The algorithmization of counterfactuals

Judea Pearl
Computer Science Department Cognitive Systems Lab 4532 Boelter Hall University of California
Los Angles, CA 90024-1596

Abstract: One of the most striking phenomenon in the study of conditionals is the ease and uniformity with which people generate, evaluate and interpret counterfactual utterance. To witness, the majority of people would accept the statement: "If Oswald didn't kill Kennedy, someone else did," but few, if any, would accept its subjunctive version: "If Oswald hadn't killed Kennedy, someone else would have."

I will present a computational model that explains how humans reach such consensus or, more concretely, what mental representation permits such consensus to emerge from the little knowledge we have about Oswald, Kennedy and 1960's Texas, and what algorithms would need to be postulated to account for the swiftness, comfort and confidence with which such judgments are issued.

The model presented is compatible with the "possible world" account of Lewis (1973), yet it enjoys the advantages of representational economy, algorithmic simplicity and conceptual clarity.

Armed with these advantages, I will then present a panoramic view of several applications where counterfactual reasoning has benefited problem areas in the empirical sciences, including policy evaluation, causal-pathways mapping, credit and blame analysis, and personal decision making.