Filtering Semantics for Counterfactuals

Paolo Santorio
University of Leeds

Synopsis. I argue that ordering/premise semantics for counterfactuals in the style of Lewis, Kratzer, and others validates inference patterns that are disconfirmed in natural language. The objection extends to all semantics employing a fixed ordering of worlds or equivalent algebras. The solution is to let the ordering shift (in a systematic way) on the basis of the antecedent. The proposed implementation starts from standard premise semantics and adds a new 'filtering' operation on the premise set. The resulting semantics is interestingly related to the semantics for counterfactuals emerging from Judea Pearl's causal models framework in computer science: filtering is a possible worlds counterpart of Pearl's interventions; my data reveals a gap in predictions between classical Lewis/Kratzer semantics and Pearl's account.

The puzzle. Consider the following scenario.

> *Love triangle.* Andy, Billy, and Charlie are in a love triangle. Billy is pursuing Andy; Charlie is pursuing Billy; and Andy is pursuing Charlie. Each of them is very annoyed by their suitor and wants to avoid them.
>
> Suppose that there is a party going on at the moment. All of them were invited. None of them went, but each of them kept appraised of the others' decisions. An occasion to spend time with the person they liked, and without their suitor being there, would have been sufficient for them to go.

The scenario gives rise to the following judgments:

(1) If Andy was at the party, Billy would be at the party. ✓

(2) If Andy was at the party, Charlie would be at the party. ✗

Generalizing ('A', 'B' and 'C' stand for the propositions that Andy, Billy, and Charlie are at the party):

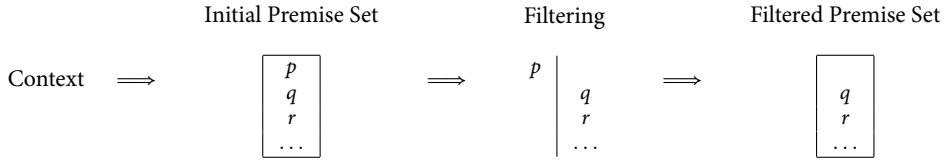
$$\begin{array}{llll} A \boxdot\!\!\rightarrow B & \checkmark & A \boxdot\!\!\rightarrow C & \times \\ B \boxdot\!\!\rightarrow C & \checkmark & B \boxdot\!\!\rightarrow A & \times \\ C \boxdot\!\!\rightarrow A & \checkmark & C \boxdot\!\!\rightarrow B & \times \end{array}$$

But this pattern cannot be vindicated by any version of ordering/premise semantics. All versions of the latter validate the following rule (which is mentioned and studied by Kraus et al. (1990)):

Loop $\quad A \boxdot\!\!\rightarrow B, B \boxdot\!\!\rightarrow C, C \boxdot\!\!\rightarrow A \vdash A \boxdot\!\!\rightarrow C$

Loop is essentially a byproduct of the fact that the comparative closeness relation $\leq_w$ is transitive. (Transitivity is essential to $\leq_w$ qualifying as an ordering.)

Account: filtering semantics. My solution starts from a Kratzer-style (1981a, 1981b) premise semantics and introduces a new operation that 'filters out' elements from the premise set:



Different antecedents filter out different premises. Hence, within the same context, counterfactuals with different antecedents are evaluated with respect to different premise sets. This is equivalent to antecedent-induced shifts in the ordering.

For illustration, consider *Love triangle*. I assume that the key premises state how one person's going to the party depends on other people going. For example (simplifying in several ways; more below):

(i) {*w*: A is at the party iff (C is at the party and B isn't at the party) in *w*} A ↔(C∧¬B)
(ii) {*w*: B is at the party iff (A is at the party and C isn't at the party) in *w*} B ↔(A∧¬C)
(iii) {*w*: C is at the party iff (B is at the party and A isn't at the party) in *w*} C ↔(B∧¬A)

Given how filtering works, different antecedents filter out different premises among (i)–(iii). E.g., (1) and (2) filter out (i). Counterfactuals with antecedents B and C filter out, respectively, (ii) and (iii). Hence counterfactuals in the Loop-invalidating sextet filter out different premises and are evaluated with respect to different premise sets. As a result, the six judgments are all vindicated.

Implementation. The semantics exploits two innovations. (1) The elements of premise sets are not propositions. They are rather pairs of a partition (construed as sets of mutually incompatible and jointly exhaustive propositions; for short, *answers*) and a proposition. E.g., the full form of (i) is:

(i) $\langle\{A, \neg A\}, A \leftrightarrow(C\wedge\neg B)\rangle$

(2) Filtering consists in removing premises from the (premise set generated by the) ordering source. (Informally: we take the smallest set(s) of answers, among all those appearing in premises in the ordering source, that entails the antecedent; we remove the corresponding premises from the ordering source.) This is encoded directly in the semantics for counterfactuals. Using '$[g|A]$' to mean that the ordering source g is filtered for antecedent A, here is a first pass:

$$[\![\Box\ [\text{if } p]\ [q]]\!]^{f,g} = [\![\Box\ q]\!]^{f\cup\{p\},[g|p]}$$

A second pass: there may be more than one way to perform filtering. (I.e. there may be more than one minimal set of answers that entails the antecedent.) For example, consider the disjunctive antecedent 'A ∨ B' in *Love triangle*. To accommodate this, we quantify over filterings:

$$[\![\Box\ [\text{if } p]\ [q]]\!]^{f,g} = \forall [g|p] \text{ s.t. } [g|p] \text{ is a filtering of } g \text{ for } p,\ [\![\Box\ q]\!]^{f\cup\{p\},[g|p]}$$

The analogy with causal models. A number of researchers in computer science (above all, Pearl (2000)) have developed a formal framework for modeling causation and causal reasoning. Pearl's framework involves a semantics for counterfactuals (in artificial languages) that is claimed to yield a logic equivalent to Lewis's (e.g. by Galles & Pearl (1998)). There are two connections between the current project and causal models literature. (1) Scenarios like *Love triangle* are counterexamples to the equivalence claim: Pearl's framework does indeed predict the failure of Loop (this claim is backed by a very recent result in Halpern (2013)). (2) *Contra* Pearl's own early claims and later literature (e.g. Kaufmann (2013)), I claim that just filtering semantics is the right way to implement Pearl's system into a compositional possible worlds semantics. In particular, filtering works as the counterpart of Pearl's intervention operation in a premise/ordering semantics.

References

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

Halpern, Joseph Y. (2013). "From Causal Models to Counterfactual Structures." *Review of Symbolic Logic*, 6(2): pp. 305–322.

Kaufmann, Stefan (2013). "Causal Premise Semantics." *Cognitive science*, 37(6): pp. 1136–1170.

Kratzer, Angelika (1981a). "The Notional Category of Modality." In H. J. Eikmeyer, and H. Rieser (eds.) *Words, Worlds, and Contexts: New Approaches to Word Semantics*, Berlin: de Gruyter.

Kratzer, Angelika (1981b). "Partition and Revision: The Semantics of Counterfactuals." *Journal of Philosophical Logic*, 10(2): pp. 201–216.

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor (1990). "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics." *Artificial intelligence*, 44(1): pp. 167–207.

Pearl, Judea (2000). *Causality: models, reasoning and inference*. Cambridge University Press.