Learning Causes: Psychological Explanations of Causal Explanation[1]

CLARK GLYMOUR
University of California at San Diego and Carnegie Mellon University

Abstract. I argue that psychologists interested in human causal judgment should understand and adopt a representation of causal mechanisms by directed graphs that encode conditional independence (screening off) relations. I illustrate the benefits of that representation, now widely used in computer science and increasingly in statistics, by (i) showing that a dispute in psychology between 'mechanist' and 'associationist' psychological theories of causation rests on a false and confused dichotomy; (ii) showing that a recent, much-cited experiment, purporting to show that human subjects, incorrectly let large causes 'overshadow' small causes, misrepresents the most likely, and warranted, causal explanation available to the subjects, in the light of which their responses were normative; (iii) showing how a recent psychological theory (due to P. Cheng) of human judgment of causal power can be considerably generalized: and (iv) suggesting a range of possible experiments comparing the cecd answe-26.

I am sure that, like everyone else, Shanks learned that correlation is not causation, but his second sentence collapses the distinction and confounds learning associations with learning how to predict and control. Knowing only the association

Minds and Machines **8:** 39–60, 1998.

between *A* and *B* doesn't usually enable us to control either *A* or *B*. The association between yellowed fingers in youth and middle age and lung cancer in later life doesn't of itself provide a way to control lung cancer. Preventing yellow fingers will do nothing to change the frequency of lung cancer unless the intervention also changes smoking frequency – making everyone wear gloves, for example, will do

In what follows I describe the elements of a representation of causal explanations – a representation now almost standard in computer science and increasingly common in statistics – that provides the basis for algorithmic solutions to these puzzles; that is, mathematical work using the representation shows how (and what) causal information may be reliably extracted from observed associations, and how that usually incomplete causal information can be used in prediction and planning.

This work has at east four kinds of implications for psychology: methodological, interpretive, analytic and substantive. The methodological issues have principally to do with old fashioned but still relevant problems, such as the justification of 'intervening variables,' and with entirely contemporary issues about techniques of data analysis and theory construction (by psychologists, not their subjects). The interpretive issues have to do with understanding the confusions in false dichotomies between 'associationist' and 'mechanist' accounts of causation, confusions so influential they threaten to eliminate from psychology any serious work on the subject of causation, and with the interpretation of experiments intended to assess whether, how much, and why, human judgment of causal relations is sub-normative. The analytical issues have to do with unfolding the hidden implications of contemporary psychological theories when they are translated into the new representation. The substantive issues have to do with the main puzzle implicit in Shanks' abstract: how do humans extract the available causal information from associations?

Of the four kinds of implications, I will ignore the methodological topics in this paper, but I will illustrate each of the others. The substantive issue, which in my view is the most interesting and important, I will deal with last.

2. Mystery "Mechanism": An Answer Too Many Psychologists Like

Possibly the most popular (among psychologists: see Ahn and Bailenson, 1996; Ahn et al., 1995; Baumrind, 1983; Schultz, 1982; White 1989, 1995; for philosophers of the same opinion see; Harré and Madden, 1975; Turner, 1987) answer to the questions I extracted from Shanks' abstract denies their presupposition: people *don't* learn causes from associations, because causes have nothing to do with associations, they have to do with 'mechanisms.' What is meant by 'mechanism' is rarely explained in this literature, but the examples make it relatively clear that to specify a 'mechanism' for a covariation is simply to specify either a sequence of causes that intervene between the candidate cause and effect, or causes that tend to bring about both the candidate cause and effect, where the causal connection posited in the 'mechanism' are of a kind that are already familiar and acknowledged. Baumrind (1983) gives the following illustration:

> The number of never-married persons in certain British villages is highly inversely correlated with the number of field mice in the surrounding meadows.

> elders until the mechanisms of transmission were finally surmised: Never-married persons bring with them a disproportionate number of cats.

Similar examples are offered by Ahn et al. (1995) and others. Mechanisms of this kind can be represented by a causal diagram or *directed graph* (i.e., a network of nodes representing features or variables and with arrows pointing from causes to effects), for example

$$\#\text{ unmarried persons} \rightarrow \#\text{ cats} \rightarrow \#\text{ mice}$$

There are two sorts of probabilistic consequences to this sort of mechanism: the mechanism implies relations among conditional probabilities, and the mechanism implies probability relations upon various interventions. The two sorts of probability implications will sometimes, as in this case, be equal, but they are not the same. In Baumrind's example, if hers is the entire mechanism behind the association, then if we were to *intervene* to hold the number of cats constant in these villages, there would be no frequency association between variations in the number of unmarried persons and variations in the number of mice. And under the same assumption, if we did not intervene at all, but simply computed the conditional probability of any number of mice

If, separately, we knew, for example, that the number of cats does not cause the number of unmarried persons, we could eliminate all but explanations 3 and 4. Both of these explanations suppose that the number of cats influences the number of mice and we could also conclude that *either* the number of unmarried persons influences the number of cats, *or* something else influences both.

So the separation of mechanisms and associations is very odd and implausible, and, to the contrary, it seems that an important part of learning causes might very well be learning mechanisms from associations together with prior knowledge. Later we will see that these inferences can be made rigorous.

Besides Baumrind's example, consider briefly the mechanism that generates the covariation between past occurrences of yellow fingers and the later occurrences of lung cancer among those who grew up in the days of unfiltered cigarettes. The mechanism behind the covariation is a common cause: smoking caused yellowed fingers and it also caused lung cancer:

yellowed fingers $\leftarrow$ smoking $\rightarrow$ lung cancer

Here again, there are two distinct kinds of probabilistic implications of the explanation. First, yellowed fingers and lung cancer are independent conditional on smoking. (More generally, when there are no other causal connections, the effects of a common cause are independent conditional on a value of the common cause (Simon, 1977)). Second, interventions that directly alter only the frequency of

compare terrestrial bodies. Mostly they don't fall or move, they just stay where they are. The intellectual ancestors of Ahn et al. would presumably have concluded that the study of motions of bodies must be marginal.

3. Another Answer: Conditioning

Classical and operant conditioning both produce an appropriate expectation in a learner, but they differ in what the learner discovers about control. In classical conditioning a subject learns an association between two kinds of events, neither of which are interventions or actions of the learner. In operant conditioning, a subject learns an association between two kinds of events, one of which is an action of the learner. In classical conditioning, unless there is other relevant knowledge, all that can be learned is an associati-us(whi.0(nawl)7(e)-1(dge,4)-293(i82(L)e'i)7(t758(r)9(l-us(whi.0(nawl)7(

4. Representation

A normative account of causal inference requires a representation of causal relations general enough to include, to good enough approximation, the great majority of causal systems we think we encounter, and a characterization of the information about causal relations that can be extracted from observed associations, or from observed associations and prior knowledge of various kinds. It does not require an *analysis* of causation or a representation or an account of inference that covers every imaginable case. (Much of the poverty of contemporary philosophy results from insisting on perfect generality, thereby avoiding the effort of investigating any unobvious consequences of any assumptions, since none are perfectly general.) What follows in this section is not mysterious, and in many respects not even difficult. It requires some patience with formal definitions and distinctions, and some elementary modern mathematics. It is an abbreviated description of the representation of causal mechanisms that has become almost standard in computer science, and is implicitly used throughout much of applied statistics. The payoff is astonishing.

In discussing the mechanist view, and Baumrind's example in particular, I introduced diagrams with nodes indicating features of a system and an arrow, or directed edge, from one node to another indicating that the feature represented by the node at the tail of the edge or arrow is a direct (relative to the features represented in the diagram) cause of the feature represented by the node at the head of an arrow or directed edge. Diagrams such as these are *directed graphs*, and carry with them obvious notions – a node at the head of an arrow is the *child* of a node at the tail, which is its *parent*; some nodes are *ancestors* of others, their *descendants*; there are *paths* from ancestors to descendants, and so on.

I will say a system *S* of variables is *causally sufficient* provided that for every pair of variables

In the directed graph for Baumrind's example, the parent set of number of mice is {number of cats} and the set of all non parent, non-descendants of number of mice is {number of unmarried persons} so the Markov condition gives exactly the conditional independence I asserted.

requirements of the associations and conditional independence will be satisfied provided a is negative, b is positive, c is negative, and $a = -bc$.

In the ignorance I imagined, this explanation will save the phenomena, but it seems inordinately – even unscientifically – complex, and in the absence of prior knowledge I think most people would reject it in preference to explanations 3 or 4 above. Notice that the conditional independence does not result from the Markov condition applied to the causal graph just given – the Markov condition applied to that graph gives no independencies whatsoever. And that observation leads to a general formulation of the intuition about simplicity the example illustrates:

> *Faithfulness Condition*: For any variables X, Y and any set of variables Z in a causally sufficient system described by a directed acyclic graph G, X is independent of Y conditional on Z if and only if the Markov condition applied to G implies that X is independent of Y conditional on Z.

Faithfulness is easier to evade than the Markov condition, but not very easy. Both for linear stems and for systems of variables each having only a finite number of values, 'almost all' probability distributions that satisfy the Markov condition for a directed acyclic graph also satisfy the faithfulness condition.[3]

The Faithfulness assumption seems to bitterly divide scientists, even when they have not formulated it explicitly. The late, eminent sociologist, Hans Zeisel, resigned from the board of supervisors of an experiment funded by the Department of Labor when the principals of the experiment saved their favorite hypothesis by forwarding an unfaithful explanation of the data. The principals in turn mounted a rather vicious *ad hominem* attack on Zeisel. (See Glymour et al., 1987 for a review and references). In cognitive psychology, recent disputes over unconscious mech-

3

This will sound ludicrous to any philosopher, statistician or social scientist familiar with confounding, but, with exceptions to be noted later, the community hmt0 c

Suppose *e* has a cause *i*, and let *a* represent all other causes of *e*. Assume *e* does not occur unless at least one of its causes occurs. Cheng reasons that the probability that *e* occurs given that *i* occurs is the probability that *i* causes *e* given that *i* occurs, plus the probability that *a* occurs given that *i* occurs times the probability that *a* causes *e* given that *a* occurs and *i* occurs, minus the probability that *a* occurs given that *i* occurs times the probability that both *a* and *i* cause *e* given that *a* and *i* both occur. She assumes the probability – for reasons that will be clear later, I denote it $P(q_{ae})$ – that a causes *e* given that *a* occurs is independent of whether *i* occurs, and likewise the probability, $P(q_{ie})$, that *i* causes *e* given that *i* occurs is independent of whether *a* occurs, and, further, that the probability $P(q_{aie})$ that *a* and *i* both cause *e* given that both occur equals $P(q_{ae})P(q_{ie})$. Hence she derives

$$\text{prob}(e = 1|i = 1) = P(q_{ie}) + P(q_{ae})\text{prob}(a = 1|i = 1) - P(q_{ie})P(q_{ae})\text{prob}(a = 1|i = 1) \quad (1)$$

which shows immediately that $P(q_{ie})$ is a conditional probability, specifically the probability that e occurs given that i occurs and a does not occur. *Cheng 's model of the power of a cause i to produce an effect e is, in this setting, the probability of e given that i occurs and that no other cause of e occurs.*

When *i* = 0

$$\text{prob}(e = 1|i = 0) = P(q_{ae})\text{prob}(a = 1|i = 0) \quad (2)$$

Now if *a* and *i* are independent, she deduces

$$\text{prob}(e = 1|i = 1) = P(q_{ie}) + P(q_{ae})\text{prob}\ (a = 1) - P(q_{ie})P(q_{ae})\text{prob}(a = 1)$$

and

$$\text{prob}(e = 1|i = 0) = P(q_{ae})\text{prob}(a = 1)$$

Hence

$$\Delta P = \text{prob}(e = 1|i ==$$

subjects are asked to estimate the power of a facilitating (rather than inhibiting) cause *i*, and they are given reason to think i and all other causes *a*, of *e*, are independent, they estimate $P(q_{ie})$.[4]

Cheng's is a theory, one of the few I know of, that at least for special cases – binary variables and direct causes of an effect – addresses the subject's model of causal structure, the relations of that causal structure to probabilities, and the aim of judgements of causal power. And, more to the good, it does not require of subjects extraordinary computational powers, tacit or explicit. The aim it supposes has a natural justification: the probability that $e = 1$ given $i = 1$ and all other causes a are absent does not depend on the frequency of other causes *a*, and so does not depend

The parameter $P(q_i = 1)$, for example can be estimated by

$$P(q_{ie} = 1) = P(e = 1|i = 1, a = 0)$$

or, when *a* is unobserved, by Cheng's formula,

$$P(q_{ie}) = \frac{\Delta P}{1 - \text{Prob}(e = 1|i = 0)} \tag{6}$$

Cheng's model of human judgment of positive causation *just is* a directed acyclic graph parameterized as a noisy-or-gate.

So what?

probability is undefined. In case (ii), the parameter $P(q_{xz})$ is not this probability at all but rather the *probability that Z* = 1 *given that* Y = 0 and U = 0 *and given an intervention to bring about X* = 1. An intervention that forces the value 1 on *X*, regardless of the value of *Y*, transforms the causal structure of case (ii) into the causal structure of case (i). In case (i) the probability (of *Z* = 1, etc.) conditional on *X* = 1 is equal to the probability given an intervention to bring about *X* = 1, but in case ii the quantities are distinct. A general theory of the transformations of causal structures and probabilities under ideal interventions is given in Spirtes (1993) and Pearl (1995).

Cases iv and vii pose another difficulty. Cheng's measure of the causal power of X to produce Y – the probability that *Z* =1 given (an intervention to produce) *X* = 1 and *all* other causes of *Z* have value 0 – is necessarily zero in case vii, because *X* only influences *Z* through effects of *X* (namely *W* and *Y*), which in turn are causes of *Z*. That suggests that an alternative measure of the causal power of *X* to produce *Z* might be the probability that *Z* = 1 given that *X* = 1 and that all other causes of *X that are not effects of X* are 0. Call the original measure *direct* causal power and the new measure *total* causal power. In case iv, the two measures of causal power are distinct, but both non-zero. In these cases any simple request for a judgment of causal power is ambiguous. In the other cases the two measures are equal.

The causes of *e* that are not effects of *c* are *f*, *h* and *g*. According to the proposition the total causal power of *c* to bring about *e* is:

$P(e = 1|f = 0, g = 0, h = 0, c = 1$ by intervention,$) = P(q_{cb}q_{be} \bigoplus q_{cd}q_{de} = 1)$
$= [P(e=1|c=1, h=0) - P(e=1|c=0, h=0)]/(1 - P(e=1|c=0, h=0)]$.

The r.h.s. of this equation is conditioned on *h*=0 because h is an observed cause of *e* but not a descendant of *e*. Recall that by the convention used in directed graphs, *f* and *g* are independent of *c*. Here is a derivation:

$$\begin{aligned} e &= q\ g \bigoplus q_b b \bigoplus q_d \\ &= q\ g \bigoplus q_b(q_f f \bigoplus q_{cb}c) \bigoplus q_d q_h h \bigoplus q_d q_{cd} c \\ &= (q\ g \bigoplus q_b q_f f \bigoplus q_d q_h h) \bigoplus C(q_b q_{cb} \bigoplus q_d q_{cd}) \end{aligned}$$

When $c = 1$ by intervention, and $h = 0$:

$$e = (q\ g \bigoplus q_b q_f f) \bigoplus (q_b q_{cb} \bigoplus q_d q_{cd})$$

When $c = 0$ and $h = 0$:

$$e = q\ g \bigoplus q_b q_f f$$

Let $\Delta P_{h=0} = P(e = 1|c = 1$ by intervention, $h = 0) - P(e = 1$

As to the first of these questions, there are experiments that support the hypothesis that people estimate causal power in accord with Cheng's model in cases (i) and (ii) above, but what happens (when the aim of judgement is disambiguated between direct and total causal power) in cases such as (iv) through (ix) does not seem to be known.

The second question has been addressed, but in a remarkably limited way. The psychological literature focuses on experiments that provide the subject with sufficient context – sufficient prior knowledge about causal structure, that only the value of a single parameter remains to be learned. Typically, some special value of that parameter corresponds to the absence of a single edge in the causal graph, an edge whose direction, if it exists, is already known to the subject. This may in fact be the only way *humans* extend their knowledge of causal structure, or it may not. It is certainly not the only possible way.

Consider the following example: data are available on the associations of four variables *X, Y, Z, W*. Nothing is otherwise known about the time order or causal relations among these variables, except that the values of these variables for a unit did not influence whether the unit was sampled. Suppose the associations show the following pattern:

X and Y are independent

X and Y are independent of W conditional on Z

No other independencies hold

Then, under the assumptions previously discussed, it follows necessarily that *Z* causes *W*. That is, *every* directed graph (and probability distribution), together satisfying the Markov and Faithfulness conditions, and which implies the three features just listed, contains a directed edge

$Z -> W.$

It doesn't matter whether the graph does or does not contain unobserved common causes of any pair of *X, Y, Z, W*, it must contain the edge from *Z* to *W*.

Or consider the following example: data are available on the associations of four variables *X, Y, Z, W*. Nothing is otherwise known about the time order or causal relations among these variables, except that for no unit did the values of these variables influence whether the unit was sampled. Suppose the associations show the following pattern:

X is independent of $\{Z,W\}$

W is independent of $\{X,Y\}$

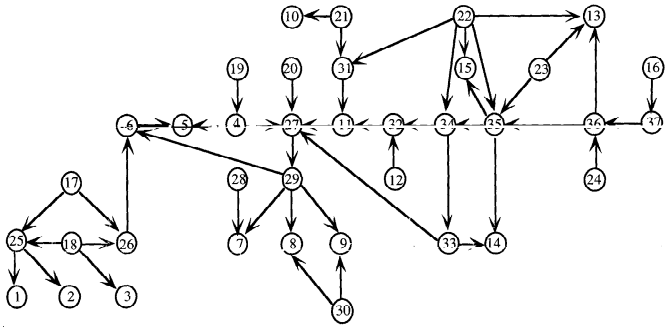
No other independencies hold

Then, under the assumptions previously discussed, it follows necessarily that there is a common cause of *Y* and *Z* other than *X* or *W*. That is, *every* directed graph and probability distribution, together satisfying the Markov and Faithfulness conditions, which implies the three features just listed contains another vertex, call it *U*, and a pair of directed paths from *U* to *Y* and to *Z*.

These simple examples illustrate that associations alone, under appropriate assumptions, sometimes suffice to determine causal connection, the direction of causation, and even the presence of unobserved or unnoticed causes. With stronger assumptions about prior knowledge, still more can be learned from the patterns of independencies and dependencies. For example, if it is known that there are no unobserved common causes of measured variables, then everything about the causal graph can be learned, except that any two graphs that have all of the same 'unshielded collider' structures – for example,

$$X \rightarrow Y \leftarrow Z$$

will be indistinguishable. For example, without the use of any prior causal information, computer algorithms are able to infer from associations most of a model of an emergency medical system show below (taken from Beinlich et al., 1989)



KEY:

1 – central venous pressure
2 – pulmonary capillary wedge pressure
3 – history of left ventricular failure
4 – total peripheral resistance
5 – blood pressure
6 – cardiac output
7 – heart rate obtained from blood pressure
8 – heart rate obtained from electrocardiogram
9 – heart rate obtained from oximeter
20 – insufficient anesthesia or analgesia
21 – pulmonary embolus
22 – intubation status
23 – kinked ventilation tube
24 – disconnected ventilatio tube
25 – left-ventricular end-diastolic volume
26 – stroke volume monitor
27 – catecholamine level
28 – error in heart rate reading due to low cardiac output

The causal graph encodes the conditional independencies that the experts specified among these variables; the search algorithms find the graphs that explain the resulting patterns in the data.

There are a number of algorithms in the computer science literature, using a variety of techniques, that recover causal structure from associations by taking advantage of these relationships between causal structure and patterns of constraints on association. The computational complexity of the procedures depends on the complexity of connections in the causal graph generating the data – strictly, on how many parents each variable has, on average. For sparse graphs the procedures are very fast, and with good data can recover a great deal of structure quite reliably. Whether humans can do the same, at least in simple cases, is essentially unknown. Only one experiment has been reported. Hashem and Cooper, 1996, gave medical students information about associations for a variety of cases involving two and three binary variables, described as disease or gender features. The subjects were asked, essentially, to recover the causal graph from the associations, and their responses were compared with those of a Bayesian inference algorithm using the same associations. The subjects did poorly in problems with three variables, but the result is uncertain for two reasons. First, because of sample size and the exclusive use of binary variables, spurious near-independencies held in the data given the subjects, so that in important cases the Bayesian search algorithm performed comparably poorly. Second, a lot of experience in psychological experiments suggests that humans do much better with frequency and independence judgements when they are not given numbers, but instead actually observe the events or can see the frequencies displayed graphically, or both.

Besides association and short of experimental intervention, the cue to causation most commonly available to us is the order of occurrence of events. Causes do not come after effects. Knowledge of time order considerably speeds up algorithmic search for structure and increases the reliability of output as well. Perhaps more important for understanding human inference, knowledge of time order may compensate for circumstances in our environment in which the Faithfulness assumption does not hold. It may be that for many of the causal relations of everyday life, relationships among observed features or variables are nearly deterministic. Faithfulness fails to hold in many deterministic systems, for technical reasons I will pass by. But when time order is known, the Faithfulness assumption is not needed for inference to causal structure in systems of deterministically related observed variables; in those contexts Faithfulness can be replaced by the weaker assumption that multiple mechanisms relating two variables do not perfectly cancel one another.

10. Conclusion

For computers anyway, there is an answer to the puzzle posed by Shanks' claim that causes are learned from associations. The answer–algorithms that represent

causal structure by networks or directed graphs and infer aspects of that structure from data about frequencies, relying on very broad assumptions connecting causal structure with conditional probabilities–raises a host of unexamined issues for experimental psychology. The same representation also provides a mathematical tool for generalizing and analyzing leading theories of human Judgement of causal power, provides the means to see and articulate important distinctions about causal influence – distinctions that, if not recognized, easily lead to erroneous interpretations of psychological experiments – and provides a coherent norm against which to measure human judgement.

Notes

1

Glymour, C. and Cheng, P. W. Causal Mechanism and Probability: A Normative Approach', in M. Oaksford and N. Chater (eds.), *Rational Models of Cognition*. Oxford, U.K.: Oxford University Press (in press).
Harré, R. and Madden, E.H. (1975), *Causal Powers: A Theory of Natural Necessity*, Totowa, New Jersey: Rowman & Littlefield.
Hashem, A.I. and Cooper, G.F. (1996), 'Human Causal Discovery From Observational Data'. *Proceedings of the 1996 symposium of the American Medical Information Association*.
Jacoby, L., Yonelinas, A. and Jennings, J., (1997), The Relation Between Conscious and Unconscious (Automatic) Influences: A Declaration of Independence, in J. Cohen and J. Schooler (eds.), *Scientific Approaches to Consciousness*., Mahwah, N.J., Lawrence Erlbaum Associates, pp. 13–47.
Jenkins, H. and Ward, W. (1965), 'Judgment of Contingency Between Responses and Outcomes'. *Psychological Monographs* 7, pp. 1–17.
Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California: Morgan Kaufmann.
Pearl, J. (1995), 'Causal Diagrams for Empirical Research', *Biomtrika* 82(4), pp. 669–709.
Price, P.C. and Yates, J.F. (1993), 'Judgmental Overshadowing: Further Evidence of Cue Interaction in Contingency Judgment'. *Memory & Cognition* 21, pp. 561–572.
Rescorla, R.A. (1968), 'Probability of Shock in the Presence and Absence of CS in Fear Conditioning'. *Journal of Comparative and Physiological Psychology* 66, pp. 1–5.
Shanks, D.R. (1995), 'Is human learning rational?' *Quarterly Journal of Experimental Psychology*, 48A, pp. 257–279.
Shultz, T.R. (1982), 'Rules of Causal Attribution'. *Monographs of the Society for Research in Child Development*, 47, (1).
Spellman, B.A. (1996a), 'Acting as Intuitive Scientists: Contingency Judgments are Made While Controlling for Alternative Potential Causes'. *Psychological Science*, 7, pp. 337–342.