26

Cause for Celebration, Cause for Concern

YOAV SHOHAM

It is truly a pleasure to contribute to this collection, celebrating Judea Pearl's scientific contributions. My focus, as well as that of several other contributors, is on his work in the area of causation and causal reasoning. Any student of these topics who ignores Judea's evolving contributions, culminating in the seminal [Pearl 2009], does so at his or her peril. In addition to the objective content of these contributions, Judea's unique energy and personality have led to his having unparalleled impact on the subject, in a diverse set of disciplines far transcending AI, his home turf. This body of work is truly a cause for celebration, and accounts for the first half of the title of this piece.

The second half of the title refers to a concern I have about the literature in AI regarding causation. As an early contributor to this literature I wade back into this area gingerly, aware of many of the complexities involved and difficulties encountered by earlier attempts to capture the notion formally. I am also aware of the fact that many developments have taken place in the past decade, indeed many associated with Judea himself, and only some of which I am familiar with. Still, it seems to me that the concern merits attention. The concern is not specific to Judea's work, and certainly applies to my own work in the area. It has to do with the yardsticks by which we judge this or that theory of causal representation or reasoning.

A number of years ago, the conference on Uncertainty in AI (UAI) held a panel on causation, chaired by Judea, in which I participated. In my remarks I listed a few requirements for a theory of causation in AI. One of the other panelists, whom I greatly respect, responded that he couldn't care less about such requirements; if the theory was useful that was good enough for him. In hindsight that was a discussion worth developing further then, and I believe it still is now.

Let us look at a specific publication, [Halpern and Pearl 2001]. This selection is arbitrary and I might as well have selected any number of other publications to illustrate my point, but it is useful to examine a concrete example. In this paper Halpern and Pearl present an account of "actual cause" (as opposed to "generic cause"; "the lighting last night caused the fire" versus "lightnings cause fire"). This account is also the basis for Chapter 10 of [Pearl 2009]. Without going into their specific (and worthwhile) account, let me focus on how they argue in its favor. In the third paragraph they say

> While it is hard to argue that our definition (or any other definition, for that matter) is the "right definition", we show that it deals well with the difficulties that have plagued other approaches in the past, especially those exemplified by

> the rather extensive compendium of [Hall 2004][1].

The reference is to a paper by a philosopher, and indeed of the thirteen references in the paper to work other than by the authors themselves, eight are to work by philosophers.

This orientation towards philosophy is evident throughout the paper, in particular in their relying strongly on particularly instructive examples that serve as test cases. This is an established philosophical tradition. The "morning star – evening star" example [Kripke 1980] catalyzed discussion of cross-world identity in first-order modal logic (you may have different beliefs regarding the star seen in the morning from those regarding the star seen in the evening, even though, unbeknownst to you, they are in fact the same star – Venus). Similarly, the example of believing that you will win the lottery and coincidentally later actually winning it served to disqualify the definition of knowledge as true belief, and a similar example argues against defining knowledge as *justified* true belief [Gettier 1963].

Such "intuition pumps" clearly guide the theory in [Halpern and Pearl 2001], as evidenced by the reference to [Hall 2004] mentioned earlier, and the fact that over four out of the paper's ten pages are devoted to examples. These examples can be highly instructive, but the question is what role they play. In philosophy they tend to serve as necessary but insufficient conditions for a theory. They are necessary in the sense that each of them is considered sufficient grounds for disqualifying a theory (namely, a theory which does not treat the example in an intuitively satisfactory manner). And they are insufficient since new examples can always be conjured up, subjecting the theory to ever-increasing demands.

This is understandable from the standpoint of philosophy, to the extent that it attempts to capture a complex, natural notion (be it knowledge or causation) it its full glory. But is this also the goal for such theories in AI? If not, what is the role of these test cases?

If taken seriously, the necessary-but-insufficient interpretation of the examples presents an impossible challenge to formal theory; a theoretician would never win in this game, in which new requirements may surface at any moment. Indeed, most of the philosophical literature is much less formal than the literature in AI, in particular [Halpern and Pearl 2001]. So where does this leave us?

This is not the first time computer scientists have faced this dilemma. Consider knowledge, for example. The S5 logic of knowledge [Fagin, Halpern, Moses, and Vardi 1994] captures well certain aspects of knowledge in idealized form, but the terms "certain" and "idealized" are important here. The logic has nothing to say about belief (as opposed to knowledge), nor about the dynamic aspects of knowledge (how it changes over time). Furthermore, even with regard to the static aspects of knowledge, it is not hard to come up with everyday counterexamples to each of its axioms.

And yet, the logic proves useful to reason about certain aspects of distributed systems, and the mismatch between the properties of the modal operator K and the everyday word "know" does not get in the way, within these confines. All this changes as one switches the context. For example, if one wishes to consider cryptographic protocols, the K axiom ($Kp \wedge K(p \supset q) \supset Kq$, valid in any normal modal logic, and here representing logical

[1] They actually refer to an earlier, unpublished version of Hall's paper from 1998.

omniscience) is blatantly inappropriate. Similarly, when one considers knowledge and belief together, axiom 5 of the logic ($\neg Kp \supset K\neg Kp$, representing negative introspection ability) seems impossible to reconcile with any reasonable notion of belief, and hence one is forced to retreat back from the S5 system to something weaker.

The upshot of all this is the following criterion for a formal theory of natural concepts: One should be explicit about the intended use of the theory, and within the scope of this intended use one should require that everyday intuition about the natural concepts be a useful guide in thinking about their formal counterparts.

A concrete interpretation of the above principle is what in [Shoham 2009] I called the *artifactual* perspective.[2] Artifactual theories attempt to shed light on the operation of a specific artifact, and use the natural notion almost as a mere visual aid. In such theories there is a precise interpretation of the natural notion, which presents a precise requirement for the formal theory. One example is indeed the use of "knowledge" to reason about protocols governing distributed systems. Another, discussed in [Shoham 2009], is the use of "intention" to reason about a database serving an AI planner.

Is there a way to instantiate the general criterion above, or more specifically the artifactual perspective, in the context of causation? I don't know the answer, but it seems to me worthy of investigation. If the answer is "yes" then we will be in a position to devise provably correct theories, and the various illustrative examples will be relegated to the secondary role of showing greater or lesser match with the everyday concept.

Acknowledgments: This work was supported by NSF grant IIS-0205633-001.

References

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1994). *Reasoning about Knowledge*. MIT Press.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis 23*, 121–123.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. MIT Press.

Halpern, J. Y. and J. Pearl (2001). Causes and explanations: A structural-model approach. part I: Causes. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, San Francisco, CA, pp. 194–202. Morgan Kaufmann.

Kripke, S. A. (1980). *Naming and necessity* (Revised and enlarged ed.). Blackwell, Oxford.

Pearl, J. (2009). *Causality*. Cambridge University Press. Second edition.

Shoham, Y. (2009). Logics of intention and the database perspective. *Journal of Philosophical Logic 38*(6), 633–648.

[2] The discussion there is done in the context of formal models of intention, but the considerations apply here just as well.