are w_4, w_5, w_6, so the outermost sphere is $S_3 = \{w, w_1, \ldots, w_6\}$. The system of spheres around w is $O_w = \{S_1, S_2, S_3\}$. The world w_7 is not in any sphere around w. The closest worlds in which p is true are w_5 and w_6, and so the smallest p-admitting sphere is S_3.

To define satisfaction of a formula φ at world w in a sphere model $\mathfrak{M}$, $\mathfrak{M}, w \Vdash \varphi$, we expand the definition for modal formulas to include a clause for $\psi \boxdot\!\!\rightarrow \chi$:

Definition min.2. $\mathfrak{M}, w \Vdash \psi \boxdot\!\!\rightarrow \chi$ iff either

con:min:sph:
sphere-vac

1. For all $u \in \bigcup O_w$, $\mathfrak{M}, u \nVdash \chi$, or

con:min:sph:
sphere-nonvac

2. For some $S \in O_w$,
   a) $\mathfrak{M}, u \Vdash \psi$ for some $u \in S$, and
   b) for all $v \in S$, either $\mathfrak{M}, v \nVdash \psi$ or $\mathfrak{M}, v \Vdash \chi$.

According to this definition, $\mathfrak{M}, w \Vdash \psi \boxdot\!\!\rightarrow \chi$ iff either the antecedent ψ is false everywhere in the spheres around w, or there is a sphere S where ψ is true, and the material conditional $\psi \rightarrow \chi$ is true at all worlds in that "ψ-admitting" sphere. Note that we didn't require in the definition that S is the *innermost* ψ-admitting sphere, contrary to what one might expect from the intuitive explanation. But if the condition in (2) is satisfied for some sphere S, then it is also satisfied for all spheres S contains, and hence in particular for the innermost sphere.

Note also that the definition of sphere models does not require that there *is* an innermost ψ-admitting sphere: we may have an infinite sequence $S_1 \supsetneq S_2 \supsetneq \cdots \supsetneq \{w\}$ of ψ-admitting spheres, and hence no innermost ψ-admitting spheres. In that case, $\mathfrak{M}, w \Vdash \psi \boxdot\!\!\rightarrow \chi$ iff $\psi \rightarrow \chi$ holds throughout the spheres S_i, S_{i+1}, ..., for some i.

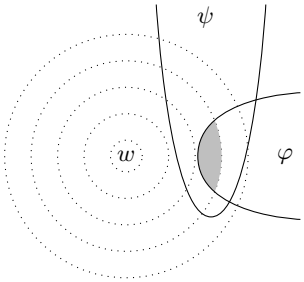


Figure min.2: Non-vacuously true counterfactual

con:min:tf:
fig:true

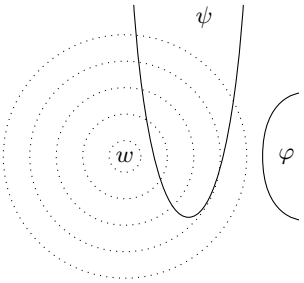


Figure min.3: Vacuously true counterfactual

con:min:tf:
fig:vacuous

min.3 Truth and Falsity of Counterfactuals

con:min:tf:
sec

A counterfactual $\varphi \boxdot\!\!\rightarrow \psi$ is (non-vacuously) true if the closest φ-worlds are all ψ-worlds, as depicted in Figure min.2. A counterfactual is also true at w if the system of spheres around w has no φ-admitting spheres at all. In that case it is *vacuously* true (see Figure min.3).

It can be false in two ways. One way is if the closest φ-worlds are not all ψ-worlds, but some of them are. In this case, $\varphi \boxdot\!\!\rightarrow \neg\psi$ is also false (see Figure min.4). If the closest φ-worlds do not overlap with the ψ-worlds at all, then $\varphi \boxdot\!\!\rightarrow \psi$. But, in this case all the closest φ-worlds are $\neg\psi$-worlds, and so $\varphi \boxdot\!\!\rightarrow \neg\psi$ is true (see Figure min.5).

In contrast to the strict conditional, counterfactuals may be contingent. Consider the sphere model in Figure min.6. The φ-worlds closest to u are all ψ-worlds, so $\mathfrak{M}, u \Vdash \varphi \boxdot\!\!\rightarrow \psi$. But there are φ-worlds closest to v which are not ψ-worlds, so $\mathfrak{M}, v \nVdash \varphi \boxdot\!\!\rightarrow \psi$.

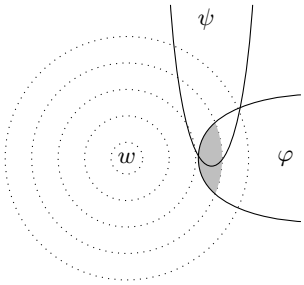


Figure min.4: False counterfactual, false opposite

con:min:tf:
fig:false

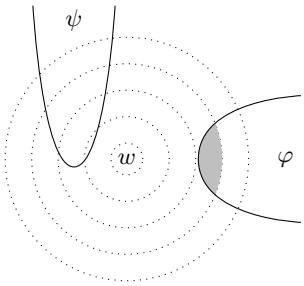


Figure min.5: False counterfactual, true opposite

con:min:tf:
fig:false-opposite

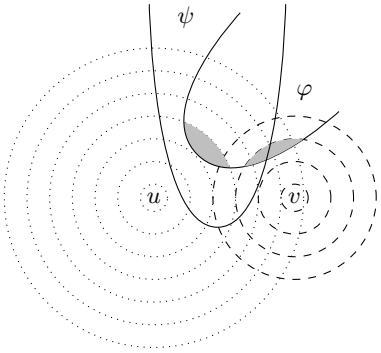


Figure min.6: Contingent counterfactual

con:min:tf:
fig:contingent

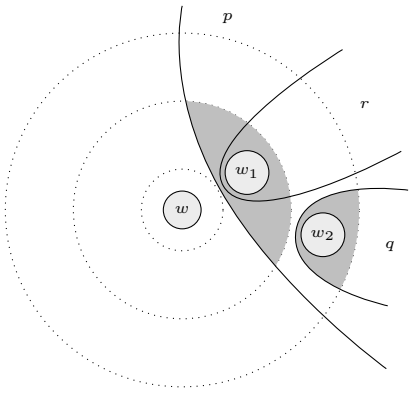


Figure min.7: Counterexample to antecedent strengthening

cnt:min:agg:
fig:antecedent-strengthening

min.4 Antecedent Strengthenng

"Strengthening the antecedent" refers to the inference $\varphi \to \chi \vDash (\varphi \wedge \psi) \to \chi$. It is valid for the material conditional, but invalid for counterfactuals. Suppose it is true that if I were to strike this match, it would light. (That means, there is nothing wrong with the match or the matchbook surface, I will not break the match, etc.) But it is not true that if I were to light this match in outer space, it would light. So the following inference is invalid:

cnt:min:agg:
sec

> I the match were struck, it would light.
>
> Therefore, if the match were struck in outer space, it would light.

The Lewis-Stalnaker account of conditionals explains this: the closest world where I light the match and I do so in outer space is much further removed from the actual world than the closest world where I light the match is. So although it's true that the match lights in the latter, it is not in the former. And that is as it schould be.

Example min.3. The sphere semantics invalidates the inference, i.e., we have $p \boxdot\!\!\!\rightarrow r \nvDash (p \wedge q) \boxdot\!\!\!\rightarrow r$. Consider the model $\mathfrak{M} = \langle W, O, V \rangle$ where $W = \{w, w_1, w_2\}$, $O_w = \{\{w\}, \{w, w_1\}, \{w, w_1, w_2\}\}$, $V(p) = \{w_1, w_2\}$, $V(q) = \{w_2\}$, and $V(r) = \{w_1\}$. There is a p-admitting sphere $S = \{w, w_1\}$ and $p \to r$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash p \boxdot\!\!\!\rightarrow r$. There is also a $(p \wedge q)$-admitting sphere $S' = \{w, w_1, w_2\}$ but $\mathfrak{M}, w_2 \nVdash (p \wedge q) \to r$, so $\mathfrak{M}, w \nVdash (p \wedge q) \boxdot\!\!\!\rightarrow r$ (see Figure min.7).

min.5 Transitivity

cnt:min:tra:
sec

For the material conditional, the chain rule holds: $\varphi \to \psi, \psi \to \chi \vDash \varphi \to \chi$. In other words, the material conditional is transitive. Is the same true for counterfactuals? Consider the following example due to Stalnaker.

> If J. Edgar Hoover had been born a Russian, he would have been a Communist.
>
> If J. Edgar Hoover were a Communist, he would have been be a traitor.
>
> Therefore, If J. Edgar Hoover had been born a Russian, he would have been be a traitor.

If Hoover had been born (at the same time he actually did), not in the United States, but in Russia, he would have grown up in the Soviet Union and become a Communist (let's assume). So the first premise is true. Likewise, the second premise, considered in isolation is true. The conclusion, however, is false: in all likelihood, Hoover would have been a fervent Communist if he had been born in the USSR, and not been a traitor (to his country). The intuitive assignment of truth values is borne out by the Stalnaker-Lewis account. The closest possible world to ours with the only change being Hoover's place of birth is the one where Hoover grows up to be a good citizen of the USSR. This is the closest possible world where the antecedent of the first premise and of the conclusion is true, and in that world Hoover is a loyal member of the Communist party, and so not a traitor. To evaluate the second premise, we have to look at a different world, however: the closest world where Hoover is a Communist, which is one where he was born in the United States, turned, and thus became a traitor.[1]

Problem min.1. Find a convincing, intuitive example for the failure of transitivity of counterfactuals.

cnt:min:tra: ex:trans-counterex

Example min.4. The sphere semantics invalidates the inference, i.e., we have $p \boxrightarrow q, q \boxrightarrow r \nvDash p \boxrightarrow r$. Consider the model $\mathfrak{M} = \langle W, O, V\rangle$ where $W = \{w, w_1, w_2\}$, $O_w = \{\{w\}, \{w, w_1\}, \{w, w_1, w_2\}\}$, $V(p) = \{w_2\}$, $V(q) = \{w_1, w_2\}$, and $V(r) = \{w_1\}$. There is a p-admitting sphere $S = \{w, w_1, w_2\}$ and $q \to q$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash p \boxrightarrow q$. There is also a q-admitting sphere $S' = \{w, w_1\}$ and $\mathfrak{M} \nVdash q \to r$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash q \boxrightarrow r$. However, the p-admitting sphere $\{w, w_1, w_2\}$ contains a world, namely w_2, where $\mathfrak{M}, w_2 \nVdash p \to r$.

Problem min.2. Draw the sphere diagram corresponding to the counterexample in Example min.4.

Problem min.3. In Example min.4, world w_2 is where Hoover is born in Russia, is a communist, and not a traitor, and w_1 is the world where Hoover is born in the US, is a communist, and a traitor. In this model, w_1 is closer to w

[1] Of course, to appreciate the force of the example we have to take on board some metaphysical and political assumptions, e.g., that it is possible that Hoover could have been born to Russian parents, or that Communists in the US of the 1950s were traitors to their country.

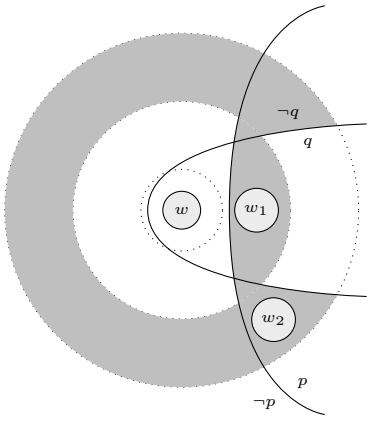


Figure min.8: Counterexample to contraposition

cnt:min:cpo: fig:contraposition

than w_2 is. Is this necessary? Can you give a counterexample that does not assume that Hoover's being born in Russia is a more remote possibility than him being a Communist?

min.6 Contraposition

cnt:min:cpo: sec

Material and strict conditionals are equivalent to their contrapositives. Counterfactuals are not. Here is an example due to Kratzer:

> If Goethe hadn't died in 1832, he would (still) be dead now.
>
> If Goethe weren't dead now, he would have died in 1832.

The first sentence is true: humans don't live hundreds of years. The second is clearly false: if Goethe weren't dead now, he would be still alive, and so couldn't have died in 1832.

cnt:min:cpo: ex:contraposition-counterex

Example min.5. The sphere semantics invalidates contraposition, i.e., we have $p \boxdot\!\!\rightarrow q \not\models \neg q \boxdot\!\!\rightarrow \neg p$. Think of p as "Goethe didn't die in 1832" and q as "Goethe is dead now." We can capture this in a model $\mathfrak{M}_1 = \langle W, O, V\rangle$ with $W = \{w, w_1, w_2\}$, $O = \{\{w\}, \{w, w_1\}, \{w, w_1, w_2\}\}$, $V(p) = \{w_1, w_2\}$ and $V(q) = \{w, w_1\}$. So w is the actual world where Goethe died in 1832 and is still dead; w_1 is the (close) world where Goethe died in, say, 1833, and is still dead; and w_2 is a (remote) world where Goethe is still alive. There is a p-admitting sphere $S = \{w, w_1\}$ and $p \rightarrow q$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash p \boxdot\!\!\rightarrow q$. However, the $\neg q$-admitting sphere $\{w, w_1, w_2\}$ contains a world, namely w_2, where q is false and p is true, so $\mathfrak{M}, w_2 \nVdash \neg q \rightarrow \neg p$.

Photo Credits

Bibliography