Computer Science Department Technical Report
University of California
Los Angeles, CA 90024-1596

PROBABILISTIC AND QUALITATIVE ABDUCTION

Judea Pearl

July 1991
CSD-910042

Presented, AAAI Spring Symposium on Abduction
Stanford, March 27-29, 1990, pp. 155-158.

Technical report R-145
February 1990

Probabilistic and Qualitative Abduction *

Judea Pearl

Computer Science Department
University of California
Los Angeles, California 90024

Abstract

This paper discusses relations between the probabilistic and qualitative approaches to abduction; it then offers a probabilistic account of the connection between causation and explanation, and proposes a non-temporal probabilistic semantics to causality.

1. Introduction

In the probabilistic approach, abduction is considered the task of finding the "most probable explanation" of the evidence observed, namely, seeking an instantiation of a set of explanatory variables that attains the highest probability, conditioned on the evidence observed. The qualitative approaches make explicit appeal to explanatory scenarios, and seek scenarios that are both coherent and parsimonious.

The major challenge for both the probabilistic and the qualitative approaches is to enforce an appropriate separation between the *prospective* and *retrospective* modes of reasoning so as to capture the intuition that prediction should not trigger suggestion. To use my favorite example: "Sprinkler On" predicts "Wet Grass," "Wet Grass" suggests "Rain," but "Sprinkler On" should not suggest "Rain." In the probabilistic approach such separation is enforced via patterns of independencies that are assumed to accompany causal relationships, cast in conditional probability judgments. In the qualitative approaches the separation is accomplished in two ways. One is to label sentences as either *causally established* (i.e., explained) or *evidentially established* (i.e., conjectured) and subject each type to a different set of inference rules [Pearl 1988a; Pearl 1988b; Geffner 1989]. The second method is to regard abduction as a specialized meta-process that operates on a causal theory [Poole 1987; Reiter 1987].

The obvious weakness of the qualitative approach is the lack of rating among competing explanations and, closely related to it, the lack of ratings of pending information sources. On the other hand, qualitative strategies demand fewer judgments in constructing the knowledge base.

In qualitative theories simplicity is enforced by explicitly encoding the preference of simple theories over complex ones, where simple and complex are given syntactical definitions, e.g., smallest number of (cohesive) propositions [Thagard 1989], minimal covering [Reiter 1987; Reggia et al. 1983]. These syntactic ratings do not always coincide with the notion of plausibility, for example, two common diseases are often more plausible than a single rare disease in explaining a given set of symptoms [Reggia 1989]. In probabilistic theories, coherence and simplicity are managed together by one basic principle — maximum posterior probability.

2. Explanation and Causation

Explanations are intimately connected to causation. When we say that "a explains b" we invariably assume the existence of a causal theory according to which "a tends to cause b" and, furthermore, that in the particular situation

* This work was supported in part by National Science Foundation Grant #IRI-88-21444 and Naval Research Laboratory Grant #N00014-89-J-2007.

where b was observed, "a actually caused b." The subtle difference between "tends to explain" and "actually caused" has been the subject of much discussion in the philosophical literature, a summary of which can be found in Skyrms & Harper [1988]. The classical example amplifying this difference is that of a skillful golfer who makes a shot with the intention of getting the ball in the hole; the shot is actually quite poor, but the ball hits a tree branch and is deflected into the hole. Here, we are likely to say that the golfer's skill and attention "tended to cause," but did not "actually cause" the ball to get in the hole. Explanation is connected with the latter, not the former; the phrase "tends to explain" is hardly in use in the language, instead, we use the phrase "is normally suggested by."

In the language of probability this distinction can be related to a difference between two conditional probabilities. If C has a tendency to cause E, then we expect $P(E|C)$ to be high. If C is identified as the event that "actually caused" E, then we expect $P(C|E, context)$ to be high where, by *context*, we mean other facts connected with the observation of E (e.g., hitting the tree in the golfer example).

In general, the probability $P(E|C)$ stands for a mental summary of a vast number of scenarios leading from C to E. Some of these scenarios involve contingencies such as trees intercepting golf balls, and some involve micro processes that can be articulated only at more refined levels of abstraction, for example, the interactions between the golf ball and the ground particles. When we confirm the sentence "C actually caused E" we normally mean that some path of contiguous micro events either can be presumed to have taken place or was actually observed. Such events are encoded in a knowledge strata more refined than the one used in the main discourse. For example, a pathologist may assert that the bullet was the "actual" cause of death only if a collection of key anatomical findings are observed confirming the existence of a contiguous physiological process leading from the bullet entry to death.

3. What's in an Explanation, a Probabilistic Proposal

If abduction is defined as "inference to the best explanation", a natural question to ask is how we define an explanation. Both the probabilistic and qualitative approaches to abduction have so far treated the term "explain" as a given primitive relationship among events, from which a "best" overall explanation is to be assembled. Both approaches have given the term "explain" a procedural semantics, attempting to match the way people use it in inference tasks, but were not concerned with what makes people believe that "a explains b," as opposed to, say, "b explains a" or "c explains both a and b." The quest for an empirical semantics of explanation has a long history in the literature of probabilistic causality, where the focus has been finding an operational definition of causation. (see Reichenbach [1956]; Simon [1957]; Good [1961]; Salmon [1984]; Suppes [1970]; Glymour et al. [1987]; Skyrms [1988]).

ith the exception of Simon [1957] and Glymour et al. [1987], temporal precedence was assumed to be essential for defining causation. For example, Reichenbach (1956, page 204) says that C is *causally relevant* to E if:

(i) $P(E|C) > P(E)$

(ii) There is no set of events *earlier* than, or simultaneous with, C such that conditional on these events E and C are probabilistically independent.

Suppes [1970] subscribes to a similar definition, with an explicit requirement that C precedes E in time.

These criteria offer a working definition for causation provided that the observed dependencies are not produced by *hidden* causes and provided that the set of events mentioned in condition (ii) is restricted to be "natural" events, excluding *artificial* events, syntactically concocted to meet condition (ii) [Good 1961; Suppes 1984].

I would like now to propose a non-temporal extension of the Reichenbach-Suppes definition of causation, one that determines the direction of causal influences without resorting to temporal information. It should be applicable, therefore, to the organization of concurrent events or events whose chronological precedence cannot be determined empirically. Such situations are common in the behavioral and medical sciences where we say, for example, that old age explains a certain disability, not the other way around, even though the two occur together (in many cases it is the disability that precedes old age). Similarly, we say that an incoming rain storm explains the falling barometer although, perceptually, the latter precedes the former in time.

The intuition behind my definition revolves around the perception of *voluntary control* [Simon 1980] and its probabilistic formulation in terms of conditional independence (see Pearl [1988], page 396). The reason we insist that the rain caused the grass to become wet and not that the wet grass caused the rain is that we can create conditions which, without disrupting the natural dependence between rain and wet grass, can get the grass wet without affecting the rain. We can, of course, also create a situation where the rain falls and the grass remain dry, say by seeding the clouds and covering the grass, but under such conditions the dependence between rain and wet grass is disrupted, which violates the symmetry between the two procedures.

As was stressed in Pearl [1988, page 396], the perception of voluntary control is not a necessary element in this asymmetry between cause and effect, but may in itself be a bi-product of dependencies observed among uncontrolled variables. In medical research, for example, we often search for a causal culprit of a disease much before attaining control over such cause.

Articulating these considerations in probabilistic terms, we come up with the following non-temporal extension of the Reichenbach-Suppes definition.

Definition: (non-temporal causation) An event C is said to be a (tentative) *direct cause* of E if

(i) $P(E|C) > P(E)$

(ii) There is no set of events such that conditional on these events E and C are independent.

(iii) There is an event C' and a set S of events not containing C, E and C' such that:

$$P(E|S,C') > P(E|S), \text{ and}$$

$$P(C|S,C') = P(C|S)$$

The set S in (iii) represents conditions needed for eliminating possible spurious dependencies between C and C'. Event C' represents our means for gaining control over E, namely, an event that can cause E without affecting C, thus providing an alternative explanation to E. Ironically, and almost circularly, explanations are defined in terms of their very destruction by other explanations; C qualifies as an explanation of E only if it can be "explained away" or rendered superfluous by some alternative explanation of C'. This is not surprising in view of the fact that people often seek an explanation for the sole purpose of ruling out others. For example, I often hope that my broker would explain the falling prices of my stock in terms of investors' panic and other transitory phenomena, so as to allay my fears of more profound explanations.

Any non-temporal definition of causation immediately raises the question of consistency, for example, is it possible that using criteria (i) through (iii) we would generate two incompatible assertions: "C cause E" and "E causes C?" It can be shown, however, that for a larger class of probability distributions these criteria are safe from such inconsistencies. Moreover, for those distributions that are unsafe, we can constrain (iii) by an additional restriction:

(iv) For every set of events S' that does not contain E and C, if there is an event E' (not in S') such that

$$P(C|S', E') > P(C|S'),$$

then

$$P(E|S', E') \neq P(E|S').$$

This restriction guarantees that we certify C as a direct cause of E only if the criterion (iii) is violated when we interchange C and E.

The definition above is a translation of that given in Pearl [1988b] to the language of Reichenbach and Suppes, where causes are propositional events having "positive" influence, hence the inequality in (i). In Pearl [1988b] these conditions were articulated in terms of *variables* rather than positively influencing *events*. A similar definition, in terms of variables, was introduced in Spirtes et al. [1989].

Another variant of this definition can be articulated using the graphical language of Bayesian networks, by considering all $n!$ orderings in which such a network can be constructed. We say that a variable C is a *direct cause* of variable E if:

(1) C and E are adjacent in all orderings, and

(2) There is an ordering in which C is a free parent of E, i.e., non-adjacent to some other parent of E, and there is no ordering in which E is a free parent of C.

This formulation reveals the type of empirical asymmetry that is responsible for evoking the perception of directionality in causal relationships.

On the practical side we also must address the question of computation complexity since, in principle, conditions (ii) and (iii) call for testing all subsets of events. It can be shown that, for a larger class of probability distributions, effective algorithms exist that determine the direction of causal influences without testing all subsets of events [Geiger 1990; Verma 1990].

A question of a more philosophical flavor concerns the relation between temporal precedence and the orientations determined by our definition: Why is it that we never observe a clash between the two? The answer, I believe, lies in the flexibility of our language; whenever the flow of dependency-based causality seems to clash with the direction of time we invent new variables (hidden causes) that reverse the former to comply with the latter.

References

Geffner, H. 1989. Default reasoning: Causal and conditional theories. Phd. dissertation. Cognitive Systems Laboratory *Technical Report (R-137)*. Computer Science Dept. University of California, Los Angeles.

Geiger, D. 1990. Graphoids: A qualitative framework for probabilistic inference. Phd. dissertation. Cognitive Systems Laboratory *Technical Report (R-142)*. Computer Science Dept. University of California, Los Angeles.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. 1987. *Discovering causal structure*. New York: Academic Press.

Good, I.J. 1961. A causal calculus. *British Journal for Philosophy of Science* Vol. 11:305-328; 12, 43-51; 13, 88; reprinted as Ch. 21 in: *Good Thinking* (University of Minnesota Press, Minneapolis, MN, 1983).

Pearl, J. 1988a. Embracing causality in formal reasoning. *Artificial Intelligence* 35(2):259-71.

Pearl, J. 1988b. *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufmann.

Poole, D. L. 1987. Defaults and conjectures: Hypothetical reasoning for explanation and prediction. Research Report CS-87-54, University of Waterloo (Kitchener, Ontario).

Reggia, J.A. 1989. Measuring the plausibility of explanatory hypotheses. *Behavioral and Brain Sciences* Vol. 12(3): 486-487.

Reggia, J. A., Nau, D. S., and Wang, Y. 1983. Diagnostic expert systems based on a set-covering model. *Intl. Journal of Man-Machine Studies* 19: 437-60.

Reichenbach, H. 1956. The direction of time. Berkeley, CA: University of California Press.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32(1): 57-95.

Salmon, W. 1984. Scientific explanation and the causal structure of the world. Princeton: Princeton University Press.

Simon, H. 1957. *Models of man*. New York: Wiley and Sons.

Simon, H.A. 1980. The meaning of causal ordering. In *Qualitative and quantitative social research*, ed. R. K. Merton, J. S. Coleman, and P. H. Rossi, 65-81. New York: Free Press.

Skyrms, B. 1988. Probability and causation. *Journal of Economics* Vol. 39, 53-68.

Skyrms, B. and Harper, W.L. 1988. *Causation, chance and credence*. Dordrecht: Kluwer Academic Publisher.

Spirtes, P. 1989. Causality from probability. Dept. of Philosophy, Carnegie-Mellon University, Report #CMU-LCL-89-4.

Suppes, P. 1970. *A probabilistic theory of causality*. Amsterdam: North Holland.

Suppes, P. 1984. *Probabilistic metaphysics* Oxford: Blackwell.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* Vol. 12(3): 435-468.

Verma, T. 1990. Learning causal structure from independence information. (in preparation).