Embracing Causality in Formal Reasoning* †

Judea Pearl

Cognitive Systems Laboratory,
UCLA Computer Science Department, Los Angeles, CA. 90024

Abstract

The purpose of this note is to draw attention to certain aspects of causal reasoning which are pervasive in ordinary discourse yet, based on the author's scan of the literature, have not received due treatment by logical formalisms of common-sense reasoning. In a nutshell, it appears that almost every default rule falls into one of two categories: *expectation-evoking* or *explanation-evoking*. The former describes association among events in the *outside* world (e.g., Fire is typically accompanied by smoke.); the latter describes how *we* reason about the world (e.g., Smoke normally suggests fire.). This distinction is consistently recognized by people and serves as a tool for controlling the invocation of new default rules. This note questions the ability of formal systems to reflect common-sense inferences without acknowledging such distinction and outlines a way in which the flow of causation can be summoned within the formal framework of default logic.

I. How Old Beliefs were Established Determines which New Beliefs are Evoked.

Let A and B stand for the following propositions:

A -- Joe is over 7 years old.
B -- Joe can read and write.

Case 1: Consider a reasoning system with the default rule

$$def_B: \; B \rightarrow A .$$

A new fact now becomes available,

e_1 -- Joe can recites passages from Shakespeare,

together with a new default rule:

$$def_1: \; e_1 \rightarrow B .$$

Case 2: Consider a reasoning system with the same default rule,

$$def_B: \; B \rightarrow A .$$

A new fact now becomes available,

e_2 -- Joe's father is a Professor of English,

together with a new default rule,

$$def_2: \; e_2 \rightarrow B .$$

(To make def_2 more plausible, one might add that Joe is known to be over 6 years old and is not a moron.)

*This work was supported in part by the National Science Foundation, Grant DCR 8644931.

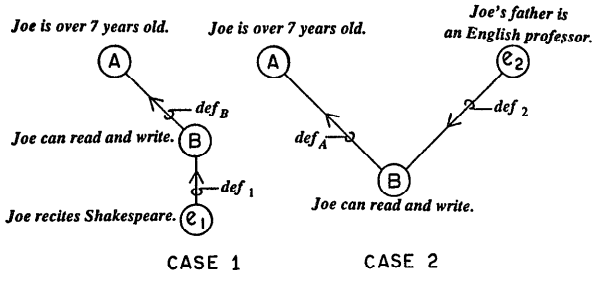

Figure 1

Common sense dictates that Case 1 should lead to conclusions opposite to those of Case 2. Learning that Joe can recite Shakespeare should evoke belief in Joe's reading ability (B) and, consequently, a correspondingly mature age (A). Learning of his father's profession, on the other hand, while still inspiring belief in Joe's reading ability, should NOT trigger the default rule $B \rightarrow A$ because it does not support the hypothesis that Joe is over 7. On the contrary; whatever evidence we had of Joe's literary skills could now be partially attributed to the specialty of his father rather than to Joe's natural state of development. Thus, if a belief were previously committed to A, and if measures of belief were permitted, it would not seem unreasonable that e_2 would somewhat *weaken* the belief in A.

From a purely syntactic viewpoint, Case 1 is identical to Case 2. In both cases we have a new fact triggering B by default. Yet, in Case 1 we wish to encourage the invocation of $B \rightarrow A$ while, in Case 2, we wish to inhibit it. Can a default-based reasoning system distinguish between the two cases?

The advocates of existing systems may argue that the proper way of inhibiting A in Case 2 would be to employ a more elaborate default rule, where more exceptions are stated explicitly. For example, rather than $B \rightarrow A$, the proper default rule should read: $B \rightarrow A \mid UNLESS \; e_2$.

Unfortunately, this cure is inadequate on two grounds. First, it requires that every default rule be burdened with an unmanageably large number of conceivable exceptions. Second, it misses the intent of the default rule $def_B: \; B \rightarrow A$, the primary aim of which was to evoke belief in A whenever the truth of B can be ascertained. Unfortunately, while correctly inhibiting A in Case 2, the UNLESS cure would also inhibit A in many other cases where it should be encouraged.

For example, suppose we actually test Joe's reading ability and find out that it is at the level of a 10-year old child, unequivocally establishing the truth of B. Are we to suppress the natural conclusion that Joe is over 7 on the basis of his father being an English professor? There are many other conditions under which even a 5-year-old boy can be expected to acquire reading abilities, yet, these should not be treated as exceptions in the default-logical sense because those same conducive conditions are also available to a seven-year old; and, consequently, they ought not to preclude the natural conclusion that a child with reading ability is, typically, over 7. They may lower, somewhat, our confidence in the conclusion but should not be allowed to totally and permanently suppress it.

To summarize, what we want is a mechanism that is sensitive to how B was established. If B is established by direct observation or by strong evidence supporting it (Case 1), the default rule $B \rightarrow A$ should be invoked. If, on the other hand, B was established by *EXPECTATION, ANTICIPATION* or *PREDICTION* (Case 2), then $B \rightarrow A$ should not be invoked, no matter how strong the expectation.

The asymmetry between expectation-evoking and explanation-evoking rules is not merely that of temporal ordering, but is more a product of human memory organization. For example, age evokes expectations of certain abilities not because it precedes them in time (in many cases it does not) but because the concept called "child of age 7" was chosen by the culture to warrant a name for a bona-fide frame, while those abilities were chosen as expectational slots in that frame. Similar asymmetries can be found in object-property, class-subclass and action-consequence relationships.

II. More on the Distinction between Causal vs. Evidential Support.

Consider the following two sentences:

1. Joe seemed unable to stand up; so, I believed he was injured.
2. Harry seemed injured; so, I believed he would be unable to stand up.

Any reasoning system that does not take into account the direction of causality or, at least, the source and mode by which beliefs are established is bound to conclude that Harry is as likely to be drunk as Joe. Our intuition, however, dictates that Joe is more likely to be drunk than Harry because Harry's inability to stand up, the only indication for drunkenness mentioned in his case, is portrayed as an expectation-based property emanating from injury, and injury is a perfectly acceptable alternative to drunkenness. In Joe's case, on the other hand, not-standing-up is described as a primary property supported by direct observations, while injury is brought up as an explanatory property, inferred by default.

Note that the difference between Joe and Harry is not attributed to a difference in our confidence in their abilities to stand up. Harry will still appear less likely to be drunk than Joe when we rephrase the sentences to read:

1. Joe showed slight difficulties standing up; so, I believed he was injured.
2. Harry seemed injured, so, I was sure he would be unable to stand up.

Notice the important role played by the word "so." It clearly designates the preceding proposition as the primary source of belief in the proposition that follows. Natural languages contain many connectives for indicating how conclusions are reached (e.g., therefore, thus, on the other hand, nevertheless, etc.). Classical logic, as well as known versions of default logic, appears to stubbornly ignore this vital information by treating all believed facts and facts derived from other believed facts on equal footing. Whether beliefs are established by external means (e.g., noisy observations), by presumptuous expectations or by quest for explanation does not matter.

But even if we are convinced of the importance of the sources of one's belief; the question remains how to store and use such information. In the Bayesian analysis of belief networks [Pearl 1986], this is accomplished using numerical parameters; each proposition is assigned two parameters, π and λ, one measuring its accrued *causal* support and the other its accrued *evidential* support. These parameters then play decisive roles in routing the impacts of new evidence throughout the network. For example, Harry's inability to stand up will accrue some causal support, emanating from injury, and zero evidential support, while Joe's story will entail the opposite support profile. As a result, having observed blood stains on the floor would contribute to a reduction in the overall belief that Joe is drunk but would not have any impact on the belief that Harry is drunk. Similarly, having found a whiskey bottle nearby would weaken the belief in Joe's injury but leave no impact on Harry's.

These inferences are in harmony with intuition. Harry's inability to stand up was a purely conjectural expectation based on his perceived injury, but it is unsupported by a confirmation of any of its own, distinct predictions. As such, it ought not to pass information between the frame of injury and the frame of drunkenness. The mental act of imagining the likely consequences of an hypothesis does not activate other, remotely related, hypotheses just because the latter could also cause the imagined consequence. For an extreme example, we would not interject the possibility of a lung cancer in the context of a car accident just because the two (accidents and cancer) could lead to the same eventual consequence -- death.

The causal/evidential support parameters are also instrumental in properly distributing the impact of newly-observed facts among those propositions which had predicted the observations. Normally, those propositions which generated strong prior expectations of the facts observed would receive the major share of the evidential support imparted by the observation. For example, having actually observed Harry unable to stand up would lend stronger support to Harry's injury

than to Harry's drunkenness. Harry's injury, presumably supported by other indicators as well, provides strong predictive support for the observation, which Harry's drunkenness, unless it accrues additional credence, cannot "explain away."

Can a non-numeric logic capture and exploit these nuances? I think, to some degree, it can. True, it can not accommodate the notions of "weak" and "strong" expectations, nor the notion of "accrued" support, but this limitation may not be too severe in some applications, e.g., one in which belief or disbelief in a proposition is triggered by just a few decisive justifications. What we can still maintain, though, is an indication of how a given belief was established -- by expectational or evidential considerations, or both, and use these indications for deciding which default rules can be activated in any given state of knowledge.

III. The C–E System: A Coarse Logical Abstraction of Causal Directionality.

Let each default rule in the system be labeled as either *C–def* (connoting "causal") or *E–def* (connoting "evidential"). The former will be distinguished by the symbol $\rightarrow_c$, as in *"FIRE* $\rightarrow_c$ *SMOKE,"* meaning *"FIRE* causes *SMOKE,"* and the latter by $\rightarrow_e$, as in *"SMOKE* $\rightarrow_e$ *FIRE,"* meaning *"SMOKE* is evidence for *FIRE."* Correspondingly, let each believed proposition be labeled by a distinguishing symbol, "$\rightarrow_c$" or "$\rightarrow_e$." A proposition P is *E–believed*, written $\rightarrow_e P$, if it is a direct consequence of some *E–def* rule. If, however, all known ways of establishing P involve *C–def* rule as the final step, it is said to be *C–believed*, i.e., supported solely by expectation or anticipation. The semantics of the C–E distinction are captured by the following three inference rules:

$$\text{(a)}\quad \frac{P \rightarrow_c Q \qquad \rightarrow_c P}{\rightarrow_c Q} \qquad \text{(b)}\quad \frac{P \rightarrow_c Q \qquad \rightarrow_e P}{\rightarrow_c Q} \qquad \text{(c)}\quad \frac{P \rightarrow_e Q \qquad \rightarrow_e P}{\rightarrow_e Q}$$

Note that we purposely precluded the inference rule:

$$\frac{P \rightarrow_e Q \qquad \rightarrow_c P}{\rightarrow_e Q}$$

which led to counter-intuitive conclusions in Case 2 of Joe's story. These inference rules imply that conclusions can only attain *E–believed* status by a chain of purely *E–def* rules. $\rightarrow_c$ conclusions, on the other hand, may be obtained from a mix of *C–def* and *E–def* rules. For example, a *E–def* rule may (viz., (c)) yield a $\rightarrow_e$ conclusion which can feed into a *C–def* rule (viz., (b)) and yield a $\rightarrow_c$ conclusion. Note, also, that the three inference rules above would license the use of loops such as $A \rightarrow B$ and $B \rightarrow A$ without falling into the circular reasoning trap. Iterative application of these two rules would never cause an *C–believed* proposition to become *E–believed* because at least one of the rules must be of type C.

The distinction between the two types of rules can be demonstrated using the following example. (See Figure 2).

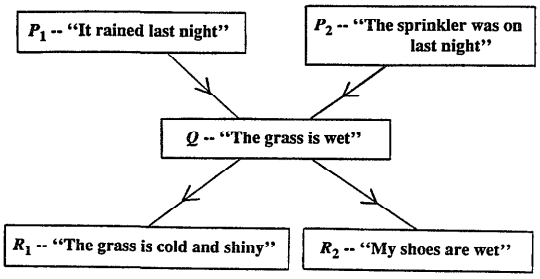

Figure 2

Let P_1, P_2, Q, R_1, and R_2 stand for the propositions:

P_1-- "It rained last night"
P_2--"The sprinkler was on last night"
Q -- "The grass is wet"
R_1--"The grass is cold and shiny"
R_2--"My shoes are wet"

The causal relationships between these propositions would be written:

$$P_1 \rightarrow_c Q \qquad Q \rightarrow_e P_1$$
$$P_2 \rightarrow_c Q \qquad Q \rightarrow_e P_2$$
$$Q \rightarrow_c R_1 \qquad R_1 \rightarrow_e Q$$
$$Q \rightarrow_c R_2 \qquad R_2 \rightarrow_e Q$$

If Q is established by an *E–def* rule such as $R_1 \rightarrow_e Q$ then it can trigger both P_1 and R_2. However, if Q is established merely by a *C–def* rule , say $P_2 \rightarrow_c Q$, then it can trigger R_2 (and R_1) but *not* P_1.

The essence of the causal assymmetry stems from the fact that two causes of a common consequence interact differently than two consequences of a common cause; the former COMPETE with each other, the latter SUPPORT each other. Moreover, the former interact when their connecting proposition is CONFIRMED, the latter interact only when their connecting proposition is UNCONFIRMED.

Let us see how this C–E system resolves the problem of Joe's age (See Fig.1.). def_B and def_1 will be classified as *E–def* rules, while def_2 will be proclaimed an *C–def* rule. All provided facts (e.g., e_1 and e_2) will naturally be *E–believed*. In Case 1, B will become *E–believed* (via rule (c)) and, subsequently, after invoking def_B in rule (c), A, too, will become *E–believed*. In case 2, however, B will only become *C–believed* (via rule (b)) and, as such, cannot invoke def_B, leaving A undetermined, as expected.

The C–E system in itself does not solve the problem of retraction; that must be handled by the mechanism of *exceptions*. For example, if in case 1 of Joe's story we are also told that e_3- "Joe is blind and always repeats what he hears" we should be inclined to retract our earlier conclusion that Joe can read and write, together with its derivative, that Joe is over

7 years old. However, the three inference rules above will not cause the negation of B unless we introduce e_3 as an exception to def_1, e.g., $e_1 \dashrightarrow_e B$ | *UNLESS* e_3. In the next section, we will touch on the prospects of implementing retraction without introducing exceptions.

IV. Can we Assemble a More Refined Non-numeric Abstraction?

E–believed status is clearly more powerful than *C–believed* status. The former can invoke both *C–def* and *E–def* rules, while the latter, no matter how strong the belief, invokes only *C–def* rules. The question may be raised whether one shouldn't dispose of this inferior, "*C*-rated" form of belief altogether and restrict a reasoning system to deal with beliefs based only on genuine evidential support[1]. The answer is that *C–def* rules, as weak as they sound, serve two functions essential for common-sensical reasoning: *predictive planning* and *implicit censorship*.

Planning is based on the desire to achieve certain expectations which can be predicted from one's current knowledge. The role of *C–def* rules is to generate those predictions from current *C–believed* and *E–believed* propositions. For example, if we consider buying Joe a birthday gift and we must decide between a book or a TV game, it would obviously be worth asking if we believe Joe can read. Such belief will affect our decision even if it is based on inferior, "*C*-rated" default rules, such as "If person Z is over 7 years old, then Z can read" or, even the weaker one yet: "If Z's father is an English professor, then Z can read." Prediction facilities are also essential in interpretive tasks such as language understanding because they help explain behavior of other planning agents around. Such facilities could be adequately served by the *C–E* system proposed earlier.

However, the prospect of using *C–def* rules as *implicit censors* of *E–def* rules is more intriguing because it is pervasive even in purely inferential tasks (e.g., diagnosis), involving no actions or planning agents whatsoever. Consider the "frame problem" in the context of car-failure diagnosis with the *E–def* rule: "If the car does not start, the battery is probably dead." Obviously, there are many exceptions to this rule, e.g., "... unless the starter is burned," "... unless someone pulled the spark plugs," "... unless the gas tank is empty," etc., and , if any of these conditions is believed to be true, people would *censor* the invocation of alternative explanations for having a car-starting problem. What is equally obvious is that people do not store all these hypothetical conditions explicitly with each conceivable explanation of car-starting problems but treat them as unattached, *implicit* censors, namely, conditions which exert their influence only upon becoming actively believed and, when they do, would uniformly inhibit *every E–def* rule having "car not starting" as its sole antecedent.

1 In Mycin [Shortliffe, 1976], for example, rules are actually restricted in this way, leading always from evidence to hypotheses.

But if the list of censors is not prepared in advance, how do people distinguish a genuine censor from one in disguise (e.g., "I hear no motor sound")? I submit that it is the *causal directionality* of the censor-censored relationship which provides the identification criterion. By what other criterion could people discriminate between the censor "The starter is burned" and the candidate censor "My wife testifies, 'The car won't start' ?" Either of these two inspires strong belief in "the car won't start" and "I'll be late for the meeting;" yet, the burned-out starter is licensed to censor the conclusion "the battery is dead," while my wife's testimony is licensed to evoke it. It is hard to see how implicit censorship could be realized, had people not been blessed with clear distinction between *explanation-evoking* and *expectation-evoking* rules. So, why blur the distinction in formal reasoning systems?

Note how convenient such a censorship scheme would be. No longer would we need to prepare the name of each potential censor next to that of a would-be censored; the connection between the two will be formed "on the fly," once the censor becomes actively believed. The mere fact that a belief in a proposition B is established by some *C–def* rule would automatically and precisely block all the rules we wished censored. More ambitiously, it could also lead to retracting all conclusions drawn from premature activation of such rules as in Truth-Maintenance Systems [Doyle, 1979]. True, to implement such a scheme we would need to label each believed proposition with the name of its (active) justifications and to augment our inference rules with instructions to correctly handle propositions which are both *E–believed* and *C–believed*. For example, Q could be *C–believed* due to P_1 and later become *E–believed* due to R_1, in which case (unlike purely *E–believed* propositions in inference-rule (c)), no $Q \rightarrow_e P_2$ rule should fire. However, this extra bookkeeping would be a meager price to pay for a facility that inhibits precisely those rules we wish inhibited and does so without circumscribing in advance under what conditions would a given proposition constitute an exception to any given rule. This is one of the computational benefits offered by the organizational instrument called causation and is fully realizable using Bayesian inference. Can it be mimicked in non-numeric systems as well?

Unfortunately the benefit of implicit censorship is hindered by a more fundamental issue, and it is not clear how it might be realized in purely categorical systems which preclude any sort of symbols for representing the degree of support that a premise imparts to a conclusion. Treating *all C–def* rules as implicit censors would be inappropriate, as was demonstrated in the starting theme of this note. In case-1 of Joe's story, we correctly felt uncomfortable letting his father's profession inhibit the *E–def* rule

CAN-READ(JOE) $\dashrightarrow_e$ OVER-7(JOE),

while now we claim that certain facts (e.g., burned starter), by virtue of having such compelling predictive influence over other facts (e.g., car not starting), should be allowed to inhibit all *E–def* rules emanating from the realization of such predictions (e.g., dead battery). Apparently there is a sharp qualitative difference between *strong C–def* rules such as

NOT-IN (Z, SPARKPLUGS) $\rightarrow_c$ WON'T-START (Z)

and *weak C–def* rules such as

ENGLISH-PROFESSOR (father (Z)) $\rightarrow_c$ CAN-READ (Z)
or
IN (Z, OLD-SPARKPLUGS) $\rightarrow_c$ WON'T-START (Z).

Strong *C–def* rules, if invoked, should inhibit all *E–def* rules emanating from their consequences. On the other hand, weak *C–def* rules should allow these *E–def* rules to fire (via rule (c)).

This distinction is exactly the role played by the parameter π which, in Bayesian inference, measures the accrued strength of causal support. It is primarily due to this strong vs. weak distinction that Bayesian inference rarely leads to counter-intuitive conclusions, and this is also why it is advisable to consult Bayes analysis as a standard for abstracting more refined logical systems which incorporate both degrees of belief and causal directionality. However, the purpose of this note is not to advocate the merits of numerical schemes but, rather, to emphasize the benefits we can draw from the distinction between causal and evidential default rules. It is quite feasible that with just a rough quantization of rule strength, the major computational benefits of causal reasoning could be tapped.

Conclusion

The distinction between *C–believed* and *E–believed* propositions allows us to properly discriminate between rules that should be invoked (e.g., case 1 of Joe's story) and those that should not (e.g., case 2 of Joe's story), without violating the original intention of the rule provider. While the full power of this distinction can, admittedly, be unleashed only in systems that are sensitive to the relative strength of the default rules, there is still a lot that causality can offer to systems lacking this sensitivity.

Acknowledgments

I thank H. Geffner, V. Lifschitz, D. McDermott, J. Minker, D. Perlis and C. Petrie for their comments on an earlier version of this paper.

References

[Doyle, 1979] Jon Doyle. "A Truth Maintenance System," *Artificial Intelligence,* 12:231-273, 1979.

[Pearl, 1986] Judea Pearl. "Fusion, Propagation and Structuring in Belief Networks," *Artificial Intelligence,* 29(3):241-288, September 1986.

[Shortliffe, 1976] Edward, H. Shortliffe. *Computer-Based Medical Consultation: MYCIN,* Elsevier, 1976.