CS112 (Fall 2018)
Counterfactual Logic II
October 12, 2018

1 Counterfactuals in a dynamic context

As previously mentioned, Lewis is not fond of the meaning of counterfactuals depending on context. At that time, not many people studied aspects of meaning beyond the sentential/truth-conditional level, such that appealing to context was to toss the problem into a "wastebasket" of vagueness. Fortunately, the situation has improved since then.

In *dynamic semantics* (not to be confused with *dynamic logic*, which we will talk about later in the course), the meaning of a sentence is related to how it changes the context of conversation. The "context", in this case, includes the model (a set of possible worlds, and an accessibility relation) currently under consideration. von Fintel's idea is that counterfactuals can change the context even before they are evaluated.

We are all familiar with the following sequence of counterfactuals (called a *Lewis-Sobel sequence*):

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

This sequence makes sense. The speaker is asserting a counterfactual conditional, but then denying the related counterfactual with a strengthened antecedent. In the closest worlds where kangaroos have no tails, they topple over, but in those where they also use crutches, they do not topple over. Now consider the reverse sequence (a *reverse Sobel sequence*):

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

This sequence is semantically infelicitous. Somehow, once you consider worlds where kangaroos have no tails and use crutches, you cannot then ignore the crutches-worlds when evaluating the second sentence.

The solution, according to von Fintel, is that instead of having a static system of spheres (or accessibility relations), there is a single accessibility relation (or *modal horizon*) that is subject to change throughout a discourse. A counterfactual sentence then potentially changes the modal horizon, and is interpreted relative to it.

von Fintel's theory builds off of earlier work by Warmbrod and Heim. They argue for a strict conditional interpretation of counterfactuals, according to a contextually supplied accessibility relation where some antecedent-worlds are accessible from the actual world. They do not specify how exactly the accessibility relation is updated when no antecedent-world is accessible; Warmbrod requires that the accessibility relation is "normal", yet constant within a discourse.

According to von Fintel's theory, the modal horizon begins as a trivial accessibility relation, in which every world is only accessible to itself (von Fintel refers to an accessibility "function" from worlds to sets of worlds, but it is easy to convert it into a relation between worlds). Then, if a conditional sentence is uttered and there are no accessible worlds where the antecedent is true, the modal horizon will expand just enough to include such a world, where the expansion proceeds according to comparative similarity of worlds. Formally:

a. Auxiliary notion: Update of f by a sentence ϕ
   For any sentence ϕ, any accessibility function f, and similarity relation $\leq$:
   $f^{\phi,\leq} = \lambda w.f(w) \cup \{w' : \forall w'' \in [[\phi]]^{f,\leq} : w' \leq_w w''\}$

   - $\lambda w.f(w)$ is the set of worlds already in $f(w)$. Then, add those worlds w' that are at least as similar to w as all worlds w'' where ϕ is true. These worlds will consist of the most similar ϕ-world, and all worlds more similar than it.

b. Context change potential of counterfactuals
   $f|\phi > \psi|^{\leq} = f^{\phi,\leq}|\phi|^{\leq}|\psi|^{\leq}$

   - To update the context according to a counterfactual $\phi > \psi$, update according to ϕ, then update according to ψ.

c. Truth-conditions
   $[[\phi > \psi]]^{f,\leq}(w) = 1$ iff
   $\forall w' \in f^{\phi,\leq}|\phi|^{\leq}(w) : [[\phi]]^{f,\leq}(w') = 1 \rightarrow [[\psi]]^{f^{\phi,\leq}|\phi|^{\leq},\leq}(w') = 1$

   - First, the antecedent updates the context: this is $f^{\phi,\leq}|\phi|^{\leq}(w)$. Then $\phi > \psi$ is true iff for all worlds w' in the updated context, if ϕ is true, then ψ is true. In other words, we are back to a strict conditional, this time using a contextually determined modal horizon as the accessibility relation.

We conclude by considering an example we saw last class:

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.

We saw that the first two premises are true, but the conclusion is false. However, according to von Fintel's system of dynamic entailment, we should not have accepted the second premise as true. Uttering the first sentence changed the context to include those worlds where Hoover were both Russian and communist. In those worlds, he would not be a traitor. But nobody complained!

One solution is that for logical arguments, we commit ourselves to a stable context. von Fintel does this as follows:

Strawson entailment
$\phi_1, ...\phi_n \models_{\text{Strawson}} \psi$ iff for all contexts c such that $c = c|\phi_1|...|\phi_n||\psi|$,
it holds that $[[\phi_1]]^c \cap ... \cap [[\phi_n]]^{c|\phi_1|...|\phi_{n-1}|} \subseteq [[\psi]]^{c|\phi_1|...|\phi_n|}$.

The problem with this definition is that it requires c to be wide enough to accommodate every sentence in the argument. So the context of the second sentence still includes worlds where Hoover were Russian and communist, but not a traitor, and thus our problem remains. Food for thought: is it possible to define Strawson entailment in such a way as to validate our example?

2 Causal and interventionist counterfactuals

In the meantime, we will turn our attention to a method of specifying exactly what it means for a world to be "more similar" than another world to the actual world. For this discussion, we will move over to the related notion of "minimal change", as in the Ramsey test (or in Pollock's theory of subjunctive conditionals).

Recall Kratzer's example of looking in a mirror. In the actual world, she is not looking in a mirror (call this fact p), and does not see her reflection (call this fact q). If she were to look in the mirror, would she see her reflection? The problem is that one can argue that worlds in which she still does not see her reflection are "more similar" to the actual world than worlds in which she does: in the first set of worlds, only one fact has changed (p), while in the second, two facts have changed (p and q).

As we previously saw, Kratzer's solution is to "lump" the two facts into a single fact $p \cap q$, such that p and q stand and fall together. Equivalently, if we let P and Q be Boolean variables for whether p and q respectively are true or false,

then we can represent this fact as the *structural equation* $P = Q$. Examples of structural equations, also called *causal laws*, include laws of nature and actual necessities (i.e., the subjunctive generalizations of Pollock).

Parallel to the work being done by philosophers and linguists on modal theories of counterfactuals, computer scientists have also been working on counterfactuals, mostly using a causal modeling approach, as exemplified by Galles and Pearl. They define a *causal model* as follows:

A causal model is a triple $M = \langle U, V, F \rangle$ where

(i) U is a set of variables, called *exogenous*, that are determined by factors outside the model.

(ii) V is a set $\{V_1, V_2, ..., V_n\}$ of variables, called *endogenous*, that are determined by variables in the model.

(iii) F is a set of functions $\{f_1, f_2, ..., f_n\}$ where each f_i is a mapping from $U \cup (V \backslash V_i)$ to V_i such that F defines a mapping from U to V. (i.e., F has a unique solution for each state u in the domain of U). Symbolically, F can be represented by writing

$$v_i = f_i(pa_i, u), \ \ i = 1, ..., n$$

where pa_i is any realization of the (unique) set of variables PA_i in V/V_i (connoting parents) that renders f_i nontrivial.

- In other words, F is a set of structural equations, one for each endogenous variable, where no variable appears on both sides of an equation.

Each causal model has an associated *causal graph*, in which the nodes are variables and for each endogenous variable V_i, the directed edges point from the parents PA_i to V_i. A causal model is *recursive* iff its causal graph has no cycles.

Let $Y_x(u)$ denote the value of a endogenous variable Y in a causal (sub)model where the structural equation for X is replaced with $X = x$, assuming constant exogenous variable values u. Then this expression encodes a counterfactual statement "If it were the case that $X = x$, then it would be the case that $Y = Y_x(u)$".

(As a brief aside, recall that Stalnaker considered—and rejected—a theory of conditionals based on a causal "connection" between the antecedent ϕ and the consequent ψ. Roughly speaking, this is equivalent to looking for a structural equation of the form $\psi = f(\phi)$ in the original causal model. In Galles and Pearl's theory, by contrast, an *intervention* is first applied to the causal model to set the value of ϕ, and then ψ is evaluated in the updated submodel.)

We note that Galles and Pearl's theory of counterfactuals is a more specific version of Lewis' theory. For the translation, we take an intervention of the form $X = x$ to be equivalent to considering the closest possible worlds where $X = x$. First, we note that two properties of causal counterfactuals for recursive models, *composition* and *effectiveness*, are sufficient to derive the axioms of Lewis' language, and vice versa:

- Composition: For any two singleton variables Y and W, and any set of variables X in a causal model, we have

$$W_x(u) = w \Longrightarrow Y_{xw}(u) = Y_x(u)$$

  - If we intervene to set $X = x$, and as a result $W = w$, then any additional intervention to set $W = w$ will have no effect on any other variable Y. As a corollary, if we intervene to set $W = w$, but $W = w$ in the actual world, then the value of Y will also remain as it is in the actual world. This means that if $A \wedge B$, then $A \mathrel{\square\!\!\rightarrow} B$, and if $A \mathrel{\square\!\!\rightarrow} B$, then $A \rightarrow B$. *Modus ponens* completes the circle: if A and $A \rightarrow B$, then B (and therefore $A \wedge B$).

- Effectiveness: For all variables X and W, $X_{xw}(u) = x$.

  - If we intervene to set $X = x$, then $X = x$, no matter what we do to W. This is equivalent to the axiom $A \mathrel{\square\!\!\rightarrow} A$—in all (closest) worlds where A is true, A is true.

However, for nonrecursive models, an additional property of Galles and Pearl, *reversibility*, has no counterpart in Lewis' theory. Reversibility states that if we set $Y = y$ and get $W = w$ as a result, and if we set $W = w$ and get $Y = y$ as a result, then $W = w$ and $Y = y$ in the original model. This is not the case in Lewis' theory, where w may be true in the closest y-worlds, and y may be true in the closest w-worlds, but neither w nor y are required to be true in the actual world.

Another limitation of Galles and Pearl's theory is that counterfactuals are restricted to the form $\mathbf{A} \rightarrow \mathbf{C}$, where $\mathbf{A}$ and $\mathbf{C}$ are conjunctions of literals. However, Lewis' theory allows us to talk about counterfactuals with arbitrary antecedents and consequents. Briggs presents an extension of Galles and Pearl's theory that allows for Boolean (non-counterfactual) antecedents and arbitrary consequents, with surprising consequences.

In Briggs' theory, if we allow the consequents of counterfactual sentences to themselves be counterfactuals, *modus ponens* is no longer valid for the counterfactual operator $\mathrel{\square\!\!\rightarrow}$. Consider the following example. In the actual world, an assassin pours poison into a king's cup of coffee. The king's bodyguard then pours an antidote into the coffee. The king drinks the coffee and survives.

Suppose the following causal laws hold. The bodyguard's action depends on what the assassin does: if the assassin poisons the coffee, the bodyguard would

apply the antidote, but if the assassin does nothing, so would the bodyguard (since there would be no need to apply the antidote in this case). Furthermore, suppose that both the poison and the antidote are deadly when taken by themselves; it is only in combination that the antidote neutralizes the poison. Let A be the variable that states whether the assassin poisons the coffee, B represent whether the bodyguard applies the antidote, and K represent whether the king survives. These causal laws can be represented by the following structural equations:

$$A = 1$$

$$B = A$$

$$K = A \leftrightarrow B$$

(Briggs also allows structural equations for exogenous variables such as A; the equation $A = 1$ states that in the actual world, the assassin poisons the coffee.)

Now consider the counterfactual sentence "If the bodyguard had applied the antidote, then if the assassin had not poisoned the coffee, the king would not have survived", or $B \mathrel{\square\!\!\rightarrow} (\neg A \mathrel{\square\!\!\rightarrow} \neg K)$. We can see that the entire sentence is true, because if the coffee contained antidote but not poison, the king would not have survived. Furthermore, the antecedent B is true (in the actual world).

However, the consequent $\neg A \mathrel{\square\!\!\rightarrow} \neg K$ is not true. Under a causal modeling semantics, this sentence applies an intervention that sets $A = 0$. To see the effect on K, we note that $B = A = 0$, and therefore $K = A \leftrightarrow B = 0 \leftrightarrow 0 = 1$. If we look at the entire sentence, however, we had previously intervened to set $B = 1$. This removes the causal dependence of B on A. As such, when we propagate $A = 0$ through the set of equations, it remains the case that $B = 1$, and therefore $K = A \leftrightarrow B = 0 \leftrightarrow 1 = 0$.

In Lewis' theory, this is not a problem for *modus ponens*, because of the definition of the counterfactual operator: $B \mathrel{\square\!\!\rightarrow} (\neg A \mathrel{\square\!\!\rightarrow} \neg K)$ is true iff there is some sphere of accessibility where all B-worlds are $(\neg A \mathrel{\square\!\!\rightarrow} \neg K)$-worlds. Because the actual world is a B-world, all spheres that contain B-worlds contain the actual world. But the actual world is not a $(\neg A \mathrel{\square\!\!\rightarrow} \neg K)$-world, at least not according to a similarity measure that ranks $(\neg A \wedge \neg B)$-worlds as more similar than $(\neg A \wedge B)$-worlds. So the entire counterfactual, by definition, is not true. In order for the entire sentence to be true under Lewis' semantics, there would have to be a sphere that includes a B-world (that is also a $(\neg A \mathrel{\square\!\!\rightarrow} \neg K)$-world), that does not include the actual world, a violation of *weak centering*.

Finally, we consider the following sentences:

If the assassin had not poisoned the coffee, then the king would have survived.
($\neg A \mathrel{\Box\!\!\rightarrow} S$)

If the assassin had not poisoned the coffee and the bodyguard had not applied the antidote, or if the assassin had not poisoned the coffee and the bodyguard had applied the antidote, then the king would have survived.
($((\neg A \wedge \neg B) \vee (\neg A \wedge B)) \mathrel{\Box\!\!\rightarrow} S$)

The antecedents of the two sentences are equivalent in propositional logic. But the two sentences have different truth values. The first sentence is true: if the assassin had not poisoned the coffee, then the bodyguard would not have applied the antidote, and the king would have survived. However, the second sentence is false: in the second scenario ($\neg A \wedge B$), the king would not have survived.

The above reasoning relied on the rule of *simplification of disjunctive antecedents (SDA)*: if $(A \vee B) \rightarrow C$ is true, then $A \rightarrow C$ and $B \rightarrow C$ are both true. This rule is valid for the material conditional $\rightarrow$, but not for Lewis' counterfactual conditional: if C is true in the nearest $(A \vee B)$-world, it is not necessarily true in the nearest A-world (because the aforementioned world could have been a B-world). Intuitively, however, SDA makes sense: when evaluating the truth of the second sentence, we consider each scenario and determine the truth of the consequent in each. This is evidence that equivalent formulas in propositional logic can not necessarily be substituted in antecedents of counterfactuals.

3 References and further reading

Primary references:

1. Rachael Briggs. Interventionist counterfactuals. *Philosophical Studies*, 160(1):139–166, 2012.

2. Kai von Fintel. Counterfactuals in a dynamic context. *Current Studies in Linguistics Series*, 36:123–152, 2001.

3. David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182, 1998.

Further reading:

4. Ivano Ciardelli, Linmin Zhang, and Lucas Champollion. Two switches in the theory of counterfactuals. *Linguistics and Philosophy*, pages 1–45, 2018.

    - More evidence, via crowdsourcing, that propositional equivalents are not equivalent in antecedents of counterfactuals. Also introduces "background semantics", which distinguishes background facts (which must stay fixed when evaluating the truth of a counterfactual) and foreground facts (which may vary).

5. Sarah Moss. On the pragmatics of counterfactuals. *Noûs*, 46(3):561–586, 2012.

    - Evidence that the context-dependency of the modal horizon is not specific to counterfactuals, but can be explained via a more general pragmatic principle.

6. Paolo Santorio. Filtering semantics for counterfactuals: Bridging causal models and premise semantics. In *Semantics and Linguistic Theory*, volume 24, pages 494–513, 2014.

    - Considers a nonrecursive causal model, and what it means for counterfactuals in natural language.