Filtering semantics for counterfactuals: bridging premise semantics with causal models

Paolo Santorio
paolosantorio@gmail.com
University of Leeds

Synopsis

I give a **puzzle for counterfactual semantics**.

- Standard premise semantics validates an inference pattern that fails in natural language.
- A connection with causal models (Pearl 2000): just this pattern marks the difference between the logics generated by premise/ordering semantics and the logics generated by causal models. (Halpern 2013)
- The proposed solution: a causal-models inspired semantics, which implements a new 'filtering' operation on the premise set.

The puzzle: Loop

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 is 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 their loved one and without their suitor would have been sufficient reason for them to go.

(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, we have this pattern:

$A \boxrightarrow B$ ✓ $\quad A \boxrightarrow C$ ✗
$B \boxrightarrow C$ ✓ $\quad B \boxrightarrow A$ ✗
$C \boxrightarrow A$ ✓ $\quad C \boxrightarrow B$ ✗

The problem: this pattern cannot be vindicated by any version of premise semantics. All premise semantics validate (Kraus et al. 1990):

LOOP
$$\frac{p \boxrightarrow q \quad q \boxrightarrow r \quad r \boxrightarrow p}{p \boxrightarrow r}$$

Proof for Stalnaker's (1968) semantics: the ordering relation $<_w$ it total, hence there is a closest world to w that is a p-world, a q-world, or an r-world. Call this world w^*. Without loss of generality, suppose w^* is a p-world. Since $p \boxrightarrow q$, w^* is a q-world. Since $q \boxrightarrow r$, and since w^* is the closest q-world, w^* is an r-world. Since the closest p-world is also an r-world, $p \boxrightarrow r$ is true. *QED.*

Counterfactual reasoning, Pearl-style

The causal models framework: a formal framework for modeling causation and causal reasoning (Pearl 2000). Causal dependencies are captured via 'directional' equations, and represented visually in a directed graph.

Example: a mechanical system where cog's 1 turning causes cog 2 to turn, which causes cog 3 to turn.

Equations:
C2 = C1
C3 = C2

cog 1 turning ⟶ **cog 2 turning** ⟶ **cog 3 turning**

Counterfactuals. Counterfactual reasoning involves **replacing equations** (and erasing arrows in the graph).

(3) If cog 2 had not turned, cog 3 would not have turned.

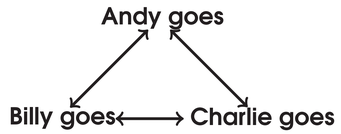
Equations:
~~C2 = C1~~
C2 = 0
C3 = C2

cog 1 turning **cog 2 turning** ⟶ **cog 3 turning**
0 0

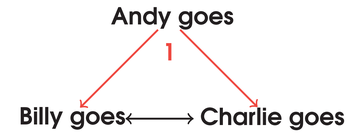
This amounts to **removing previous information about causal dependencies** and replacing it with new information. This replacement operation is the basis of the departure from standard counterfactual logics.

Below is a model for the love triangle example. Notice that the verdicts match the judgments in the data:

Equations:
A = (C ∧¬B)
B = (A ∧¬C)
C = (B ∧¬A)



Equations:
~~A = (C ∧¬B)~~
A = 1
B = (A ∧¬C)
C = (B ∧¬A)



Implementation: Filtering Semantics

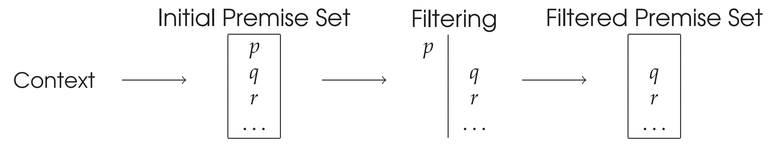
Standard counterfactual semantics (Lewis 1973, Kratzer 1981a, 1981b, 1986): we check whether **all maximal consistent sets** generated by adding ordering source premises to the antecedent entail the consequent.

$[\![p \boxrightarrow q]\!]^{w,f,g} = 1$ iff, for every maximal consistent superset S of $\{p\}$ relative to $g(w)$, $S \vDash [\![q]\!]^{w,f,g}$

Filtering semantics (first pass): we check whether **the set resulting from adding the antecedent to the ordering source and removing some relevant premise** entails the consequent.

$[\![p \boxrightarrow q]\!]^{w,f,g} = 1$ iff $\{p\} \cup$ ($g(w)$ filtered for p) entails q

The novelty: the **filtering operation**, which selectively removes elements from the premise set.



To determine what is filtered out, premises encode more information (i.e. info about **what determines what**).

- We model this by taking premises to be pairs of a (Groenendijk & Stokhof) question and a proposition.
- First pass: a premise is filtered out by p just in case p settles the answer to the question.

Initial premise set:
⟨{A, ¬A}, A ↔ (C ∧¬B)⟩
⟨{B, ¬B}, B ↔ (A ∧¬C)⟩
⟨{C, ¬C}, C ↔ (B ∧¬A)⟩

⟶

Premise set after filtering for A:
~~⟨{A, ¬A}, A ↔ (C ∧¬B)⟩~~ ⟨{A, ¬A} A⟩
⟨{B, ¬B}, B ↔ (A ∧¬C)⟩
⟨{C, ¬C}, C ↔ (B ∧¬A)⟩

In terms of ordering semantics, this is equivalent to an **antecedent-induced ordering shift**.

More detail: many filterings

A problem: an antecedent may be made true by settling different combinations of answers.

(4) If Andy or Billy was at the party, ...

The way to capture this: we check **all (minimal) ways to settle questions in the premise set that make the antecedent true**.

Initial premise set:
⟨{A, ¬A}, A ↔ (C ∧¬B)⟩
⟨{B, ¬B}, B ↔ (A ∧¬C)⟩
⟨{C, ¬C}, C ↔ (B ∧¬A)⟩

Permissible filterings for A∨B:
~~⟨{A, ¬A}, A ↔ (C ∧¬B)⟩~~
⟨{A, ¬A} A⟩
...
~~⟨{B, ¬B}, B ↔ (A ∧¬C)⟩~~
⟨{B, ¬B} B⟩
...

Filtering semantics (second pass): we check whether **all sets resulting from adding the antecedent to the ordering source and removing some other relevant premise** entail the consequent.

$[\![p \boxrightarrow q]\!]^{w,f,g} = 1$ iff, $\forall S$: S is a permissible filtering of $\{p\} \cup g(w)$ for p, S entails q

Extras: disjunctive antecedents

A seemingly valid pattern (Fine 1975):

SIMPLIFICATION
$$\frac{p \vee q \boxrightarrow r}{p \boxrightarrow r,\ q \boxrightarrow r}$$

(5) If Mary or Sue came, the party would be fun.
⟹ If Mary came, the party would be fun.
⟹ If Sue came, the party would be fun.

SIMPLIFICATION is invalid on standard semantics. But, on a suitable choice of premises, we validate:

RESTRICTED SIMPLIFICATION
$$\frac{p \vee q \boxrightarrow r}{p \boxrightarrow r,\ q \boxrightarrow r}$$

Provided $p \vee q \not\vDash p$ and $p \vee q \not\vDash q$

On the agenda:

- Formulating a (hyperintensional) version of the semantics that fully validates **SIMPLIFICATION**.
- Exploring connections with related scalarity phenomena (Santorio 2014).

References

Fine (1975) *Review of Lewis' "Counterfactuals"*; Halpern (2013) *From Causal Models to Counterfactual Structures*; Kaufmann (2013) *Causal Premise Semantics*; Kratzer (1981a) *The Notional Category of Modality*, (1981b) *Partition and Revision: The Semantics of Counterfactuals*; (1986) *Conditionals*; Kraus, Lehmann, and Magidor (1990), *Nonmonotonic Reasoning, Preferential Models and Cumulative Logics*; Lewis (1973) *Counterfactuals*; Pearl (2000), *Causality*; Santorio (2014), *Exhaustified Counterfactuals*; Stalnaker (1968) *A Theory of Conditionals.*