Informative Counterfactuals

Adam Bjorndahl & Todd Snider
Cornell University

PHLINC2
February 14-15, 2014

- We use counterfactuals all the time:

(1) If Alice had gone to the party, Bob would have stayed home.

(2) If the movie had been any good, I wouldn't have fallen asleep.

(3) Even if there hadn't been traffic, we still would have been late.

- We can use them to talk about things we know to be false or things we're uncertain about
- (1) usually means that Alice didn't go to the party and that Bob did.
- It also communicates some connection between the two events.

- There are different ways for the events in (1) to be connected; different ways for this counterfactual to be informative.

(1) If Alice had gone to the party, Bob would have stayed home.

- Does Bob try to avoid Alice?
  - Maybe he's shy.
  - Maybe he doesn't like her.
  - Maybe he doesn't like her perfume.
- Do other circumstances prevent them from attending parties together?
  - Maybe they're a couple on a tight budget.
  - Maybe Bob is actually Alice in disguise.
- Does Alice try to avoid Bob?
  - Unlike the other scenarios, this one does not seem to jive with (1)...

- Consider a world where Alice and Bob are married, and live with their young son Doug

(1) If Alice had gone to the party, Bob would have stayed home.

(4) If Alice had gone to the party, Doug would have been home alone.

- (1) and (4) are each felicitous individually
- A felicitous utterance of one precludes a felicitous utterance of the other
- Any account of how we update our knowledge with counterfactuals should explain this

- There have been two main approaches to accounting for counterfactuals
- The classical approach ascribes structure between worlds in the form of a similarity relation
- The structured possible world approach ascribes structure within worlds
- We'll be using the latter
  - As we'll see, this allows us to represent distinct interpretations of a given counterfactual, what we call *explanatory strategies*
  - It also provides a principled account of the incompatibility between (1) and (4)

- Most of the counterfactuals literature focuses on defining *truth conditions*
  - Lewis 1973, 1979a,b; Tichý 1976; Kratzer 1989; Pearl 2000; Hiddleston 2005; Kment 2006
- We focus instead on *informativity*:
  - On hearing a counterfactual, how do we update our knowledge/beliefs with it?

Outline

1 Overview

2 Some preliminaries
- Informativity
- The framework

3 Our proposal
- Understanding a counterfactual
  - Three explanatory strategies
- Integrating a counterfactual with our knowledge

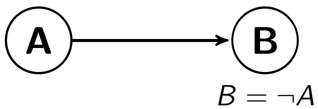
4 Conclusion

What does it mean to be informative?

- An assertion is informative if it excludes some but not all worlds in the context set
  - Gives us a smaller (but non-empty) set of candidate worlds
- If worlds are sets of events, their truth values, and **dependencies** among events, then we can use these dependencies to partition worlds
  - We don't need to gain information that is counter to fact
  - We can retain knowledge about the factual state of events
  - We can learn about the ways in which events are related
- Asserting the existence of a specific dependency excludes worlds without that dependency

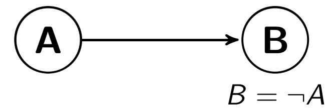
Structural Equation Modeling (SEM)

- As far back as Wright 1921, but formalized in Pearl 2000
- Allows for the modeling not only of variables but also dependencies
- Models consist of:
  - Nodes   Circles   Variables/Events
  - Edges   Arrows   Dependencies
    - Labeled with equations



Structural Equation Modeling (SEM)

- For convenience and simplicity, our examples are
  - Two-valued
  - Deterministic
- This framework and analysis also handles multi-valued and/or probabilistic systems



What do counterfactuals do?

- They assert some degree of covariance between the antecedent and consequent
  - Not necessarily perfect covariance
- They implicate a direct (causal) dependence of consequent on antecedent ($C = A$)
- This implicature can be canceled (or strengthened):

(5) If I push this button then the rocket will launch.

(6) If I push this button then the rocket will launch, but my pushing this button doesn't directly cause the rocket to launch.

(7) If I push this button then the rocket will launch, and my pushing this button directly causes the rocket to launch.

- This implicated direct dependency is enough to make many counterfactuals informative
  - (5) excludes worlds where the button and launch never covary
  - If the implicature isn't canceled, the hearer updates with this simple direct dependency
    - We'll return to how this update works in a bit
- For some counterfactuals this direct dependency is problematic

Rejecting explanations

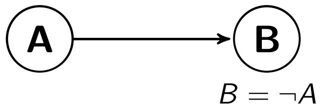
- Many reasons to reject an explanation (including the implicated simple dependency)
  - It might contradict prior knowledge
  - It might violate a law of good explanations
    - e.g. by positing an effect that is temporally prior to its cause
  - It might not satisfy the contextual parameter for specificity

- Any of these reasons might make us reject the simple direct dependency of the consequent on the antecedent
  - In other words, we reject the $C = A$ edge
- But the counterfactual stipulates some covariance
- Trying to maintain the cooperativity of the speaker's contribution, we search for an explanation to make the counterfactual true
- Three possible ways to deal with this problematic dependence:
  - Positing an Additional cause
  - Positing a Common cause
  - Positing an Intermediate cause
- Call these **explanatory strategies**

- To understand these explanatory strategies, it will be helpful to have an example:

(1) If Alice had gone to the party, Bob would have stayed home.

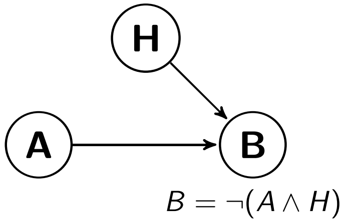
- The implicated simple dependency of (1) is captured in this model



- This model is unsatisfying
- Alice's attendance doesn't literally cause Bob to be elsewhere
- What's missing is an **explanation**

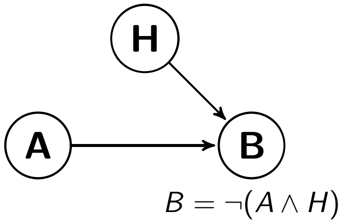
Additional cause

- The hearer might suppose that the consequent is dependent not solely on the antecedent but also on some additional cause
- For example, a common interpretation of (1) might lead one to believe that Bob hates Alice
- We can consider Bob's hatred of Alice as an additional node in our model



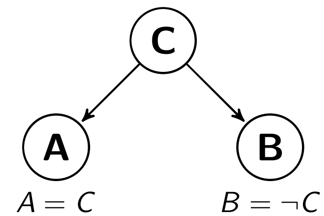
Additional cause

- The dependence of B on A is still present, but it's been modified
  - The $B = \neg A$ edge is no longer part of the model
- The antecedent and consequent covary only in the right H-conditions



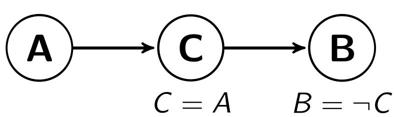
Common cause

- The hearer might suppose that the consequent isn't dependent upon the antecedent at all
- Instead, both antecedent and consequent depend on some common cause
- They still covary, but have no interdependence
- For example, imagine that Alice & Bob flip a coin to determine who attends



Intermediate cause

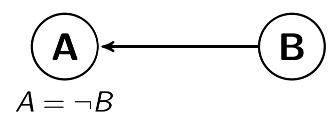
- The hearer might suppose that the consequent depends on the antecedent only by means of some intermediate cause
- The antecedent and consequent still covary, but without positing a direct causal dependency
- For example, imagine that Alice brings her cat wherever she goes, and Bob is deathly allergic to cats



A fourth explanatory strategy?

- Reversing the simple causal relationship also allows the antecedent and consequent to covary

(1) If Alice had gone to the party, Bob would have stayed home.

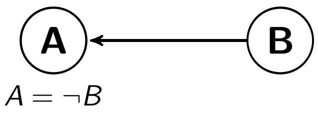


- This classical *backtracker* has the consequent as the cause
- This model is rejected as an interpretation of (1)
- It can be licensed by a double-auxiliary construction, as in (8)

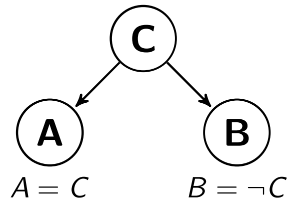
(8) If Alice had gone to the party, Bob would have had to have stayed home.

A note on *backtracking*

- Two different things referred to as *backtracking*



- Reversing causal direction
- Classic philosophy literature
- Needs double-aux licensing



- 'Upstream' reasoning
- Recent psychology literature
- Doesn't need licensing

How do we update with what we've learned?

- Once an acceptable explanation is found, we have to integrate it with our extant body of knowledge
- With structured possible worlds, our knowledge must include not just facts about variables but also dependencies
- We can model our knowledge as one persistent SEM
- Integrating an informative counterfactual is consolidating a new explanatory SEM with the persistent one

- When consolidating, we integrate dependencies, not variable values
- Counterfactuals *can* inform us about actual values via presupposition
- Okay to accommodate these actual world facts
  - Accommodation in the Stalnaker 1974 sense
  - This can be done prior to explanation
- We don't want to update with Alice's counterfactual attendance

- While not yet formalized, there are at least two operations required for consolidation

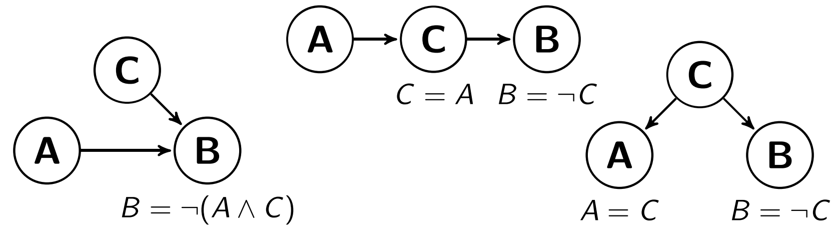
1. Addition
   - For extending the graph
   - Possibly add new nodes
   - Add new dependencies among nodes
2. Explosion
   - For looking deeper into the internal mechanism of a single node
   - Explode one node into multiple nodes
   - Retains incoming/outgoing dependencies of the original node

- At least these two operations, possibly others
- After consolidation, deduce values of new nodes, if necessary

- This consolidation process gives us insight into interactions between counterfactuals
  - (1) If Alice had gone to the party, Bob would have stayed home.
  - (4) If Alice had gone to the party, Doug would have been home alone.
- Updating with (1) adds a covariance between A and $\neg B$ to our knowledge base
  - Alice and Bob have opposite party-attendance values
- Updating with (4) requires that A and B have the same value
- Consolidating either (1) or (4) with one's persistent SEM makes the other contradictory

Conclusion

- We can use structured possible worlds to model dependencies, not just facts
- We propose using them to model informative counterfactuals
- Doing so gets us a natural way to represent the three explanatory strategies



Conclusion

- Our analysis also neatly captures the distinction between different senses of *backtracking*
  - Classical philosophical backtrackers reverse the generally implicated direction of dependence
  - Recent psychological uses of the term refer to explanations including at least one instance of COMMON CAUSE
- It accounts for mutually infelicitous counterfactuals
  - Each updates our internal SEM in a way that precludes the other

Eric Hiddleston. A causal theory of counterfactuals. *Noûs*, 39(4):632–657, 2005.

Boris Kment. Counterfactuals and explanation. *Mind*, 115(458):261–310, 2006.

Angelika Kratzer. An investigation of the lumps of thought. *Linguistics and Philosophy*, 12(5):607–653, 1989.

David K. Lewis. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2(4):418–446, 1973.

David K. Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13:455–476, 1979a.

David K. Lewis. Conditionals: Ordering semantics & Kratzer's semantics. Unpublished manuscript, Princeton University., 1979b.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge Univ Press, 2000.

Robert Stalnaker. Pragmatic presuppositions. In Peter K. Unger Milton K. Munitz, editor, *Semantics and Philosophy*, pages 197–214. New York: New York University Press, 1974.

Pavel Tichý. A counterexample to the Stalnaker-Lewis analysis of counterfactuals. *Philosophical Studies*, 29:271–273, 1976.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7): 557–585, 1921.