Kaibo Xie, University of Amsterdam

Belief and Causality in the Epistemic Reading of Counterfactuals

The entanglement between belief and counterfactual conditionals has been discussed at large within the philosophical literature on the semantics of conditionals. The starting point is the famous *Ramsey's Test*, which Stalnaker (1968) reformulates as "First, hypothetically add the antecedent to your stock of beliefs; second, make adjustments to maintain consistency without modifying the hypothetical belief; finally, consider whether or not the consequent is true". Many philosophers believe that this 'conditional is belief revision' idea can only be applied to indicative conditionals. An example often used to support this position is the famous Kennedy example from Adams (1970). Adams concludes that indicative and counterfactual conditionals "are logically distinct species." Lewis (1973) says on this matter: "Therefore there really are two different sorts of conditional; not a single conditional that can appear as indicative or as counterfactual depending on the speaker's opinion about the truth of the antecedent". Thus for counterfactuals we need a different approach than belief revision. For a long time the similarity approach of Stalnaker and Lewis was the standard way to account for counterfactual conditionals. But there has been a very successful recent proposal by Pearl (2002) building on his account of causality in terms of structural equations. However, people have also argued that, next to the *ontic* reading of counterfactuals described by the similarity approach and Pearl's recent proposal, there is a second *epistemic* reading. Kratzer (1989) illustrates the latter with the King Ludwig example.

Example 1. *King Ludwig of Bavaria likes to spend his weekends in Leoni Castle. Whenever the king is in the castle, the lights will be on and the royal flag will be up. A traveler watches the castle from a distance and sees that the lights are on. The flag, however, is not up. He says:*

(1) If the flag had been up, the king would have been in the castle.

This counterfactual seems intuitively to be true in the given context. However, the dominant approaches to counterfactuals at the moment, like Pearl (2002)'s approach, cannot account for this intuition. (1) is simply false according to Pearl's account: as the flag has no causal effect on the king, changing the status of the flag will not actually make the king appear in the castle. Pearls approach captures reasoning from cause to effect, but in this example we clearly reason in the other direction. Belief must be involved in such reasoning because modifying the effect only leads to belief change of the cause instead of change of the cause. An alternative way to explain our intuition for (1) is the 'belief revision' approach, according to which a counterfactual is true in case the consequent holds after revising one's belief with the antecedent. However, according to the standard approaches to belief revision, this approach also fails for (1). The problem is that it cannot explain why the light should remain on even if the status of the flag is modified: it does not take the causal independence into account. If we want to solve this problem, the model our account based on must take two aspects into account: on one hand, it must be able to express belief state in order to make the reasoning from effect to cause possible; on the other hand it must encode the causal information.

In this paper I will build a "causal epistemic model" as an extension of the causal model in Pearl (2002). On top of the information encoded already in Pearl's models, these epistemic causal models will also represent the belief states of an agent, using a plausibility order over possible worlds. I will also extend language with belief operators which allows us to define the condition of accepting a counterfactual under epistemic readings.

A structural equation model is an elegant and descriptively strong account of causality in terms of structural equations, functions that deterministically pine down the value of a variable given the value of other variables that have causal effect on it (Pearl, 2002). The model divides causal variables into exogenous variables and endogenous variables. Given values of exogenous variables, structural equations determine values of all variables when any change of value is made to some of them, so the model predicts the result of modifying certain values. Pearls modeling accounts not only for deterministic causality but also for non-deterministic one by simply adding variables representing disturbance and probabilistic distribution over variables.

The first step of the proposal is to enrich these models with a representation of the beliefs of an agent. One could object here that this is superfluous, because the models already contain a probability distribution over the exogenous variables. But using probabilities to interpret a belief operator is problematic. One could try to define beliefs in terms of probability. However there is a problem of finding a numerical threshold for the probability of ϕ for an agent to believe ϕ. To avoid this problem we will replace the probability functions of Pearl's approach with a qualitative representation of the beliefs of an agent. We follow two steps. In the first, we will lift Pearl's model, based on causal variables, into models based on possible worlds. In these new models, probability is assigned to possible worlds instead of causal variables. The second step builds on Baltag and Smets (2008), translating the probability distribution over possible worlds into a plausibility ordering over them. That gives us a qualitative representation of the quantitative information in the original model. Thus, we have a purely qualitative plausibility model together with the deterministic structural equations from the original one.

Pearl's approach to the ontic reading of counterfactuals can be restated using the models introduced above. Halpern uses the formal language $[\vec{X} = \vec{x}]\phi$ to express that ϕ holds after modifying the value of $\vec{X}$ to $\vec{x}$. The truth condition of "$[\vec{X} = \vec{x}]\phi$ holds at a possible world w" is as follows: first set values of $\vec{X}$ to $\vec{x}$, then the structural equations uniquely determines values of all endogenous variables (given the values of the exogenous variables at w); second, as values of all causal variables are determined in the first step, we can check whether ϕ holds under such setting of causal variables. The operator $[\vec{X} = \vec{x}]$ is known as "intervention". In our model, since each possible world is a full assignment of values on variables, intervention can be seen as a transition between possible worlds. Thus, a counterfactual statement is true on a possible world if and only if its consequent holds on the possible world that is reached by the transition corresponding to the antecedent.

Our new models allow us to add a belief operator Bel to our formal language. With counterfactuals evaluated locally, belief operator can be applied to counterfactuals as well and belief in a counterfactual: the agent believes ϕ if and only if ϕ holds on the most plausible worlds.

The richer language introduced this way allows us to distinguish two readings of a counterfactual in the form of (2) "if $[\vec{X} = \vec{x}]$ had been the case, ϕ would have been the case", both based on an interventionist interpretation of the antecedent: (2) is accepted at a possible world w under epistemic reading iff $Bel[\vec{X} = \vec{x}]Bel\phi$ holds at w; (2) is accepted under the ontic reading, iff $Bel[\vec{X} = \vec{x}]\phi$ holds at w. We will see that according to this definition, the acceptance condition under the ontic reading coincides with (Pearl, 2002)'s account. And the causal epistemic model predicts the desired truth value of (1) in King Ludwig's example: on one hand the structural equations guarantee that the light is still on after taking the flag up; on the other hand, our setting of plausibility ordering guarantees that the agent considers a possible world in which "the king is in the castle with flag up and light on" to be more plausible than a possible world in which "the king is not in the castle but with flag up and light on", because the latter violates the causal rules. I will also show that such interpretation based on causal epistemic models is generalizable: it keeps right predictions in other examples with counterfactuals.

References

Adams, E. W. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, pages 89–94.

Baltag, A. and Smets, S. (2008). Probabilistic dynamic belief revision. *Synthese*, 165(2):179–202.

Halpern, J. Y. (2013). From causal models to counterfactual structures. *The Review of Symbolic Logic*, 6(2):305–322.

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

Lewis, D. (1973). Counterfactuals and comparative possibility. In *Ifs*, pages 57–85. Springer.

Pearl, J. (2002). Causality: models, reasoning, and inference. *IIE Transactions*, 34(6):583–589.

Stalnaker, R. C. (1968). A theory of conditionals. In *Ifs*, pages 41–55. Springer.