On Counterfactuals and Cognitive Science: Rumlhart Prize Symposium in Honor of Judea Pearl

Nick Chater (nick.chater@wbs.ac.uk)
Behavioural Science Group, Warwick Business School
Lance Rips (rips@northwestern.edu)
Psychology Department Northwestern University, 2029 Sheridan Road
Evanston IL 60208
Jim Joyce (jjoyce@umich.edu)
Department of Philosophy University of Michigan
Ann Arbor, MI 48109-1003
Stefan Kaufmann (kaufmann@northwestern.edu)
Department of Linguistics Northwestern University, 2016 Sheridan Rd.
Evanston IL 60208-4090
Steven Sloman (Steven_Sloman@brown.edu)
Cognitive, Linguistic, & Psychological Sciences Brown University, Box 1821
Providence, RI 02912
Judea Pearl (judea@cs.ucla.edu)
Computer Science Department Cognitive Systems Lab UCLA, 4532 Boelter Hall

Keywords: counterfactuals; causal reasoning, explanation; belief revision; linguistics

Motivation

Counterfactuals are the building blocks of scientific thought and the oxygen of moral behavior. The ability to reflect back on one's past actions and envision alternative scenarios is the basis of free will, responsibility, and social adaptation. Recent progress in the algorithmization of counterfactuals has advanced our understanding of this mode of reasoning and has brought us a step closer toward equipping machines with similar capabilities. I hope this symposium will inspire cognitive scientists to empower themselves with these new tools, and to tackle some of the more difficult problems that counterfactuals present: why evolution has endowed humans with the illusion of free will and how it manages to keep that illusion so vivid in our brain.

Title 1: Mental mechanisms: Reasoning About How the World Might Have Been

Author: Nick Chater

Abstract : Pearl (2000) argues that intelligent systems must primarily represent the causal structure of the world, and provides a revolutionary theory of reasoning with causal knowledge. Applying these ideas to the psychology of reasoning, Mike Oaksford and I have argued that reasoning occurs over local "mental mechanisms," rather than propositional or purely probabilistic representations. These mechanisms are, instead, analogous to programs; and automatically embody counterfactual claims (e.g., if a particular register were reset to value x, at time t, then…). Reasoning about a situation requires constructing, and then reasoning about a mental mechanism; reasoning errors may be generated at both steps. This perspective has implications for empirical and theoretical research on learning and reasoning.

Title 2: Counterfactual States and Explanatory Search

Authors: Lance J. Rips and Brian J. Edwards

Abstract : We report the results from studies of how people answer counterfactual questions about simple machines. Participants learned about devices that have a specific configuration of components, and they answered questions of the form "If component X had not operated, would component Y have operated?" We compare the results of these decisions to the predictions of rival Bayes-net theories of counterfactual reasoning (e.g., Hiddleston, 2005; Pearl, 2000) and describe some departures from these models. The results suggest that people try to construct an explanation for the counterfactual state—why component X had not operated—while attempting to preserve the device's operating principles. Participants tended

to prefer simpler explanation—explanations that require fewer changes to the device—but they sometimes had trouble tracing the logical implications of these changes. These difficulties help predict the pattern of results.

Title 3: Counterfactuals and Belief Revision

Author: Jim Joyce

Abstract : Counterfactuals fall within a circle of interconnected concepts, which includes laws of nature, cause and effect relations, physical propensities and objective single-case chances. As Pearl stresses, these concepts are indispensable both to statistical inference and to decision making. Moreover, none of them can be reductively analyzed in terms of correlations, associations or any other non-model notions. I will discuss two recent approaches to counterfactual reasoning, and will explore their impact on belief revision: Pearl's treatment, which makes use of "mutilated" causal models, emphasizes the importance of structural equations; David Lewis's "imaging" approach emphasizes similarities between actual and counterfactual situations. On most questions, there turns out to be less difference between these approaches than there might first seem, but the two can diverge when disjunctions are counterfactually supposed. While it is not entirely clear what to say about such cases, I will briefly lay out some of the pros and cons of each position. I will also, if time permits, say something about the central place of counterfactual reasoning in decision making.

Title 4: Talking About Interventions: New Questions for Linguistic Research

Authors: Stefan Kaufmann

Abstract : Causal (Bayesian) Networks are increasingly being used in the semantic analysis of counterfactual conditionals and related linguistic constructions. How are these constructions interpreted relative to Causal Networks? There are really two sides to this question: First, how are Causal Networks utilized and operated on in counterfactual reasoning? There is widespread agreement that some form of local intervention is involved, but also mounting evidence that this operation is not as simple and uniform as severing a variable from its parents. Second, inasmuch as intervention is flexible and context-dependent, how is it driven in particular cases by the linguistic properties of the sentence in question? These questions must be kept distinct but inform each other. I discuss ways for linguistic theory to both benefit from and contribute to research on causal and counterfactual reasoning.