# CS112 (Fall 2018)
# Counterfactual Logic I
# October 9, 2018

## 1 Introduction

We have previously seen the *logical problem of conditionals*. To review: in propositional logic, a conditional sentence $p \rightarrow q$ is true if $p$ is false or $q$ is true. This problem is especially apparent when considering counterfactual sentences, whose antecedents are presupposed to be false. For example, the translations of the following two sentences into propositional logic are both true:

If kangaroos had no tails, they would topple over.
If kangaroos had no tails, they would not topple over.

But these sentences cannot both be true; in fact, they can be seen as negations of each other. Therefore, a different analysis is needed.

We have also previously seen one such analysis, using modal logic. Namely, natural language conditionals are to be interpreted as strict conditionals $\Box(p \rightarrow q)$. Under this analysis, $\Box(p \rightarrow q)$ is true if $q$ follows from $p$ in all possible worlds (accessible from the current world).

## 2 The Stalnaker and Lewis approach

The question now becomes: What is the accessibility relation? First, we note that not every world should be accessible from the current world. Consider a world where kangaroos had no tails, but they used crutches. In that world, they would not topple over. But we still consider the counterfactual "If kangaroos had no tails, they would topple over" to be true.

According to Stalnaker and Lewis, worlds in which kangaroos use crutches are "too far away" from the actual world to be relevant to the counterfactual above. This goes back to a suggestion by Ramsey (the "Ramsey test"): to determine the truth of a conditional, add the antecedent to your stock of beliefs, and determine the truth of the consequent. In the case of a counterfactual, adding

the antecedent to your beliefs results in a contradiction, so you must modify your beliefs to resolve the contradiction (in this case, removing the belief that kangaroos have tails). But there is no need to change your belief about whether kangaroos use crutches or not—that can remain the same as in the actual world.

Lewis therefore argues that the accessibility relation should be framed in terms of *comparative similarity of possible worlds*. On one extreme, there is the accessibility relation in which for each world $i$, every other world is accessible, corresponding to logical necessity as above. On the other hand, it is possible for no worlds to be accessible, or for the only accessible world from some world $i$ to be $i$ itself. For some given accessibility relation and some world $i$, it is possible to arrange the worlds accessible to $i$ in a *sphere of accessibility* around $i$. Then the different accessiblity relations form a system of spheres around $i$.

In order for some system of spheres $\$_i$ around some world $i$ to adequately represent comparative similarity, it must satisfy the following conditions:

- $\$_i$ is *centered on* $i$: there exists a sphere in $\$_i$ that contains only $i$.
- $\$_i$ is *nested*: if $S$ and $T$ are spheres in $\$_i$, either $S$ is included in $T$ or $T$ is included in $S$.
  - The first two conditions say that no world is more similar to $i$ than $i$ itself, and that the spheres in $\$_i$ are totally ordered with respect to each other.
- $\$_i$ is *closed under union*.
- $\$_i$ is *closed under intersection*.
  - The last two conditions say that any set of worlds $S$ such that any world within $S$ is more similar to $i$ than any world outside of $S$, is a member of $\$_i$.

Finally, Lewis defines the counterfactual conditional $\phi \mathrel{\Box\!\!\rightarrow} \psi$ as follows:
$\phi \mathrel{\Box\!\!\rightarrow} \psi$ is true at a world $i$ (according to a system of spheres $\$_i$) if and only if either:

1. no $\phi$-world belongs to any sphere $S$ in $\$_i$, or
   - In other words, counterfactuals with impossible antecedents are vacuously true.
2. some sphere $S$ in $\$_i$ does contain at least one $\phi$-world, and $\phi \rightarrow \psi$ holds at every world in $S$.
   - Lewis' counterfactuals are *variably strict conditionals*. Instead of a single accessibility relation, there is a set of accessibility relations (i.e. spheres), each of which corresponds to a certain "distance" from the actual world. Then counterfactuals are strict conditionals over some relation where some $\phi$-world is accessible.

Stalnaker arrives at his theory of conditionals via a different approach, but we will see that his theory is very similar to Lewis'. Stalnaker says little about the accessibility relation; instead, he introduces a *selection function* $f$, that ends up serving the same purpose. The selection function $f$ is a function of a proposition $A$ and a possible world $\alpha$ that returns the most similar world to $\alpha$ where $A$ is true. The selection function has the following properties:

1. For all antecedents $A$ and base worlds $\alpha$, $A$ must be true in $f(A, \alpha)$.
2. For all antecedents $A$ and base worlds $\alpha$, $f(A, \alpha) = \lambda$ only if there is no world possible with respect to $\alpha$ where $A$ is true.
    - $\lambda$ is the *absurd world*, where everything is true, even $\perp$. This property ensures that counterfactuals with impossible antecedents are vacuously true, as in Lewis' theory.
3. For all base worlds $\alpha$ and antecedents $A$, if $A$ is true in $\alpha$, then $f(A, \alpha) = \alpha$.
    - This formalizes the notion that no world is more similar to $\alpha$ than $\alpha$ itself.
4. For all base worlds $\alpha$ and antecedents $B$ and $B'$, if $B$ is true in $f(B', \alpha)$ and $B'$ is true in $f(B, \alpha)$, then $f(B, \alpha) = f(B', \alpha)$.
    - This property allows for the possible worlds to be totally ordered according to similarity to $\alpha$.

Stalnaker defines the counterfactual conditional (which he calls $>$) as follows:

$A > B$ is true in $\alpha$ if $B$ is true in $f(A, \alpha)$;
$A > B$ is false in $\alpha$ if $B$ is false in $f(A, \alpha)$.

We note that if a sentence $A > B$ is true (false) in Stalnaker's theory, then the equivalent sentence $A \mathrel{\square\!\!\rightarrow} B$ is true (false) in Lewis' theory. First, if the antecedent $A$ is impossible, then the counterfactual is vacuously true in both theories.

Now consider the case where $A > B$ is true. We note that $A$ and $B$ are both true in $f(A, \alpha)$, and that $A$ is false in all worlds that are more similar to $\alpha$ than $f(A, \alpha)$ (because if $A$ were true in one of those worlds, that world would have been selected instead). Now consider the set $S$ containing $f(A, \alpha)$ and all worlds more similar to $\alpha$. Because the worlds form a total order, and $\$_\alpha$ (the system of spheres around $\alpha$) is closed under union and intersection, $S$ forms a sphere in $\$_\alpha$. Furthermore, $A \rightarrow B$ is true in all worlds in $S$ (because $A$ and $B$ are true in $f(A, \alpha)$, and $A$ is false in all other worlds in $S$). Therefore, if $A > B$ is true, then $A \mathrel{\square\!\!\rightarrow} B$ is true.

Finally, consider the case where $A > B$ is false, and define the set $S$ as above. We note that because $A$ is true but $B$ is false in $f(A, \alpha)$, it is not the case that

$A \rightarrow B$ is true in all worlds in $S$. Any sphere smaller than $S$ cannot contain an $A$-world (because then that world would have been selected instead). Furthermore, it cannot be the case that $A \rightarrow B$ is true in all worlds in some sphere larger than $S$, because $\$_\alpha$ is nested, so such a sphere also contains $f(A, \alpha)$. Therefore, there is no sphere that satisfies the necessary requirements, and so $A \boxrightarrow B$ is false.

What about the reverse: if a sentence $\phi \boxrightarrow \psi$ is true (false) in Lewis' theory, is the equivalent sentence $\phi > \psi$ true (false) in Stalnaker's theory? Here we run into a bit of a problem. In Lewis' theory, it is possible for there to be two distinct worlds that are members of exactly the same spheres around some world $i$, and therefore equally similar to $i$. This is not possible in Stalnaker's theory, where ties are not possible. However, if it is possible to break ties, then the two semantics are equivalent again (for a proof, see chapter 3.4 of Lewis). Therefore, in a sense, Lewis' theory is a more "general" version of Stalnaker's theory.

## 2.1 Modal properties and false inferences

Both Stalnaker and Lewis define the modal operators $\Box$ and $\Diamond$ in terms of their conditional operators:

$\Box A := \neg A > A$
$\Diamond A := \neg(A > \neg A)$

(Lewis uses the "might" counterfactual, defined as $\phi \diamondrightarrow \psi := \neg(\phi \boxrightarrow \neg\psi)$, in his definitions; it is easily verified that his definitions are equivalent to Stalnaker's.) We can see that these modal operators are equivalent to the standard ones we have seen, if we take the accessibility relation to be Stalnaker's original accessibility relation, or the outermost sphere in Lewis' theory.

There are certain inferences that apply to the material and strict conditionals, but not to Lewis and Stalnaker's conditionals. We have already seen the failure of *strengthening the antecedent* (if $p \rightarrow q$, then $(p \wedge r) \rightarrow q$), as evidenced by the following sentences:

If kangaroos had no tails, they would topple over.
If kangaroos had no tails but used crutches, they would not topple over.

More generally, the counterfactual conditional is not *transitive*: if $p > q$ and $q > r$, then it is not necessarily the case that $p > r$. Consider the following sentences:

If J. Edgar Hoover had been born a Russian, then he would today be a communist.

If J. Edgar Hoover were today a communist, then he would be a traitor.

If J. Edgar Hoover had been born a Russian, he would be a traitor.

The first two sentences are true, but the third is false. According to Lewis, communist-worlds are more similar to the actual world than Russian-worlds, so it is possible for all the nearest communist-worlds to be traitor-worlds, but the nearest Russian-worlds to include some non-traitor worlds.

Finally, we have a failure of *contraposition*: if $p > q$, then it is not necessarily the case that $\neg q > \neg p$. Consider the following sentences:

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

If Olga had not gone, Boris would still not have gone.

Suppose that in the actual world, Olga went to the party but Boris did not. In the next most similar world, both went to the party, and in the world after that, Boris went to the party but Olga did not. In such a model, the first sentence is true, but the second is false.

# 3 Kratzer's premise semantics

Finally, we turn our attention to Kratzer's theory of counterfactuals, called premise semantics. Formulated differently from Stalnaker and Lewis' theories, Kratzer's theory ends up being similar in many ways, yet more amenable to certain extensions of counterfactual theory we will discuss later.

In Kratzer's theory, propositions are defined as sets of possible worlds (in which they are true). Let $f$ be a function that maps a possible world $w$ to the set of propositions true at $w$. Then we can define the proposition $p$ corresponding to a sentence "If it were the case that $q$, then it would be the case that $r$" as the set of worlds $w$ such that:

If $A_w(q)$ is the set of all consistent subsets of $f(w) \cup \{q\}$ which contain $q$, then every set $X$ in $A_w(q)$ has a superset $Y$ in $A_w(q)$ such that $\bigcap Y \subseteq r$.

In the above definition, since $f(w)$ is the set of propositions true at $w$, $A_w(q)$ can be seen as the set of sets of propositions in $w$ consistent with $q$. Then the requirement that $\bigcap Y \subseteq r$ can be seen as requiring that if you start from $\{q\}$ and add as many propositions from $w$ as consistency permits, $r$ follows. This analysis results in the following *critical truth-conditions*:

- If $q$ is true, then $p$ is true if and only if $r$ is true (in the world under consideration).
- If $q$ is false, then $p$ is true (in the world under consideration) if and only if $r$ follows from $q$ (i.e., $q \subseteq r$).

We can see that $p$ reduces to a material conditional $q \rightarrow r$ if $q$ is true and a strict conditional $\Box(q \rightarrow r)$ (according to the accessibility relation corresponding to logical necessity) if $q$ is false. The first condition is fine (recall that Stalnaker and Lewis' counterfactual operators also satisfy this condition). However, we have previously seen that the strict conditional (at least with that accessibility relation) is not what we want.

Kratzer's solution is to redefine $f$ (the *partition function*) such that it maps a possible world $w$ to the set of *facts* true at $w$. A set of facts true at $w$ is a set of propositions that uniquely characterizes $w$; it can consist of the set of all propositions true at $w$ as above, the set $\{w\}$, or something in between. A choice of facts represents a choice of how similar we require some possible world to be to $w$ (since we are adding as many facts as consistency permits). In fact, it is possible (see Lewis 1981) to prove that Kratzer's theory is a more "general" version of Lewis' theory.

Here we note two things. First, not all facts are created equal. Pollock suggests that laws of nature (or *strong subjunctive generalizations*) are more important than actual necessities (or *weak subjunctive generalizations*), which are more important than other facts. In Kratzer's theory, a similar effect is obtained by "lumping" facts together. We will study a related proposal in the next lecture.

Second, the partition function $f$ depends on *context*. Counterfactuals are vague, and some of this vagueness comes from the question of which partition function to use. Stalnaker alludes to this problem when he discusses the *pragmatic problem of counterfactuals*. Lewis is dismissive of the role of context, saying that at most, the context is the antecedent itself, and that it is "defeatist" to appeal to context in this way. However, we will study ways in which the meaning of a counterfactual depends on context (even beyond the antecedent) in the next lecture.

# 4 References and further reading

Primary references:


1. Angelika Kratzer. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, 10(2):201–216, 1981.

2. David Lewis. *Counterfactuals*. Blackwell, 1973.

3. Robert C. Stalnaker. *A Theory of Conditionals*, pages 41–55. Springer Netherlands, Dordrecht, 1968.


Further reading:


4. David Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10(2):217–234, 1981.

    - Connects Kratzer's theory to a more general version of Lewis' theory, namely, Pollock's theory (see below).

5. John Pollock. *Subjunctive Reasoning*. 1976.

    - Another related theory of counterfactuals, and other subjunctive conditionals. Also discusses subjunctive generalizations.