Causal Explanation as Constraint Satisfaction : A Critique and a Feedforward Connectionist Alternative

Frank Van Overwalle (FRANK.VANOVERWALLE@VUB.AC.BE)
Department of Psychology
Vrije Universiteit Brussel
B-1050 Brussel, Belgium

Introduction

Read and Marcus-Newhall (1993) proposed that causal explanation should be viewed from the perspective of a constraint satisfaction process, in which people choose a cause among several alternative explanations that is the most *coherent* with the events to be explained. Based on Thagard's (1992) theory of explanatory coherence, they put forward four principles that underlie the coherence of social explanations : Perceivers should prefer the explanation that accounts for most of the evidence (*breadth*), that is most parsimonious (*simplicity*), and that is explained by other causes (*being explained*). Moreover, the strength of an explanation also depends on the availability of alternative explanations (*competition*). In a series of studies, Read and Marcus-Newhall (1993) found not only strong empirical support for Thagard's four principles of coherence, but were also able to closely simulate subjects' explanatory ratings with Thagard's ECHO constraint satisfaction program.

Limitations of ECHO

However, ECHO has several important shortcomings. First, ECHO fails to be sensitive to covariation or contingency (even in the most simple case with one cause and one event) which is crucial to causality. This can be demonstrated by simulations as well as by mathematical proof. Second, ECHO does not learn from experience. Hence, it requires arbitrary assumptions about the strength of diverse explanations based on prior ratings of subjects or good sense. As another consequence, to simulate the principle of competition (or discounting), it requires that true and discounted causes are always simultaneously activated, which is psychologically implausible. Third, it cannot deal efficiently with conjunctions of explanations. Therefore, two slightly different network architectures are needed that include either individual causes or their conjunction but not both, which is also very implausible psychologically.

Feedforward Alternative

To remedy these limitations, an alternative feedforward connectionist model is proposed which has several advantages. First, the feedforward model is implemented with the delta algorithm which allows it to learn from exposures to causes and their effects. This algorithm is part of a broad theory of learning, not only of causal learning but also of categorization, and is formally equivalent to the Rescorla-Wagner model of animal conditioning. Second, it has been shown mathematically that the delta algorithm is susceptible to covariation given appropriate encoding specifications. Moreover, the model can replicate other classical attribution principles of discounting and augmentation. Third, the model can handle both individual causes as well as their conjunctions by adopting Gluck and Bower's (1988) configural-cue architecture.

Simulating Four Principles of Coherence

I also explored to what extent the feedforward model is capable of simulating the four principles of coherence, by running novel simulations of Read and Marcus-Newhall's data. The results showed that for all four principles of coherence, the feedforward network was able to mimic the predicted effects. Although the feedforward simulations did not replicate the human data from Read and Marcus-Newhall (1993) perfectly, the results were often better than those of the ECHO constraint satisfaction network.

Conclusion

Although the constraint satisfaction approach may provide insights in some domains, its contribution cannot lie in causal explanation, at least not in the format of ECHO as proposed by Thagard (1992) and Read and Marcus-Newhall (1993). In contrast, my theoretical analysis and simulations suggest that the feedforward model provides a more promising alternative for the principles of coherence and causality.

References

Gluck, M. A. & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166-195.

Read, S. J. & Marcus-Newhall, A. (1993) Explanatory coherence in social explanations : A parallel distributed processing account. *Journal of Personality and Social Psychology, 65,* 429-447.

Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ : Princeton University Press.