A PROBABILISTIC TREATMENT OF THE YALE SHOOTING PROBLEM

Judea Pearl

December 1987
CSD-870068

TECHNICAL REPORT

CSD-8700XX

R-100

September 1987

A PROBABILISTIC TREATMENT OF THE YALE SHOOTING PROBLEM*

Judea Pearl

Cognitive Systems Laboratory

Computer Science Department

University of California, Los Angeles, CA. 90024-1596

ABSTRACT

This paper uses the Yale Shooting episode as a test bed for presenting the probabilistic approach to default reasoning. Using a probabilistic interpretation of causality and irrelevance, the approach provides a powerful logic for constructing sound qualitative arguments.

* This work was supported in part by NSF Grants #DCR 83-13875 and #IRI 86-10155.

A PROBABILISTIC TREATMENT OF THE YALE SHOOTING PROBLEM

Judea Pearl

1. INTRODUCTION

The so-called "Yale Shooting" problem [1] is regarded as the fuse that triggered McDermott's recent disenchantment with the logicist program in AI and has served as a focal point for discussions on the merit of this program. This paper presents a probabilistic treatment of the problem with the following objectives in mind:

1. To focus the logicists-probabilists debate on a concrete example.

2. To convince logicists that probability theory has more to it than number crunching. Taken as a logic for manipulating contexts, probability theory provides a powerful methodology for constructing sound qualitative arguments.

3. To convince probabilists that probability theory proper is insufficient for handling common sense reasoning. It can overcome some of the hurdles faced by the logicist approach only upon invoking the auxiliary notions of causation and relevance, in their appropriate probabilistic interpretations.

2. THE PROBLEM AND ITS SOLUTION

A simplified version of the Yale Shooting episode goes like this: suppose you load a gun at time t_1, wait for a while, then shoot someone at time t_2. The shooting is supposed to make the victim dead at time t_3, despite the normal tendency of "alive at t_2" to persist over long time periods. Yet, surprisingly, the logical formulation of the episode reveals an alternative, perfectly symmetrical version of reality, whereby the persistence of "alive" is retained while the persistence of "loaded" is interrupted, yielding the unintended conclusion that the victim is alive at time t_3. The question is what information people extract from the story that makes them prefer the persistence of "loaded" over the persistence of "alive".

The analysis of the shooting episode will be facilitated by the following definitions:

LD_1 = The gun is loaded at time t_1 $\quad$ AL_2 = The victim is alive at time t_2

LD_2 = The gun is loaded at time t_2 $\quad$ AL_3 = The victim is alive at time t_3

SH_2 = You shoot the gun (i.e., pull the trigger) at time t_2.

The story contains three known facts LD_1, AL_2, SH_2, and the problem is to infer the truth of $\neg AL_3$ (and LD_2). Domain knowledge is given by four default rules:

d_1: $LD_1 \rightarrow LD_2$ $\quad$ d_3: $AL_2 \wedge SH_2 \wedge LD_2 \rightarrow \neg AL_3$

d_2: $AL_2 \rightarrow AL_3$ $\quad$ d_4: $AL_2 \wedge SH_2 \wedge \neg LD_2 \rightarrow AL_3$

d_1 rule, for example, states that under normal circumstances a gun is expected to remain loaded, while d_2 asserts the natural tendency of life to persist over time. These rules can be given the following probabilistic interpretation:

d'_1: $P(LD_2 \mid LD_1) = high = 1 - \varepsilon_1$ $\quad$ d'_3: $P(AL_3 \mid AL_2, SH_2, LD_2) = low = \varepsilon_3$

d'_2: $P(AL_3 \mid AL_2) = high = 1 - \varepsilon_2$ $\quad$ d'_4: $P(AL_3 \mid AL_2, SH_2, LD_2) = high = 1 - \varepsilon_4$

where the ε's are small positive quantities whose exact values turn out to be insignificant.

Using these inputs, our task is to derive the conclusion that, given the stated facts $\{LD_1, AL_2, SH_2\}$, the victim is unlikely to remain alive at t_3, namely,

$$P(AL_3 \mid LD_1, AL_2, SH_2) = low. \tag{1}$$

Unlike their logicist colleagues, probabilists can discover *immediately* that the information given does not specify a complete probabilistic model and, so, is insufficient for deriving the intended conclusion (1) nor its negation. Moreover, the assumptions needed for completing the model can be identified and given precise formulation within the language of conditional probabilities.

Since the context of (1) differs from that of d'_3, the natural step is to refine the former by conditioning over the two possible states of LD_2:

$$P(AL_3 \mid LD_1, AL_2, SH_2) = P(AL_3 \mid LD_2, LD_1, AL_2, SH_2)P(LD_2 \mid LD_1, AL_2, SH_2)$$
$$+ P(AL_3 \mid \neg LD_2, LD_1, AL_2, SH_2)\, P(\neg LD_2 \mid LD_1, AL_2, SH_2) \tag{2}$$

Clearly, to be able to use the given default rules, the first and last term in (2) must undergo the following two transformations of context:

$$P(AL_3 \mid LD_2, LD_1, AL_2, SH_2) = P(AL_3 \mid LD_2, AL_2, SH_2) = \varepsilon_3 \tag{3}$$

$$P(\neg LD_2 \mid LD_1, AL_2, SH_2) = P(\neg LD_2 \mid LD_1) = \varepsilon_1 \tag{4}$$

The first states that the effect of the shooting depends only on the state of the gun at time t_2, not on its previous history. The second asserts that the truth of AL_2 and SH_2 does not diminish the likelihood of the gun to remain loaded at t_2, given that it is loaded at t_1.

Assuming that (3) and (4) are permissible (justification will follow), the desired conclusion (1) is obtained immediately. Substituting (3) and (4) in (2) yields:

$$P(AL_3 \mid LD_1, AL_2, SH_2) = \varepsilon_3(1 - \varepsilon_1) + P(AL_3 \mid \neg LD_2, LD_1, AL_2, SH_2)\,\varepsilon_1$$
$$\leq \varepsilon_3 + \varepsilon_1$$
$$= low$$

which confirms (1).

One can easily imagine situations where (3) or (4) are violated, e.g., that the gun user is known to be an extra cautious individual and would not pull the trigger (SH_2) before making sure that the gun is unloaded at t_2. However, the main point is not to invent fanciful violations of expectation but rather to formulate the general principles which govern our normal expectation. In other words, what general principles allow us to posit the validity of (3) and (4), while rejecting the alternative yet symmetrical assumption:

$$P(AL_3 \mid LD_1, AL_2, SH_2) = P(AL_3 \mid AL_2) = 1 - \varepsilon_2 \tag{5}$$

reflecting the persistence of life under normal conditions. Such principles have not been explicated in the probabilistic literature, where it is often assumed that all conditional probabilities are either available or derivable from a complete distribution function.

Cast in probabilistic terms, three such principles can be identified:

$P-1$: propositions not mentioned explicitly in the default rules represent possibilities which are summarized in the numerical values of the probabilities involved; e.g., the possibility that someone has emptied the gun between t_1 and t_2 is summarized by ε_1.

$P-2$: Dependencies not mentioned explicitly are presumed to be *independencies* (provided they are consistent with mentioned dependencies); e.g., AL_3 is presumed to be independent of LD_1, given LD_2, AL_2 and SH_2, (thus justifying (3)), because no direct influence between LD_1 and AL_3 is given explicitly. However, the two cannot be assumed to be unconditionally independent; that will violate the dependencies embodied in d_1 and d_3.

$P-3$: The directionality of the default rules is presumed to represent a *causal structure*. Probabilistically interpreted, this means that there exists some total order θ of the propositions in the system, consistent with the orientation of the default rules, such that propositions mentioned as direct justifications (antecedents) of an event E render E conditionally independent of all its predecessors in θ. θ can be thought of as a temporal precedence, along which the present is presumed to be sufficiently detailed to render the future independent of the past. However, an identical interpretation also applies to non-temporal hierarchies of property inheritance.

This latter principle has far reaching ramifications, stemming from the logic of conditional independence [2]. One of its corollaries is that the existence of *one* ordering θ satisfying the independence conditions of $P-3$ guarantees that the conditions are satisfied in *every* ordering consistent with the orientation of the rules. In other words, we do not have to know the actual chronological order of events in the system; given the truth of all propositions mentioned as antecedents of event E, the probability that E will materialize is not affected by any other proposition in the system except, of course, by E's own consequences. For example, LD_2 is presumed to be independent of AL_2 and SH_2, given LD_1, because LD_1 is mentioned as

the only cause (justification) of LD_2 while AL_2 and SH_2 are not mentioned as consequences of LD_2. On the other hand, AL_3 is *not* independent of SH_2 given AL_2, because SH_2 is explicitly mention as a direct cause of $\neg AL_3$ in rule d_3. Thus, the transformations (3) and (4) are licensed by principles $P-1$ to $P-3$ while (5) is rejected.

3. CONCLUDING DIALOGUE

Logicist: I am quite intrigued by the $P(\cdot|\cdot)$ notation you employ to keep track of varying contexts, it reminds me of how TMS's keep track of justifications. But, getting to the bottom of things, what really makes your system prefer the persistence of "loaded" over the persistence of "alive"?

Probabilist: My system hates to interrupt the persistence of life in much the same way that it tries to minimize all abnormal events. But, as you well know, simply minimizing the number of abnormal events is a bad policy; what needs to be minimized is *conditional* abnormality, namely, abnormality in the context of all known facts. Under normal circumstances, clipping one's life is indeed abnormal. But we are not dealing here with normal circumstances because two input facts are known to have occurred "shoot" and "gun loaded at t_1", and there is no rule stating that life tends to persist in this, more refined context.

Logicist: But circumstances are hardly ever "normal"; in the course of any reasoning activity we are always going to have new facts floating around that were not

explicitly spelled out by the rules. How do you ever get to use any of the rules if its specified context does not match exactly the context created by all the new facts.

Probabilist: I have the logic of probabilistic *independence* here to help me. It permits me to identify and prune away irrelevant facts from the current context so as to match it with the context specified by the rules. This is how I managed to show that ''shoot'' is irrelevant to the persistence of ''loaded'' (Eq. (4)). I could not show, though, that ''shoot'' is irrelevant to the persistence of ''alive'' because the rules (e.g., d_3) tell me that "shoot" is capable (together with ''loaded'') of interfering with ''alive''.

Logicist: I meant to ask you about this biased treatment. The rules also tell you that ''shoot'' is capable (together with other facts) of interfering with ''loaded''; if we find the victim alive at t_3 then, by virtue of knowing ''shoot'', we can conclude ''unloaded''. Thus, it seems to me that, contrary to (4), ''shoot'' and ''loaded'' are not entirely independent. Now, since we have no rule stating that guns tend to remain loaded under contexts involving ''shoot'', shouldn't the persistence of ''loaded'' be questioned the way the persistence of ''alive'' was?

Probabilist: Here is where causality comes in as yet another information source about relevancies. Writing rule d_3 with ''shoot'' and ''loaded'' as antecedents makes me assume that the two are causally affecting ''alive''. Now, we

have strict laws of how to interpret causal information in terms of independence relations. One of these laws tells us that an event with no antecedents is independent of all other events except its own consequences. This means that "shoot" and "loaded" are independent events while "shoot" and "alive" are not. [It would be worthwhile if you could spend few minutes examining the logic of causal-dependencies [2]; it is really quite simple].

Logicist: You mean to tell me that you draw all causal information from the directions of the rules? This means that they must be acyclic and that I have to be very careful about using contraposition.

Probabilist: I would much rather extract causal information directly from temporal precedence, like you folks are doing; it would make things much easier for me. But if temporal information is not available, I rely on the directionality of the rules, as most people would do, and then, yes, one must be careful. For example, had you written rule d_3 in its contrapositive form

$$alive(t_2) \wedge alive(t_3) \wedge shoot(t_2) \rightarrow unloaded(t_2)$$

without warning me that the rule now conveys diagnostic rather than causal information. I would be led to believe that "shoot" has some causal influence over "loaded". Moreover, finding no arrow from "shoot" to "alive", I would also conclude that "shoot" and "alive" are independent events, i.e., "shoot" being incapable of clipping "alive".

Logicist: When you come down to it, the bottom line reason why you ruled out "shoot" as a potential interference with "loaded" is because the two interact only if the victim was seen alive at t_3 and one of your independence laws says that interactions mediated via unconfirmed future events can be discounted. Isn't this equivalent to Shoham's scheme of "Chronological Ignorance" [3] whereby one sweeps forward in time and minimizes the number of abnormal events while ignoring, as much as possible, the effect of future events.

Probabilist: Yes. The right to ignore unconfirmed future events is definitely a common feature of both schemes, but I am not sure at this point whether "Chronological Ignorance" captures *all* the context transformations licensed by the probabilistic interpretation of causality; the latter also teaches us how to manage facts that can't be ignored by chronological considerations. Nevertheless, the logic of probabilistic independence does give Shoham's scheme its operational and probabilistic legitimacy.

4. REFERENCES

[1] Hanks, S., & McDermott, D., "Default Reasoning, Nonmonotonic Logics, and the Frame Problem," *Proc. AAAI-86,* Philadelphia, 1986, pp. 328-333.

[2] Pearl, J. and Verma, A., "The Logic of Representing Dependencies by Directed Graphs," in *Proc. AAAI-87, Seattle, WA,* July 1987, pp. 374-379.

[3] Shoham, Y., "Chronological Ignorance: Time, Nonmonotonicity, Necessity, and Causal Theories," *Proc. AAAI-86,* Philadelphia, 1986, pp. 389-393.