Causation and Counterfactuals in Medical Decision-Making: Philosophical, Artificial Intelligence and Psychological Perspectives on Mapping Theory onto Application

Abstract

How do individuals reason about causation? What does it mean for theory, and application? Few scholars, I believe, would contest the scope, relevance and significance of these questions when it comes to developing insights about the human condition and the mind/brain mystery. Although theories and applications of causation exist as integral elements in many mono-disciplinary academic frameworks, this paper takes a uniquely cognitive science perspective, and concentrates on the point where philosophy, artificial intelligence and psychology intersect.

Some philosophical groundwork is in order. For there to be inferences of causation, individuals must structure their understanding around the axioms of causal relations and mechanisms. However, the quandary is that the causal mechanisms that are occurring may not always be subject to direct observation, and thus, there is an inherent prerequisite for counterfactuals to be incorporated (Lewis, 1973). Counterfactual reasoning, via hypothetical suppositions, attempts to elucidate how matters actually are, by contrasting them to what they are *not*, or could *not* be (Rescher, 2001). In other words, counterfactual reasoning relies on a nexus of speculative propositions contrary to the reality of the state of affairs. Counterfactual reasoning is the cornerstone of knowledge advancement, given that determining what something is *not*, or *cannot* be, may very well be as urgently vital to our understanding of a phenomenon as comprehending what it is. In terms of understanding causation, counterfactual reasoning is indelibly constrained by how individuals reason about causation, with the caveat that mere correlations are not counterfactual.

However, reasoning about causation by way of counterfactuals is somewhat problematic. Take for example, what is known as the base-rate fallacy, or base-rate neglect. It is a robust finding in psychology, where in the process of calculating probabilities, individuals have a tendency to underutilize base-rates and instead overemphasize irrelevant information (e.g., Bar-Hillel, 1980; Kahneman & Tversky, 1973). Cohen (1986) argued that the manifestation of base-rate fallacies can be attributed to the tendency of counterfactualizable probabilities to dominate over non-counterfactualizable ones. According to Cohen (1986), a probability is considered counterfactualizable if, and only if, it applies not only to entities that comprise the reference-class but also to those that do not. As such, counterfactualizable probabilities will persist, even if numbers of members in the reference class change, and is therefore considered an intrinsic probability that is stable. Further, counterfactualizable probabilities have the potential to inform decisions, given that it is applicable to infinite possibilities and thus allows for predictive power. A non-counterfactualizable probability, on the other hand, is an accidental property of its reference-class and as such, has very restricted applicability and minimal predictive power. Given that individuals are predisposed to adopt the probability with the greatest predictive power, Cohen (1986) concludes that mistakes in logic are commonplace.

To address this human fallibility, artificial intelligence researchers have selectively modernized the semantics of counterfactuals, and used technology to customize this traditionally philosophical tool. For example, Ginsberg (1986), influenced by the work of Lewis (1973) and Stanlnaker (1968), has proposed that counterfactuals can be understood by way of a database

based on simple predicate calculus. Jackson (1989), on the other hand argues that it is Winslett's (1988) possible models approach that provides the accurate and useful foundation for understanding the semantics of counterfactuals. One of the manifestations of this approach is modification of Bayes theorem (Bayes, 1763). In essence, Bayes' theorem is built on a network of probabilities, which dually represents a set of variables, (known as nodes in artificial intelligence) and their conditional independencies (connectors) (Howson & Urbach, 1993; Yudkowsky, 2003). Bayesian networks were originally designed based on probabilistic functionality, but since their inception, have evolved into modifiable structural models, which allow for the additional computation of effects of new actions and counterfactual statements (Balke & Pearl, 1997)

When the above causation-and-counterfactual theoretical template is mapped onto application, one context in which to explore causation and counterfactual reasoning is that of decision-making. Specifically, reference is made to medical decision-making, in the sense of diagnosing diseases from the interpretation of symptoms presented. First, from a philosophical perspective, counterfactual reasoning is inextricably linked to medical decision-making because the diagnosis of disease is limited by how individuals reason about causation. That is, in order to infer causation (e.g., symptom B is caused by disease A), individuals must reason that if disease A causes symptom B, then symptom B would not have materialized if disease A had not been present, even if disease A did occur (Sloman, 2005). Second, from a psychological standpoint, mistakes are made in medical decision-making, in ways that Bar-Hillel (1980), Kahneman and Tversky (1973) have described. Obviously, causal inference by way of counterfactual reasoning may be deficient in this aspect (Cohen, 1986). Third, from an AI mindset, the recent trend toward decision support systems (DSSs) that rely on Bayes' theorem is driving the impetus to find deeper insights into counterfactual reasoning in technical systems. Using the semantics of counterfactuals, a modern Bayesian diagnostic model may operate by estimating the probabilistic relationship between a myriad of diseases and symptoms. In other words, given a specific symptom, the Bayesian network performs probabilistic computations to diagnose the presence of particular diseases (Lindgaard, 1985; 2005; 2007). Within the Bayesian framework, "diagnosticity is a measure of the specificity of a given sign or symptom in support of a particular hypothesis. Specificity is determined by the relative frequency of occurrence of the sign or symptom in question under a range of diseases" (Lindgaard, 2007, ¶ 10). The notion that biases in medical decision making can be reduced by the incorporation of a Bayesian diagnostic network is an intuitively appealing one.

The mapping of theory onto application is imperfect - arguably, more research is needed to enable the growth and maturation and improvement of causation-and-counterfactual theory and practice. It is critical that this is grounded in a cognitive science perspective, given that the information gleaned will have extensive theoretical and practical implications for counterfactual reasoning approaches in philosophy, artificial intelligence and psychology. As such, an interdisciplinary approach is strongly, and urgently, advocated.

References

Balke, A.A., & Pearl, J. (1997). Probabilistic Counterfactuals: Semantics, Computation, and Applications (Defense Technical Information Centre Rep. No. ADA332296). Los Angeles, CA: California University.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211 – 233.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions, Giving Some Account of the Present Undertakings, Studies and Labours of the Ingenious in Many Considerable Parts of the World 53,* 370 - 418.

Cohen, J.L. (1986). *The dialogue of reason: An analysis of analytic philosophy.* New York, NY: Oxford University Press, USA.

Howson, C., & Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach (Third Edition).* Chicago, Illinois: Open Court Books.

Jackson, P. (1989). On the semantics of counterfactuals. In N. S. Sridharan (Ed.). *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI-89).* (pp. 1382 – 1387). Detroit, MI: Morgan Kaufmann Publishers.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80,* 237 - 251.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70* (17), 556 – 567.

Lindgaard, G.. (1985). *Weighting of individuating information elements and base rates in a nursing decision-making task involving nondiagnostic case information*, Unpublished MSc thesis, Department of Psychology, Monash University, Melbourne, Australia.

Lindgaard, G. (2005). Human judgment, decision theory, and technology: Applications of Bayes' Theorem, *History and philosophy of psychology Bulletin*, *17*(2), 29 - 39.

Lindgaard, G. (2007). Bayesian models in decision support systems in traditional medicine and e-helath., *HOT Topics*, 6 (4). Retrieved November 23, 2007, from http://hot.carleton.ca/hot-topics/articles/bayesian-models-in-decision-support-systems/

Rescher, N. (2001). *Philosophical reasoning: A study in the methodology of philosophizing*. Boston, MA: Blackwell Publishing Limited

Sloman, S.A. (2005). Causal models: How people think about the world and its alternatives. New York, NY: Oxford University Press, USA.

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in Logical Theory.* Oxford, U.K.: Oxford University Press.

Winslett, M.A (1988). Reasoning about action using a possible models approach. In R. G. Smith and T. M. Mitchell (Eds.), *Proceedings of the 7th National Conference on Artificial Intelligence: August, 1988,* (pp 89 – 93). St Paul, MN: Morgan Kaufmann Publishers

Yudkowsky, E.S. (2003). An Intuitive Explanation of Bayesian Reasoning. Retrieved November 23, 2007, from http://yudkowsky.net/bayes/bayes.html