Causal Mechanism and Probability:
A Normative Approach

Clark Glymour
University of California, San Diego
&
Carnegie Mellon University

Patricia W. Cheng
University of California, Los Angeles

Abstract

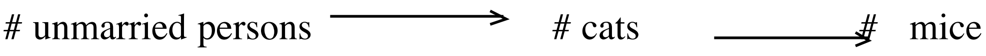
The rationality of human causal judgments has been the focus of a great deal of recent research. We argue against two major trends in this research, and for a quite different way of thinking about causal mechanisms and probabilistic data. Our position rejects a false dichotomy between "mechanistic" and "probabilistic" analyses of causal inference -- a dichotomy that both overlooks the nature of the evidence that supports the induction of mechanisms and misses some important probabilistic implications of mechanisms. This dichotomy has obscured an alternative conception of causal learning: for discrete events, a central adaptive task is to induce causal mechanisms in the environment from probabilistic data and prior knowledge. Viewed from this perspective, it is apparent that the probabilistic norms assumed in the human causal judgment literature often do not map onto the mechanisms generating the probabilities. Our alternative conception of causal judgment is more congruent with both scientific uses of the notion of causation and observed causal judgments of untutored reasoners. We illustrate some of the relevant variables under this conception, using a framework for causal representation now widely adopted in computer science and, increasingly, in statistics. We also review the formulation and evidence for a theory of human causal induction (Cheng, 1997) that adopts this alternative conception.

1. The Old Mechanism Approach

A long and still popular tradition in the study of human causal reasoning insists on a dramatic bifurcation between "mechanistic" conceptions of causal inference and "probabilistic" or "covariational" conceptions of this process (e.g., Ahn & Bailenson, 1996; Ahn, Kalish, Medin & Gelman, 1995; Baumrind, 1983 [Clark, please supply reference]; Harré & Madden, 1975; Turner, 1987; Schultz, 1982; White, 1989, 1995). What is meant by "mechanism" is rarely specified in this literature, but the examples given 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 connections 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. [Marriage] was considered an established cause of field mice by the village 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 graph with nodes and arrows pointing from causes to effects), for example



Those who frame issues in term of this dichotomy are themselves proponents of the mechanistic approach. Although these researchers do not explicitly state that theirs is the more normative approach, such appears to be their tacit assumption (as should be clear from Baumrind's example).

There are probabilistic consequences to this sort of mechanism. In Baumrind's example, the number of unmarried persons is independent of the number of mice conditional on the number of cats. More generally, remote nodes in a causal chain are independent conditional on any set of values of the intervening causes (Pearl, 1988; Spirtes, Glymour, & Scheines, 1993). And by almost any measure of covariation, the (negative) covariation between unmarried persons and mice should be weaker than the (negative) covariation between cats and mice. Or consider the mechanism that generates the covariation between past occurrences of yellow fingers and the later occurrences of lung cancer among people of the age of the first author, 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 ← smoking → lung cancer

Here again, there is a probabilistic implication of the explanation: yellowed fingers and lung cancer are independent conditional on smoking. More generally, the effects of a common cause are independent conditional on a value of the common cause (e.g., Simon, 195?; Pearl, 1988; Waldmann, Holyoak, & Fratianne, 1995).

As these examples illustrate, there are intricate connections between mechanisms and probabilistic patterns, and a fruitful mechanistic approach to understanding human judgment about causation might try to understand those patterns and investigate the ways humans use them, or can learn to use them, to infer causal mechanisms~~in learning and planning~~. But, a disconnection between mechanisms, on the one hand, and probabilistic patterns, on the other, puts everything on a false footing.

Those who contrast mechanistic and probabilistic analyses of causation are not concerned with understanding these intricate connections between probability relations and mechanisms. We will take Ahn et al. (1995) as an example, although many others are roughly equivalent. Our aim is not to examine their paper in any detail, but instead to present general arguments not specific to their paper, that show why their results do not tell against a probabilistic approach.

The substance of their views seems to have several components:

1. People are more likely to judge a covariational relation to be causal when they can know, or can plausibly conjecture, a "mechanism."

2. When asked for an explanation of how a cause produces an effect, people are likely to propose a "mechanism."

3. When free to seek information in order to decide whether a relation is causal, people ask for information about the "mechanism" rather than for probabilistic information about covariations.

4. Information on covariation is generally not necessary for learning causal relations. With rare exceptions, people do not learn causal relations from covariations but from applying prior knowledge of "mechanisms."

Given that "mechanisms" necessarily reflect ~~prior~~ knowledge about causal connections, theses 1 and 2 are consistent with almost any sensible theory of human judgment of causation. When the question is whether a covariation between candidate cause c and effect e is due to the influence of c on e (as against, for example, chance, or a common cause of both, or sample selection bias), it is only rational to give greater weight to answers that are coherent with other~~prior~~ knowledge. And what could an explication of a causal connection *be* except the specification of intervening or confounding causes? [Clark, I don't understand why you want to include confounding causes when the question is how a cause produces an effect (#2 above) rather than whether a regularity is causal.] These two theses make sense as refutations of the probabilistic approach, as proponents of the dichotomy apparently intend, only if one assumes that researchers who adopt the probabilistic approach -- whose goal is to study how causal beliefs emerge from probabilistic information -- deny that people have a conception of causality. We are not aware of any psychologist or philosopher who denies this evident fact.

Establishing the third mechanistic thesis is one the chief aims of the experiments reported by Ahn et al. (1995). This thesis is irrelevant to the probabilistic approach, however, because this approach does not aim at addressing the issue of how frequently people conjecture that the causal knowledge relevant to a novel situation is familiar to them. Instead, the probabilistic approach is concerned with the origin of such knowledge whenever it is acquired. If a reasoner has already acquired a considerable amount of causal knowledge, it is eminently plausible that he or she might check whether some of this knowledge applies in a novel situation, rather than acquiring that knowledge afresh from probabilistic information on each new encounter. Causal induction would not be of much use if the induced causal knowledge cannot be applied subsequently. ~~Contrary to an apparent motivation for the dichotomy between the two types of information people seek, researchers who adopt the probabilistic approach have no reason to~~ Unless probabilistic models assume complete forgetting, seeking information about “mechanisms” rather than probabilistic information is consistent with such models, although irrelevant to their evaluation.

The fourth mechanistic thesis leads to either a theoretical vacuum or the conclusion that all causal knowledge is innate.~~is theoretically problematic with regard to the origin of causal knowledge~~. Consider the links that must be interpolated between candidate cause and effect according to the mechanistic account. Each is a piece of causal knowledge, which often the reasoner cannot subdivide into still further links. The mechanism view cannot possibly require that causal relations always have sub-mechanisms known to the reasoner, in infinite regress. Consistently, it can only require that causal relations have intervening links that reduce eventually to some basis set known to the reasoner. How, on the mechanistic account of the human understanding of causation, are these fundamental causal relations known? Either, it seems, they must be known *a priori*--innately--or they must be learned. If learned, they must be learned without imposing on them the requirement of further mechanistic explanation, and thus by procedures that the mechanistic conception does not illuminate. By rejecting the thesis that causal relations are ultimately learned through the use of observable information such as probabilities, the mechanistic account advocates a vacuum as the answer to the question of causal learning. If *a* *priori*, then the mechanistic account must then embrace the view that all possible causal relations are known *a priori*, at least implicitly, and must hold that people learn only two things about causation: the consequences of what they already know, and which relations are instantiated where in the world. Cognitive psychology becomes fully Platonic, and the model for developmental psychology becomes the *Meno*.

The most serious objection to the fourth thesis, however, is that it overlooks the likely possibility that people can and do learn new causal relations by giving a causal interpretion to frequency information under constraints provided by prior knowledge. Theoretical and experimental explorations of that possibility require an understanding of how such inferences are possible in principle. That is the subject of the rest of this paper.

2. Evaluating the Normativeness of Contingency by Causal Mechanisms

The most common "normative" standard used in psychology for evaluating the influence of a candidate cause c on an effect e (Allan, 1980; Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Chapman & Robbins, 1990; Price & Yates, 1993; Rescorla, 1968; Shanks, 1991) is the "contingency" of e on c:

$$\Delta P = P(e \mid c) - P(e \mid {\sim}c),$$

where $P(e \mid c)$ is the probability of e occurring given the presence of c and $P(e \mid {\sim}c)$ is the probability of e occurring given the absence of c. ΔP is a measure of deviation from independence: e and c are independent in probability when and only when $\Delta P = 0$. However, there are contexts in which ΔP is clearly an inappropriate estimator. These contexts have to do with the mechanism, or causal process, by which the covariation between c and e comes about. For example, consider again the past occurrences of yellowed fingers and the later occurrences of lung cancer.

Smoking caused yellowed fingers and it also caused lung cancer. The probability of lung cancer given yellowed fingers was greater than the probability of lung cancer given not-yellowed fingers, and not equal to the probability of lung cancer averaged over those with and without yellowed fingers. But the probability of lung cancer if *everyone* had been required--and so acted--to wear gloves throughout the waking day, thus having not-yellowed fingers, would have been the same as the probability of lung cancer originally, because acting to prevent smoking from discoloring fingers, without acting to prevent smoking, would have no effect on the likelihood of lung cancer. Notice that from the perspective of the unlikely actuary, who has information about the distribution of yellowed fingers in the population, and is interested only in predicting who will and will not die,

ΔP = P(lung cancer | yellowed fingers) – P(lung cancer | not-yellowed fingers)

is an informative quantity. But, from the perspective of the Surgeon General, who is interested in what causes lung cancer, and in the changes in the probability of lung cancer that various interventions would bring about, ΔP is irrelevant. Quantities more relevant to these causal questions are the probabilities of lung cancer conditional on yellowed fingers and smoking, and conditional on non-yellowed fingers and smoking, and the *conditional contingencies*,

$\Delta P_{\text{Smoking}}$ = P(lung cancer | yellowed fingers, smoking) – P(lung cancer | not-yellowed fingers, smoking)

$\Delta P_{\text{Not-smoking}}$ = P(lung cancer | yellowed fingers, not-smoking) – P(lung cancer | not-yellowed fingers, not smoking)

both of which are zero. If the surgeon general wanted to predict the effects on lung cancer of altering the probability of yellowed fingers without changing the probability of smoking, these are relevant quantities, because (as we explain later) the zero conditional ΔPs reveal that the association is produced by a common cause of yellowed fingers and lung cancers, and not by any direct influence of one on the other.

In the example just given, ΔP is irrelevant to assessing the influence of yellowed fingers on lung cancer because a third variable, smoking, is a common cause of both. But, depending on context, ΔP can also be irrelevant when the third variable is an effect rather than a common cause. A textbook case (Freedman, Pisani & Purves, 1978) involves the admission of women to graduate school at the Universtiy of California, Berkeley. It was found that the rejection rate for women was much higher than that for men, suggesting a systematic bias against women in admission policies. Further investigation showed, however, that women applied more often than men to programs that had lower than average acceptance rates regardless of the gender of the applicants. Conditional on program applied to, gender and rejection were nearly independent. In judging the influence of sexism on admission decisions in this context, ΔP is clearly the wrong quantity. Arguably the conditional ΔP of rejection rate on gender, controlling for graduate department, would be more relevant.

Notice that the conditional contingencies are only relevant to predicting the effects of an intervention in these cases because they happen to reflect the particular mechanism involved. In still other cases both the contingency and the conditional ΔP might be irrelevant. Suppose that parents' intelligence influences child's intelligence and both variables influence the years of education the child receives. Then the conditional ΔP between parental and child intelligence, controlling for child's education, would be the wrong quantity altogether to use to estimate the influence of parents' intelligence on child's intelligence. Conditionalizing on a common effect yields a probabilistic dependency between alternative causes of the effect even when no causal connection exists between these causes.

Some psychologists have assumed that ΔP is the optimal probabilistic model of causal induction and tested it as a model of human causal inference (Baker, et al., 1993; Chapman, 1991; Chapman & Robbins, 1990; Price & Yates, 1993; Shanks, 1991). They concluded that observed causal judgments deviate systematically from ΔP. The findings in these experiments, which involve situations with multiple varying candidate causes, can be and indeed have been reinterpreted in terms of conditional ΔP along the same lines as our yellowed fingers and graduate admissions examples (see Melz, Cheng, Holyoak, & Waldmann, 1993; Shanks, 1995; Spellman, 1996a). These and other studies (Fratianne & Cheng, 1995; Park & Cheng, under review; Spellman, 1996a; Yarlas, Cheng, & Holyoak, 1995) show that untutored reasoners prefer to judge the causal influence of c on e conditional on the absence of other potential causes of e that are correlated with c. This finding has led some psychologists to suggest that ΔP conditional on alternative causes describes human causal judgments (e.g., Cheng & Holyoak, 1995; Melz, Cheng, Holyoak, & Waldmann, 1993; Spellman, 1996a & b; Waldmann & Holyoak, 1992). It turns out that ΔP conditional on alternative causes is exactly what is computed asymptotically by the Rescorla-Wagner model (1972) for many experimental designs to which the model has been applied (see Cheng, 1997, for a proof of the conditions under which this model computes conditional ΔPs). The Rescorla-Wagner model incorporates a version of the popular "delta rule" in connectionist models and is the most prominent associationist model of Pavlovian conditioning and of causal induction.

The type of deviation from ΔP just discussed is only one among a diverse set of such deviations. Even in situations in which alternative causes (that are not along a candidate causal pathway) are controlled, robust deviations from conditional ΔP have been observed in many psychological experiments:

- When $P(e \mid c) = P(e \mid \sim c) = 1$, people do not judge that c does not produce e; rather, they tend to withhold causal judgment altogether (Fratianne & Cheng, 1995; Waldmann & Holyoak, 1992; Wu & Cheng, 1997).

- When $P(e \mid c) = P(e \mid {\sim}c) = 0$, people do not judge that c does not prevent e; rather, they tend to withhold causal judgment altogether (Wu & Cheng, 1997; Yarlas et al., 1995).

- If $P(e \mid c) - P(e \mid {\sim}c) = P(e \mid c') - P(e \mid {\sim}c') > 0$, and $P(e \mid {\sim}c) > P(e \mid {\sim}c')$, people tend to judge that c has greater power than c' to produce e (Buehner & Cheng, 1997; Vallée-Tourangeau, Murphy, & Baker, 1996).

- In contrast, if $P(e \mid c) - P(e \mid {\sim}c) = P(e \mid c') - P(e \mid {\sim}c') < 0$, and $P(e \mid {\sim}c) > P(e \mid c')$, then people tend to judge that c' has more power than c to prevent e (Buëhner & Cheng, 1997; Vallee-Tourangeau et al., 1996).

- People tend to weigh frequencies of events that estimate $P(e \mid c)$ more than those that estimate $P(e \mid {\sim}c)$ (e.g., Anderson & Sheu, 1995; Baron, 1994; Dickinson & Shanks, 1986; Schustack & Sternberg, 1981; Wasserman et al., 1993).

Why are there these deviations? And are they normative?

3. An Alternative Causal Mechanism Approach

We have argued that information about causal mechanisms and probabilities are not mutually exclusive. Instead, patterns of probabilities are manifestations of the operations of causal mechanisms. In our alternative conception of the connection between causal mechanisms and patterns of probabilities, all of the above deviations can be normative. This conception is congruent with scientific uses of the notion of causation, and has been adopted by researchers in philosophy, computer science, statistics, and psychology (e.g., Cheng, 1997; Pearl, 1988; Spirtes et al., 1993; other references). Under this conception, a central goal of causal inference is to make accurate predictions about the consequences of actions in novel as well as familiar situations. The achievement of this goal requires the estimation of relational properties that are as independent of context as available information allows. It also requires the acknowledgement of some constraints. First, it is important to understand how causal inference can start from observations alone, or from observations and very limited constraints such as time order. Second, ~~because there are infinitely many causal structures that are consistent with any pattern of probabilistic dependence and independence,~~ causal inference from patterns of probabilities requires the assumption that the influence of a cause on an effect can be represented as a probability~~, fundamental to experimental design, that probabilistic independence reflects causal irrelevance~~. [Clark, I haven't thought through whether we need an additional assumption: in the absence of known causes, the absence of an effect in question reflects the absence of all causes.] Third, other things equal, causal inference requires a preference for simpler causal explanations over more complex explanations that equally account for observed associations. And fourth, ordinarily causal inference uses prior knowledge.

Under this conception, there are statistics that are more informative than either ΔP or conditional ΔP. Consider the assessment of the consequences of smoking cigarettes. Suppose some tobacco company announces that it has found a human population in which smokers and nonsmokers are impeccably matched on all relevant variables, such as genetic disposition to lung cancer and asbestos level in the environment. This population therefore allows a more direct assessment of the power of smoking to produce lung cancer in humans than ever before possible. Now, in this population, P(lung cancer | smoking) – P(lung cancer | no smoking), controlling for alternative causes, is fairly small, say .05. However, the asbestos level in this population is uniformly high, so that P(lung cancer | smoking)=0.95 and P(lung cancer | no smoking) =0.90. If conditional ΔP were a normative criterion for causal inference, one would conclude that smoking poses a relatively small risk for lung cancer. Many smokers might then feel that the benefit of smoking warrants the risk.

In contrast, under our alternative framework, the same information from this population leads to the assessment that smoking has the *power* to produce lung cancer with a probability of .50, ten times higher than the value of conditional ΔP suggests. Smoking having a power of .50 to produce lung cancer in humans means that, when other causes of lung cancer are absent, conditional on an intervention that results in smoking (e.g., being assigned to smoke), lung cancer occurs in humans with a probability of .50. In the example, it is obviously infeasible to measure this power by conducting an experiment. We will return to explain how such powers are estimated under our approach. Unlike conditional ΔP, which is only relevant to the particular population with high asbestos level, the causal power of smoking to produce lung cancer is generalizable to other human populations. For example, in an environment in which there are no causes of lung cancer other than smoking, one would estimate that smoking raises the probability of lung cancer from 0 to 0.5. In an environment in which all causes of lung cancer other than smoking jointly produce lung cancer with a probability of .2, if smoking can be assumed to produce lung cancer independently of other causes, then it is estimated to raise the probability of lung cancer by .4 (from .2 to .6). This estimate follows because among those in this population who would not have contracted lung cancer from the other causes, smoking would be estimated to produce lung cancer in half of them (.8 x .5 = .4).

Some informative variables under this framework are:

I. the probability of e conditional on an action to produce c and on the absence of all other causes of e (the power variable illustrated in the smoking example).

II. the probability that e does not occur conditional on (1) an action to produce a cause c that prevents e, (2) the nonexistence of all other preventive causes of e, and (3) e occurring with certainty (due to the influence of some generative cause or causes) if not for c. Quantities I and II apply to direct causes.

III. the probability of e conditional on an action to produce c and on the absence of all other causes of e that are not effects of c.

IV. the probability that e does not occur given an action to produce an inhibiting cause c, and given that all positive causes of e that are not influenced by c occur. Quantities III and IV are more general than I and II in that they apply to *multi-layered* causal networks, causal structures in which there is at least one causal chain that contains two or more causal arrows pointing in the same direction.

V. the probability of e conditional on an action to produce c.

VI. the probability of e conditional on an action to produce c, minus the probability of e conditional on an action to produce the absence of c.

A formula for computing quantity I in a parametric class of mechanisms, when unobserved causes are independent of the cause to be assessed, is given by Cheng (1997). She also gives a formula for estimating quantity II in a related class of systems, when the cause at issue prevents, rather than generates, the effect. Quantities I and II have been shown to be psychologically relevant. A theory proposed by Cheng (1997) explains a range of psychological phenomena by postulating that people estimate ordinal properties of quantities I and II in situations in which they are willing to assume that a candidate cause and a composite of all causes alternative to it influence e independently. These phenomena include all the deviations from ΔP and conditional ΔP mentioned earlier. Cheng's formula for estimating I, the power of a direct generative cause, is generalized to any multi-layered mechanism of the same parametric class, under the same independence assumption, by Glymour (1997), illustrating how to calculate III for a class of parametrizations. No one has yet investigated IV, so far as we know. An algorithm for computing V (or VI) in many problems is given by Spirtes et al. (1993), and rules for its calculation are developed in Pearl (1995).

The model of an emergency medical system depicted in Figure 1 (taken from Beinlich et al., 1989) is an example of a causal mechanism that a computer algorithm based on this framework (see Spirtes et al., 1993) is able to infer from probabilistic data, without the use of any prior causal information.

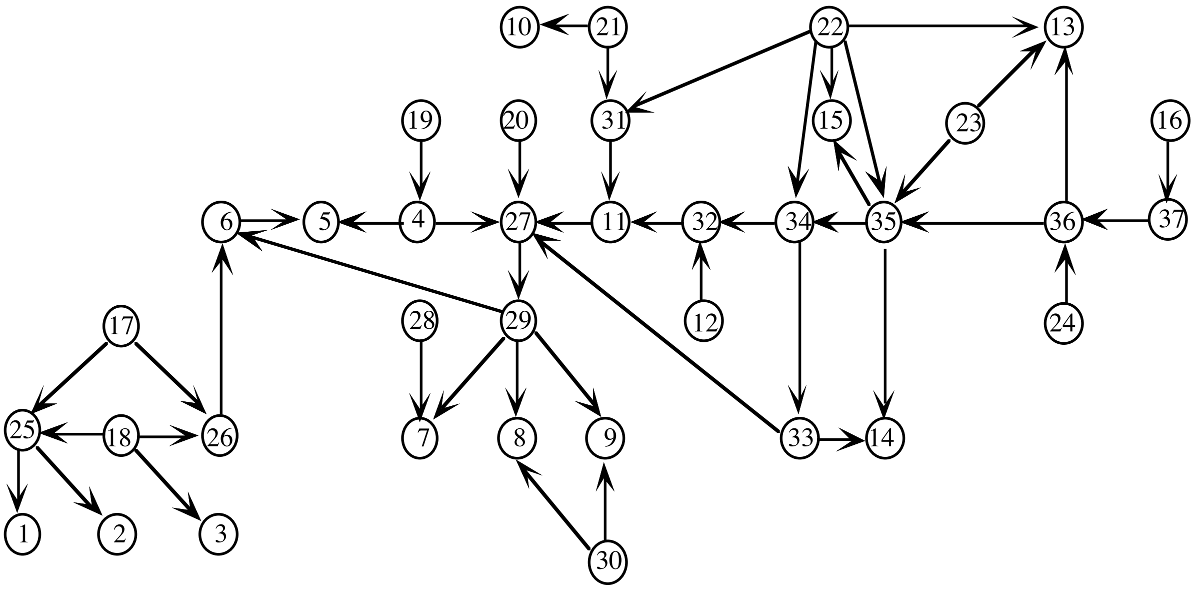


Figure 1. The ALARM belief network

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 monitor

8 - heart rate obtained from electrocardiogram

9 - heart rate obtained from oximeter

10 - pulmonary artery pressure

11 - arterial-blood oxygen saturation

12 - fraction of oxygen in inspired gas

13 - ventilation pressure

14 - carbon-dioxide content of expired gas

15 - minute volume, measured

16 - minute volume, calculated

20 - insufficient anesthesia or analgesia

21 - pulmonary embolus

22 - intubation status

23 - kinked ventilation tube

24 - disconnected ventilation tube

25 - left-ventricular end-diastolic volume

26 - stroke volume

27 - catecholamine level

28 - error in heart rate reading due to low cardiac output

29 - true heart rate

30 - error in heart rate reading due to electrocautery device

31 - shunt

32 - pulmonary-artery oxygen saturation

33 - arterial carbon-dioxide content

34 - alveolar ventilation

17 - hypovolemia
18 - left-ventricular failure
19 - anaphylaxis

35 - pulmonary ventilation
36 - ventilation measured at endotracheal tube
37 - minute ventilation measured at the ventilator

Clark, say what role the quantities play in the algorithms that infer networks such as the above.

4. Computing Quantities I and II.

Let e be an event, or event type, and a, i be events, or event types, that may cause e, and let them be all of the possible causes of e, so that if neither a nor i occurs, e does not occur. We can for formal convenience replace e, a, i by binary variables taking the value 1 if the corresponding event occurs and the value 0 otherwise (for example, e = 1 if e occurs and e = 0 otherwise). There will then be cases in which i, for example, occurs but does not cause e, and cases in which i occurs and does cause e. We can distinguish between those cases by introducing additional parameters, for example by introducing a binary variable q_i which indicates when the occurrence of i will produce e. The probability that $q_i = 1$ then gives the probability that, given that i occurs, that occurence causes e. Thus,

(1) $e = f(iq_i; aq_a)$

where the functional form f is yet to be specified. A simple example common in computer science is the "noisy *or* gate" given by the equation

(2) $e = iq_i \oplus aq_a$

where e, i, a, q_i , and q_a are binary variables and the $\oplus$ is Boolean addition. (It will become clear in a later section how the $\oplus$notation is useful for simplying the algebra.) Assume {i, a, q_i, q_a} to be jointly independent. Then for the noisy *or* gate,

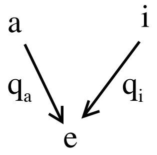
(3) $P(e=1) = P(iq_i \oplus aq_a=1) = P(i=1)P(q_i=1) + P(a=1)P(q_a=1) - P(a=1)P(q_a=1)P(i=1)P(q_i=1)$.

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

(4) $P(q_i=1) = P(e=1 \mid i= 1, a = 0)$.

That is, the parameter is quantity I of section 2.

The noisy *or* gate can be represented graphically as:



The q parameters are associated with their respective arrows.

Cheng (1997) shows that $P(q_i=1)$ can be estimated *without observing a* by

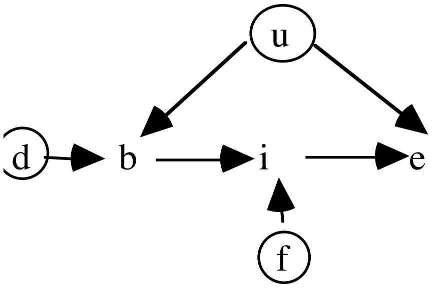
(5) $P(q_i = 1) = [P(e=1 \mid i=1) - P(e=1 \mid i = 0)] / [1 - P(e=1 \mid i = 0)]$.

Recall that $\{i, a, q_i, q_a\}$ are jointly independent. Because $P(q_i=1)$ is estimated without confounding by alternative causes, it corresponds to the relevant probability of e conditional on an intervention to produce i.

In her psychological theory, the noisy *or* gate illustrated in the diagram is a psychological construct in which i represents the candidate cause in question and a represents a composite of all causes alternative to i. This construct is a filter through which the reasoner can view any causal structure with arbitrarily many layers and any topology, as long as all *"ancestors"* of e can be divided into two parts, corresponding respectively to i and its ancestors on one hand and a on the other, and i and a produce e independently. Ancestors of a node are direct or remote causes of the node. It is crucial that to infer the power of i, the theory does not require knowledge of the power of a. In most cases (including those in which causal learning does occur), it is not possible to have an exhaustive list of all alternative causes, not to say their causal powers.

Formula 5 is derived from explaining each of the two conditional probabilities in ΔP by this or-gate construct. Thus, P(e | i) is the probability of the union of two events: (1) e produced by i and (2) e produced by a when i is present, and P(e | ~i) is the probability of e produced by a alone when i is absent. Formula 5 follows from this explanation of ΔP in the special case in which a occurs independently of i. A variant of (5) applies even if i and a are confounded by some observed common cause b. One need only also condition on b = 0 in every term on the right hand side in (5). The conditionalizing recreates the independence assumption.

The~~se~~ requirement that the candidate cause~~ whose power is assessed~~ be independent of all unobserved causes of e is sufficient but not necessary. For many mechanisms, parameters such as $P(q_i = 1)$ can be estimated from frequencies even when there are confounding unobserved causes. For example, representing the mechanism again by a diagram in the form of a directed acyclic graph, consider the structure:



where the variables in circles are unobserved. A convention adopted in the representation of directed graphs is that the absence of a common cause node implies that there is no such common cause. The problem is to estimate $P(q_{ie} = 1) = P(e = 1 \mid i = 1, u = 0)$. Note that i and u are not independent but are independent conditional on any value of b. More generally, as a consequence of structures such as a noisy *or* gate, in which the value of a node is determined by the values of the parameters of arrows and nodes leading into it, an effect is independent of its remote causes conditional on a more direct cause. In this case, the parameter $P(q_{ie} = 1)$ can be estimated by conditioning on b = 1 in Cheng's formula, that is:

$$P(q_{ie} = 1) = [P(e = 1 \mid i = 1, b = 1) - P(e = 1 \mid i = 0, b = 1)] / [1 - P(e = 1 \mid i = 0, b = 1)].$$

In general, to estimate the influence of i on e it suffices to condition on a measured variable on each path from each confounder, such as u, to i. More generally, still, if there are a set of measured variables conditional on which a direct cause i of e is independent of all other causes of e (save ancestors of i), $P(q_i=1)$ can be estimated. These measured variables allow the selection of cases from the data to allow an estimation of causal power. Algorithms for calculating the independencies implied by a mechanism have been published in several places (Lauritzen, 1997 [Clark, supply reference]; Pearl, 1988; Spirtes et al., 1993); implementations are available in commercial software programs, and are accessible for free on the web.

Because the variables i and a in the noisy *or* gate both potentially produce e, the noisy *or* gate implies that $P(e =1 \mid i = 1) \geq P(e = 1 \mid i = 0)$. It follows that only nonnegative ΔPs can be interpreted in terms of a noisy *or* gate, a construct that allows the evaluation of whether a candidate cause generates or produces an effect. Cheng (1997) introduces a noisy *and* gate for interpreting nonpositive ΔPs to evaluate whether a candidate cause *prevents* an effect. For candidate preventive cause, i, when a generative cause a is also present,

(6) $e = aq_a(1 - iq_i)$

The noisy *and* gate can be represented by the same diagram as a noisy *or* gate, but q_i has a quite different meaning. Using the noisy *and* gate in (6) to explain ΔP, just as she did using the noisy *or* gate, Cheng derived that

(7) $P(e = 1 \mid i = 1) = P(a = 1 \mid i = 1)P(q_a=1)(1 - P(q_i=1))$.

In this case,

(8) $P(e = 0 \mid a = 1, q_a=1, i = 1) = P(q_i = 1)$,

so that for a noisy *and* gate, $P(q_i = 1)$ is quantity II of section 2, where $P(q_i= 1)$ now measures the power of i to prevent e from occurring. When a and i are independent Cheng shows that $P(q_i = 1)$ can be estimated without observing a because

(9) $P(q_i =1) = - [P(e =1 \mid i =1) - P(e = 1 \mid i = 0)] / P(e =1 \mid i = 0)$.

We now return to explain the deviations from ΔP and conditional ΔP mentioned in Section 2. Recall that Formula 5 follows from the noisy *or* gate explanation of ΔP in the special case in which a occurs independently of i. In other cases (see Cheng, 1997), this explanation implies that

(10) $P(q_i =1) = [\Delta P - P(q_a=1) P(a=1 \mid i=1) + P(q_a=1) P(a=1 \mid i=0)] / [1 - P(q_a=1) P(a=1 \mid i=1)]$.

Formula 10 shows why covariation does not in general imply causation. The denominator on the right-hand-side has both a negative term and a positive term in addition to the ΔP term. It can be shown that ΔP can either overestimate $P(q_i =1)$ or underestimate it. Thus, ΔP does not provide an interpretable estimate of $P(q_i =1)$ when a does not occur independently of i, but ΔP does estimate $P(q_i =1)$ according to Formula 5 otherwise. A formula analogous to 10 follows from the noisy-*and*-gate explanation of ΔP. For a reasoner whose goal is to estimate $P(q_i = 1)$, it would therefore be rational to prefer to judge the causal influence of c on e conditional on the absence of other potential causes of e that are correlated with c, as our yellowed fingers example and the results of many psychological experiments indicate.

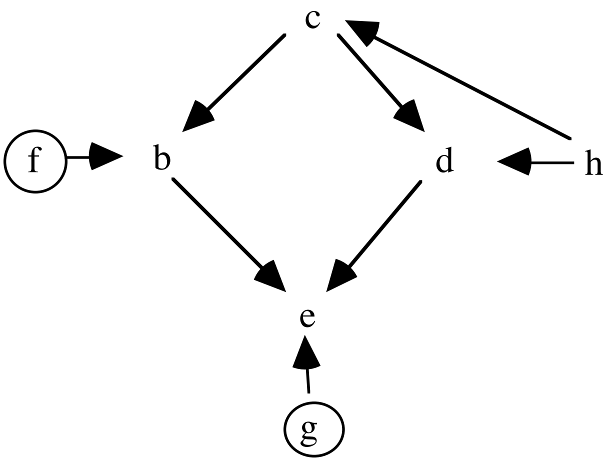
Formulae 5 and 9 explain the rest of the findings mentioned. According to Formula 5, for judgments regarding the generative power of c, $P(q_i = 1)$ is undefined when $P(e = 1 \mid i = 0) = 1$, leading to the prediction that the reasoner should withhold causal judgment in this situation. Likewise, according to this formula, if $P(e \mid c) - P(e \mid \sim c) = P(e \mid c') - P(e \mid \sim c')$ and $P(e \mid \sim c) > P(e \mid \sim c')$, then $P(q_c = 1) > P(q_{c'} = 1)$. According to Formula 9, however, for judgments regarding the preventive power of c, $P(q_i = 1)$ is undefined when $P(e = 1 \mid i = 0) = 0$, leading to the opposite prediction about when the reasoner should withhold causal judgment. Moreover, according to this formula, if $P(e \mid c) - P(e \mid \sim c) = P(e \mid c') - P(e \mid \sim c')$ and $P(e \mid \sim c) > P(e \mid \sim c')$, then $P(q_c = 1) < P(q_{c'} = 1)$.

The well-known finding that people tend to weigh frequencies of events that estimate $P(e \mid c)$ more heavily than those that estimate $P(e \mid \sim c)$ also follows directly from Formulae 5 and 9 (see Cheng, 1997, for the derivation). These formulae predict an exception to this tendency: for changes in ΔP between a zero and a non-zero value, these kinds of frequencies should be weighed equally. We do not know of empirical tests of this predicted exception.

As should be clear, some of the deviations from ΔP observed in experiments conducted on untutored reasoners correspond to standard principles of experimental design. Cheng's (1997) theory therefore provides an explanation of scientific as well as everyday uses of the notion of causation.

5. Calculating Quantity III.

We have considered the noisy *or* gate theory as a mental construct that consists of direct causes. This theory can be generalized to any mechanism, with arbitrarily many levels and any topology, so long as there are no cycles in the corresponding directed graph (Glymour, 1997). Cheng's formula for evaluating the generative power of c on e generalizes as follows. For any directed graph representing the mechanism involving c and e, consider each directed path (each causal pathway) from c to e. Form the product of the q coefficients associated with the links on each path, then take the Boolean sum of these products over all paths from c to e. The probability of that Boolean sum is the parameter whose value is the probability of c given e and given that all other causes of e, that are not themselves effects of c, are absent. To illustrate, consider the structure in the next figure, in which the variables in circles are assumed to be unobserved.



The causes of e that are not effects of c are f, h and g. According to the theorem just given,

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

The RHS of this equation is conditionalized on h = 0 because h is a confounder of the influence of c on e. Recall that by the convention used in directed graphs, f and g are independent of c.

We will illustrate the rather tedious algebra for this case:

$e = q_g g \oplus q_b b \oplus q_d d$

$= q_g g \oplus q_b(q_f f \oplus q_{cb} c) \oplus q_d q_h h \oplus q_d\, q_{cd} c$

$= (q_g g \oplus q_b q_f f \oplus q_d q_h h) \oplus c(q_b q_{cb} \oplus q_d q_{cd})$

When c = 1, and h = 0:

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

When c = 0 and h = 0:

$e = q_g g \oplus q_b q_f f$

$\Delta p = P(e = 1 \mid c = 1, h = 0) - P(e = 1 \mid c = 0, h = 0)$

$= P(q_b q_{cb} \oplus q_d q_{cd}) - P(q_g g \oplus q_b q_f f) \bullet P(q_b q_{cb} \oplus q_d q_{cd})$

$= P(q_b q_{cb} \oplus q_{cd} q_d)\ [1 - P(e =1 \mid c = 0, h = 0)].$

Note that a single-layered noisy *or* gate filter superimposed over a multi-layered causal mechanism and the decomposition of this filter into a multi-layered structure yield internally consistent causal powers. The assumptions underlying the two analyses are identical except for the grain size of the directed graphs. It therefore would not make sense to suggest, as does the dichotomy between mechanisms and probabilities, that "mechanisms" must consist of a multi-layered causal structure, and are conceptually distinct from causal relations involving a single link. It is true, however, that multi-layered mechanisms allow inferences not possible in single-layered mechanisms, as illustrated in some of our examples. For this reason, multi-layered mechanisms allow tests for internal consistency that do not apply to a single causal link.

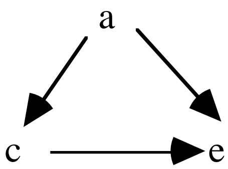
Although there are experimental studies in the psychological literature that investigate causal judgments in multi-layered structures (e.g., Busemeyer, McDaniel, & Byun, 1997; Spellman, in press), there seems to be none that investigates judgments of causal power in such structures. Nor, with one exception (Haseem & Cooper, 1997), are there studies of human ability to learn the mechanism represented by a multi-layered directed graph from various combinations of observation, intervention and prior causal knowledge. There are a number of computer algorithms that learn such structures, or features of them, from frequency data alone (Spirtes et al., 1993).

We have not investigated the analogous generalization of noisy *and* gates.

6. Computing V and VI.

In contexts in which alternative actions (say, making c = 0, or c= 1, or c = 2, where the values of c are mutually exclusive) may be taken, and the probability of a resulting event is important (say the probability that e = 1), the quantities of interest are likely to be P(e = 1 | an action that makes c = i). In general, these quantities, and algebraic combinations of them, are distinct from P(e = 1 | c =i) (recall the yellow fingers and gloves example) and also from quantities calculated in previous sections, e.g., P(e = 1 | c = i and all other causes of e are absent), and from algebraic combinations of these quantities.

Given a mechanism, represented by a directed acyclic graph, for which there are no unobserved common causes of a causal factor c and a variable e that c influences, and given the conditional probability distribution of each variable conditional on each set of values of each immediate cause (each parent of the variable in the graph), the probability distribution for e given an action that forces a value c' on c can be easily calculated. One simply ignores any influences on c, and uses the probabilities conditional on c = c'. For example, if the mechanism is represented by:

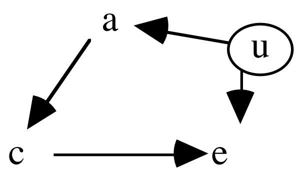


then the probability that e = 1 given an intervention to force c = c' is

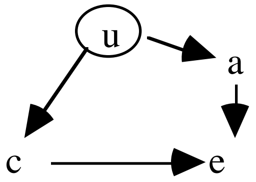
$$\Sigma_a P(e = 1 | c', a)P(a)$$

where the sum is over all values of a and the probabilities are those that hold before any intervention. The formula holds no matter how many values e, c and a may take on. Note that when all variables are binary, the formula is distinct from the probability that e = 1 given an action to force c = c' and given a = 0. P(e) is summed over all values of a in V but conditional on the absence value of a in I. [Clark, explain why V and VI are better than the quantities we bashed.]

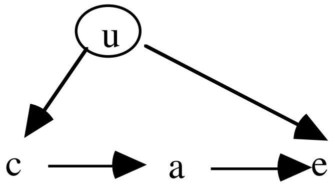
For appropriate mechanisms, the effect on e of an intervention on c can be predicted even when c and e have unobserved common causes, as in the mechanisms represented by the diagram



or the diagram



or the diagram



Spirtes et al. (1993) give an algorithm for determining for any mechanism whether the distribution of an effect e results from an intervention to fix a value of another variable c, and for computing the consequent distribution of e from the probabilities before the intervention. Pearl (1995) gives rules of calculation. It is not known if the algorithm or the rules are complete.

7. Conclusion

We have argued that the false~~a~~ dichotomy between mechanistic and probabilistic approaches to causal reasoning obscures a conception of probabilities as manifestations of causal mechanisms. This alternative conception of the relation between probabilities and mechanisms supports the estimation of relational properties in the world that allow the explanation and prediction of the consequences of actions in novel as well as familiar situations. Such estimates are not merely normative; there is considerable evidence indicating that some of these estimates describe untutored human causal induction. In fact, many robust observed deviations from ΔP and conditional ΔP are normative consequences of estimating quantities under this framework.

To us, proponents of the dichotomy are not addressing the issue of how an intelligent system discovers causal knowledge, but are instead attempting to document evidence for the use of prior causal knowledge. There is consensus that reasoners should find prior causal knowledge useful wherever it is applicable. Our alternative framework is not only compatible with the use of prior causal knowledge, it~~but~~ also allows explicit formulations of how prior causal knowledge normatively interacts with novel observations and knowledge of interventions. Psychological work adopting this framework is mostly limited to mental constructs involving single-layered causal mechanisms. Whether people intuitively make use of normative algorithms for inferring multi-layered causal mechanisms and whether they normatively combine information regarding observations and interventions and prior causal knowledge remain to be investigated.

References

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*.

Ahn, W., Kalish, C.W., Medin, D.L., & Gelman, S.A. (1995). The role of covariation versus mechanism information in causal attribution, *Cognition*, *54,* 299-352.

Allan, L.G. (1980). A note on measurements of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*, 147-149.

Allan, L.G., & Jenkins, H.M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14***,** 381-405.

Anderson, J.R., & Sheu, C-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, *23*, 510-524.

Baker, A.G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: The presence of a strong causal factor may reduce judgments of a weaker one, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 414-432.

Baron, J. (1994). *Thinking and deciding*. (Second Edition). New York: Cambridge University Press.

Beinlich, I, Suermondt, H. Chavez, R. and Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief netowrks.
Proceeds of the Second European Conference on Artificial Intelligence in Medicine, London, England, 247-256.

Buehner, M. J. & Cheng, P.W. (1997). The influence of the base rate of the effect on causal judgments of candidate causes with the same contingency. *Proceedings of the Nineteenth Annual Cognitive Science Society*.

Busemeyer, J., McDaniel, M.A. & Byun, E. (1997). Multiple input-output causal environments. *Cognitive Psychology*.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 837-854.

Chapman, G. B., & Robbins, S. I. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18*, 537-545.

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Cheng, P.W., & Novick, L.R. (1992). Covariation in natural causal induction. *Psychological Review*, 99, 365-382.

Dickinson, A., & Shanks, D.R. (1986). The role of selective attribution in causality judgment. In D.J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality*. Brighton: Harvester Press.

Dickinson, A., Shanks, D.R., & Evenden, J. L. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Pspychology*, *36A*, 29-50.

Fratianne, A. & Cheng, P.W. (1995). Assessing causal relations by dynamic hypothesis testing. Department of Psychology, UCLA.

Freedman, D, Pisani, R, & Purves, R. (1978)*Statistics*.. New York : Norton.

Glymour, C. (1997). Learning causes. *Minds and Machines.*

Harré, R., & Madden, E.H. (1975). *Causal powers: A theory of natural necessity*, Totowa, New Jersey: Rowman & Littlefield.

Haseem, A. I. & Cooper, G.F. (1996). Human causal discovery from observational data. Proceedings of the 1996 symposium of the American Medical Information Association.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 7*, 1-17.

Melz, E.R., Cheng, P.W., Holyoak, K.J., & Waldmann, M.R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner learning rule? Comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1398-1410.

Park, J., & Cheng, P. W. (under review). Boundary conditions on "overexpectation" in causal learning with discrete trials: A test of the power PC theory. Department of Psychology, UCLA.

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)*, 669-709, 1995.

Price, P.C., & Yates, J.F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory & Cognition, 21*, 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*, 1-5.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 433-443.

Shanks, D.R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, *48A*, 257-279.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*, No. 1.

Spellman, B.A. (1996a). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7,* 337-342.

Spellman, B.A. (1996b). Conditionalizing causality. In D.R. Shanks, K.J. Holyoak, D.L. Medin (Eds.), *The Psychology of Learning and Motivation*, vol *34*: Causal learning (pp. 167-207). San Diego: Academic Press.

Spellman, B. A. (in press). Crediting causality. *Journal of Experimental Psychology: General.*

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search*. Springer-Verlag: New York.

Turner, M. (1987). *Death is the mother of beauty: Mind, metaphor, criticism.* Chicago: University of Chicago Press.

Vallée-Tourangeau, F., Murphy, R.A., Baker, A.G. (1996). Judging the contingency of a constant cue: Contrasting predictions from an associative and a statistical model. Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society, G W. Cottrell (Ed.), pp. 447-452.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222-236.

Wasserman, E.A., Chatlosh, D.L., & Neunaber, D.J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation*, *14*, 406-432.

Wasserman, E.A., Elek, S.M., Chatlosh, D.L., & Baker, A.G. (1993). Rating causal relations: The role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Motivation, and Cognition*, *19*, 174-188.

White, P.A. (1989). A theory of causal processing. *British Journal of Psychology*, *80*, 431-454.

White, P.A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition, 23*, 243-254.

Yarlas, A. S., Cheng, P. W., & Holyoak, K. J. (1995). Alternative approaches to causal induction: The probabilistic contrast versus the Rescorla-Wagner model. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 431-436). Hillsdale, NJ: Erlbaum.