TECHNICAL REPORT
R-512
November 2021

Causation and Decision Theory – An Introduction

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

November 9, 2021

All accounts of rational behavior presuppose knowledge of how actions affect the state of the world and how the world would change had alternative actions been taken. If the options available to an agent are specified in terms of their immediate consequences, as in "make him laugh," "paint the wall red," "raise taxes" or, in general, $do(X = x)$, then a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum_y P_x(y)U(y) \tag{1}$$

over all options x. Here, $U(y)$ stands for the utility of outcome $Y = y$ and $P_x(y)$ – the focus of this paper - stands for the (subjective) probability that outcome $Y = y$ would prevail, had action $do(X = x)$ been performed so as to establish condition $X = x$.

Most studies of decision theory have dealt with the utility function $U(y)$, its behavior under various shades of uncertainty, and the adequacy of the expectation operator in Eq. (1). Relatively little has been said about the probability $P_x(y)$ that governs outcomes $Y = y$ when an action $do(X = x)$ is contemplated. Yet regardless of what criterion one adopts for rational behavior, it must incorporate knowledge of how our actions affect the world. We must therefore define the function $P_x(y)$ and explicate the process by which it is assessed or inferred, be it from empirical data or from world knowledge. We must also ask what mental representation and thought processes would permit

a rational agent to combine world knowledge with empirical observations and compute $P_x(y)$.

It has long been recognized that Bayesian conditionalization, i.e., $P_x(y) = P(y|x)$, is inappropriate for serving in Eq. (1), for it leads to paradoxical results of several kinds (see Pearl 2000, pp. 108–9; Skyrms 1980). For example, patients would avoid going to the doctor to reduce the probability that one is seriously ill; barometers would be manipulated to reduce the chance of storms; doctors would recommend a drug to male and female patients, but not to patients with undisclosed gender, and so on. Yet the question of what function should substitute for $P_x(y)$, despite decades of thoughtful debates (Cartwright, 1983; Harper et al., 1981; Jeffrey, 1965) seems to still baffle philosophers in the 21st century (Arlo-Costa, 2007; Weirich, 2020; Woodward, 2003). Modern discussion over evidential vs. causal decision theory (Weirich, 2020) echo these debates.

Causal inference research in the past 3 decades has settled this debate by defining $P_x(y)$ in terms of Structural Causal Models (SCM) which allows one to compute counterfactuals directly from the way people store knowledge about the world.

The theory that emerges from SCM offers several conceptual and operational advantages over Lewis's closest-world semantics of counterfactuals. First, it does not rest on a metaphysical notion of "similarity," which may differ from person to person and, thus, could not explain the uniformity with which people interpret causal utterances. Instead, causal relations are defined in terms of our scientific understanding of how variables interact with one another. Second, it offers a plausible resolution of the "mental representation" puzzle: How do humans represent "possible worlds" in their minds and compute the closest one, when the number of possibilities is far beyond the capacity of the human brain? Any credible theory of rationality must account for the astonishing ease with which humans comprehend, derive and communicate counterfactual information. Finally, it results in practical algorithms for solving some of the most critical and difficult causal problems that have challenged data analysts and experimental researchers in the past century (see Pearl and Mackenzie (2018) for extensive historical account). These include: the control confounding and predicting effects of interventions and policies, defining and estimating direct and indirect effects, generating explanations and estimating causes of effect, managing missing data, and generalizing empirical results across diverse environments.

Acknowledgement

This research was supported in parts by grants from the National Science Foundation [#IIS1704932], Office of Naval Research [#N00014-17-S-12091 and #N00014-21-1-2351] and Toyota Research Institute of North America [#PO-000897].

References

Arlo-Costa, H. (2007). The Logic of Conditionals. In *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Winter 2007 ed. https://plato.stanford.edu/archives/win2007/entries/logic-conditionals/, Metaphysics Research Lab, Stanford University.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Harper, W., Stalnaker, R. and Pearce, G. (1981). *Ifs*. D. Reidel, Dordrecht.

Jeffrey, R. (1965). *The Logic of Decision*. McGraw-Hill, New York.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

Skyrms, B. (1980). *Causal Necessity*. Yale University Press, New Haven.

Weirich, P. (2020). Causal Decision Theory. In *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Winter 2020 ed. https://plato.stanford.edu/archives/win2020/entries/decision-causal/, Metaphysics Research Lab, Stanford University.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.