A dynamic account for backtracking counterfactuals

KATRIN SCHULZ
(ILLC, University of Amsterdam)

This project combines recent research in linguistics, logic and psychology on the topic of counterfactual reasoning. Here, we are particularly interested in backtracking. Backtracking counterfactuals ask us to reason backward in time, as in (1).

(1) If the pavement had been wet, it would have rained last night.

The prominent approaches towards the meaning of counterfactuals are all based on the idea that in order to evaluate a counterfactual, we have to change the course of history just slightly before the antecedent became false, make the antecedent true and then run the natural course of events and check whether the consequent becomes true as well [Lewis 1979, Pearl 2013]. Such approaches predict (1) to be false. But various philosophers and linguists observe that sometimes backtracking seem intuitively acceptable [Arregui 2005, Lewis 1979]. The goal of this paper is to account for backtracking, but also explain the intuition driving the standard approach.

Backtracking. Recently, a number of interesting empirical studies on backtracking were published. Sloman & Lagnado (2005) confirm that people rarely backtrack. However, Rips & Edwards (2013) observe in very similar settings a frequent occurrence of backtracking. Trying to account for these different findings Gerstenberg et al. (2013) observe that the studies (Sloman & Lagnado, 2005) and (Rips & Edwards 2013) present their questions to the participants in a different order. Both studies use a scenario (Figure 1) with 4 propositional variables A, B, C and D, where the truth of A causes B and C to be true and that, in turn, causes D to be true (see the equations in Figure 1). The participants were told that A, B, C and D are all true and then were asked to consider what would have been the case, if B hadn't been true. The study (Sloman & Lagnado, 2005) asks the participants first whether D would hold, if B was false, and then turns to A. The study (Rips & Edwards 2013), on the other hand, asks about the consequences on A, C and D at the same time, but the order of the questions on the screen invites the participants to first consider A before they turn to C and D. In this case backtracking occurs frequently. Gerstenberg et al. (2013) empirically confirm this dependency of backtracking on the order of questions.

The picture emerging from these results is that given the right setting of the context, backtracking is possible. The crucial contextual parameter on which backtracking depends seems to be what the interpreter is paying attention to, and that changes dynamically in discourse. Backtracking is only available, if the attention is explicitly drawn to the causal past of the antecedent.
The Proposal. We propose that the interpretation of counterfactuals doesn't involve a change in the course of history, whether by intervention (Pearl 2013), miracle (Lewis 1979) or other means. Instead, a counterfactual is always interpreted with respect to a submodel of the full set of causal laws. This submodel normally doesn't contain the causal history of the antecedent. Therefore, normally backtracking doesn't occur. Thus, the forward-looking orientation of counterfactuals is not due to cutting off the causal history of the antecedent, but due to a tendency to ignore it. This perspective allows us to account for backtracking, in case the attention is explicitly drawn to the causal history of the antecedent. Furthermore, by allowing the relevant submodel to grow in a counterfactual discourse (similar to von Fintel, 2001), we can capture the dynamic effects reported above.

This approach can be spelled out using a dynamic version of the causal network approach (Pearl, 2013). It correctly predicts the observations of Sloman & Lagnado (2005), Rips & Edwards (2013) and Gerstenberg et al. (2013) concerning Figure 1: if the interpreter first evaluates whether A would be the case, then C, then D, backtracking occurs and A is predicted to be false. However, if the interpreter considers first whether D would be the case, then considers C, and finally turns her attention to A, backtracking does not occur.

Testing. At the moment we are running an empirical study testing the proposal, using a setting that involves indirect backtracking (Figure 2). In this example A causes B and B prevents A from causing C as well. The truth/acceptability of the counterfactual (2) in Figure 2 changes depending on whether the interpreter backtracks from ¬B to ¬A or holds the truth of A fixed. We predict that even though interpreting (2) involves A, this doesn't draw attention to A as the cause of B. Therefore, no backtracking occurs and (2) should be judged true. This prediction was confirmed (66% of the participants judged (2) true). Furthermore, we predict an order effect: if asked first about the truth of (2) and then about (3), backtracking should be out and (3) judged false. But asking first about (3) should make backtracking possible. The order, however, did turn out to have a significant effect on the possibility of backtracking. In our setting backtracking occurred much less frequent than observed in Rips & Edwards (2013).