Probabilistic Semantics for Qualitative Influences

Michael P. Wellman[1]
MIT Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139

Abstract

What's in an influence link? To answer this foundational question, I propose a semantics for qualitative influences: *a* positively influences *b* if and only if the posterior distribution for *b* given *a* increases with *a* in the sense of first-order stochastic dominance. By requiring that this condition hold in all contexts, we gain the ability to perform inference across chains of qualitative influences. Under sets of basic desiderata, the proposed definition is necessary as well as sufficient for this desirable computational property.

I. Introduction

Innumerable AI programs incorporate constructs that are intended to capture the notion that one variable "causes" or influences another in some particular fashion, at precisions ranging from the mere direction of influence to exact numerical relationships. Although such terms as "cause" and "influence" are often defined rather loosely in knowledge language specifications, any inference procedure that manipulates models containing these terms imposes constraints on their possible meaning.

In the sections below, I investigate the constraints imposed on a semantics for qualitative probabilistic influences by the most basic properties of typical inference algorithms. *Qualitative* influences are those at the imprecise end of the spectrum, asserting only a direction of association among variables. In looking for a *probabilistic* semantics we admit models where the directions are not guaranteed, and the functional relationships are not deterministically fixed.

II. Example: The Digitalis Therapy Advisor

Our discussion of qualitative influences is set in the context of a simple causal model taken from Swartout's program for digitalis therapy [Swartout, 1983]. The model, shown in Figure 1, is a fragment of the knowledge base that Swartout used to re-implement the Digitalis Therapy Advisor [Gorry *et al.*, 1978] via an automatic programmer.

[1]Supported by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine.

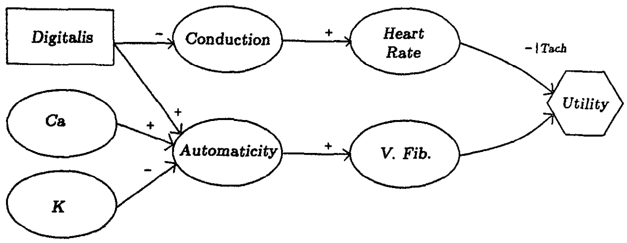

Figure 1: Part of the causal model for digitalis therapy. The direction on a link from *a* to *b* indicates the effect of an increase in *a* on *b*.

In the figure the elliptical nodes represent random variables. The rectangular node is a decision variable, in this case the dosage of digitalis administered to the patient. The hexagonal node is called the value node and represents the utility of the outcome to the patient. This terminology and notation are adapted from *influence diagrams* [Shachter, 1986], a probabilistic modeling formalism similar to Bayes networks [Pearl, 1986a].

Influences among the variables are indicated by dependence links, annotated with a sign denoting the direction of influence. Thus digitalis negatively influences conduction and positively influences automaticity. The former is the desired effect of the drug, because a decrease in conduction *decreases* the heart rate, which is considered beneficial for patients with tachycardia (tach), the population of interest here. The desirability of lower heart rates is represented by the negative influence on the value node (given tach), asserting that lower rates increase expected utility. The increase in automaticity is an undesired side-effect of digitalis because this variable is positively related to the probability of ventricular fibrillation (v. fib.), a life-threatening cardiac state. Calcium (Ca) and potassium (K) levels also influence the level of automaticity.

One of the primary advantages of encoding the digitalis model qualitatively is *modularity*, a knowledge representation issue of particular concern in the case of uncertainty [Heckerman and Horvitz, 1987]. While the exact probabilistic relationships among these variables vary from patient to patient, the *direction* of the relations are reliably taken as constant. Conclusions drawn from this model are therefore valid for a much broader class of patients.

The conclusions we would like our programs to derive from the digitalis model are those taken for granted in the description above. For example, we unthinkingly assumed that the effects of digitalis on conduction and of conduction on heart rate would combine to imply that digitalis reduces the heart rate. Further, because lower heart rates are desirable, digitalis is *therapeutic* along the upper path. Similarly, it is *toxic* along its lower path to the value node. The tradeoff between therapy and toxicity cannot be resolved by the mere qualitative influences in the model.

The remainder of this paper develops a semantics for qualitative influences that justifies the kinds of inferences we require while providing the maximum possible degree of modularity. A formalism for qualitative influences among binary events was introduced in a previous paper [Wellman, 1987]. In the sections below I present the basic definitions, extending them to cover multi-valued parameters. The resulting definition is shown to be the weakest that satisfies our inference desiderata.

III. Qualitative Influences Defined

Consider two random variables, a and b. Informally, when a and b are dichotomous events, a qualitative influence is a statement of the form "a makes b more (or less) likely." This binary case is easy to capture in a probabilistic assertion. Let A and $\bar{A}$ denote the assertions $a = \textbf{true}$ and $a = \textbf{false}$, respectively, and similarly, B and $\bar{B}$. Then we say "a positively influences b," written $S^+(a,b)$, if and only if

$$\forall x \ \Pr(B|Ax) \geq \Pr(B|\bar{A}x). \tag{1}$$

In the equation x ranges over all assignments to the other event variables consistent with both A and $\bar{A}$. The quantification is necessary to assert that the influence holds in all contexts, not just marginally. Because of this context variable, S^+ holds in a particular influence network programs that alter the structure of the network may exhibit non-monotonicity in S^+ [Grosof, 1987]. Conditions analogous to (1) and those following serve to define negative and zero influences, omitted here for brevity.

For the dichotomous case, Bayes's rule implies that (1) is equivalent to

$$\forall x \ \Pr(A|Bx) \geq \Pr(A|\bar{B}x). \tag{2}$$

In the terminology of Bayesian revision (1) is a condition on *posteriors*, while (2) is a condition on *likelihoods*. Notice that $S^+(a,b)$ is simply an assertion that the *likelihood ratio* is greater than or equal to unity.

Formalizing S^+ is not quite so straightforward when a and b take on more than two values. In such cases we want to capture the idea that "higher values of a make higher values of b more likely." An obvious prerequisite for such statements is some interpretation of "higher." Therefore, we require that each random variable be associated with an order $\geq$ on its values. For numeric variables such as "potassium concentration," this relation has the usual interpretation; for variables like "automaticity" a measurement scale and ordering relation must be contrived.

The more troublesome part of defining positive influences is in specifying what it means to "make higher values of b more likely." Intuitively, we want a statement that the probability distribution for b shifts toward higher values as a increases. To make such a statement, we need an ordering that, given any two cumulative probability density functions (CDFs) G_1 and G_2 over b, determines whether G_1 is "higher" than G_2.

However, not all probability distributions can be easily ordered according to the size of the random variable. Different rankings are obtained through comparing distributions by median, mean, or mean-log, for example. We require an ordering that is robust to changes of these measures because the random variables need be described by merely *ordinal* scales [Krantz *et al.*, 1971]. An assertion that calcium concentration positively influences automaticity should hold whether calcium is measured on an absolute or logarithmic scale, and regardless of how automaticity is measured.

An ordering criterion with the robustness we desire is *first-order stochastic dominance* (FSD) [Whitmore and Findlay, 1978]. FSD holds for G_1 over G_2 if and only if the mean of G_1 is greater than the mean of G_2 for *any* monotonic transform of b. That is, for all monotonically increasing functions ϕ,

$$\int \phi(b)dG_1(b) \geq \int \phi(b)dG_2(b). \tag{3}$$

A necessary and sufficient condition for (3) is

$$\forall b \ G_1(b) \leq G_2(b). \tag{4}$$

That is, for any given value of b the probability of obtaining b or less is smaller for G_1 than for G_2. For further discussion and a proof that (4) is equivalent to (3), see [Fishburn and Vickson, 1978].

We are now ready to define qualitative influences. Let $F(b|a_i x)$ be the CDF for b given $a = a_i$ and context x.[2] Then $S^+(a,b)$ if and only if

$$\forall a_1, a_2, x \quad a_1 \geq a_2 \Rightarrow F(b|a_1 x) \text{ FSD } F(b|a_2 x). \tag{5}$$

Adopting the convention for binary events that $\textbf{true} > \textbf{false}$, we can verify that (5) is a generalization of (1).

Like (1), (5) is a condition on posteriors. Milgrom [Milgrom, 1981] proves that the equivalent likelihood condition is the Monotone Likelihood Ratio Property (MLRP) from statistics [Berger, 1985].

[2] As above, x is an assignment to the remaining random variables consistent with the condition $a = a_i$. We need to include x here and in the definitions below because these conditions will be applied in situations where x is partially or totally known. If we had stated the conditions in marginal terms ("on average, a positively influences b"), it would not be valid to apply them in specific contexts.

Finally, we need a special definition for influences on the value node. The variable a positively influences utility, $U^+(a)$, if and only if

$$\forall a_1, a_2, x \quad a_1 \geq a_2 \Rightarrow u(a_1, x) \geq u(a_2, x), \tag{6}$$

where u is a utility function [Savage, 1972] defined over the event space.

IV. Chains of Influence

In an earlier paper on qualitative influences [Wellman, 1987], I considered networks of variables connected by influence links describing the direction of probabilistic dependence. There, I demonstrated for the binary case that, in the absence of direct links from a to b, $S^+(a,b) \wedge S^+(b,c) \Rightarrow S^+(a,c)$. From a computational perspective the ability to perform inference across influence chains is an essential property of a qualitative algebra. From the digitalis model, for example, we would like to deduce that increasing the dose of digitalis decreases the heart rate but increases the likelihood of v. fib. Indeed, most programs with models like this would make such an inference. Fortunately, the definition offered above for S^+ implies transitivity for multi-valued as well as binary variables.

Proposition 1 *If a and c are not connected by any direct links, $S^+(a,b)$, and $S^+(b,c)$, then $S^+(a,c)$ holds in the network obtained by removing b.*

Proof: Choose a_1 and a_2 such that $a_1 \geq a_2$, and an x_0 consistent with a_1, a_2, and all b. Let G denote the conditional CDF for c and $\underline{c}$ the minimal value of the variable. By the definition of cumulative probability we have

$$G(c_0|a_i x_0) = \int_{\underline{c}}^{c_0} \int f_{bc}(bc|a_i x_0) db dc. \tag{7}$$

Changing the order of integration and decomposing the joint probability yields

$$G(c_0|a_i x_0) = \int \int_{\underline{c}}^{c_0} f_c(c|a_i b x_0) f_b(b|a_i x_0) dc db. \tag{8}$$

Because a and c are conditionally independent given b and x, we can remove a_i from the f_c expression.[3] Rewriting the density function as the derivative of a cumulative, we get

$$G(c_0|a_i x_0) = \int \int_{\underline{c}}^{c_0} f_c(c|b x_0) dc dF(b|a_i x_0). \tag{9}$$

The inner integral is simply the CDF for c given b.

$$G(c_0|a_i x_0) = \int G(c_0|b x_0) dF(b|a_i x_0). \tag{10}$$

[3]The conditional independence follows from separation in the influence network. See Pearl [Pearl, 1986b] for a discussion of the independence properties of graphical probability representations.

Because b positively influences c, the pointwise FSD condition (4) implies that for any c_0, $G(c_0|b x_0)$ is a *decreasing* function of b. And $S^+(a,b)$ entails FSD of $F(b|a_1 x_0)$ over $F(b|a_2 x_0)$. Therefore, (3) applies with the inequality reversed (negating $G(c_0|b x_0)$ yields an increasing function), leading to the conclusion

$$\forall c_0 \quad G(c_0|a_1 x_0) \leq G(c_0|a_2 x_0), \tag{11}$$

implying FSD. Because a_1, a_2, and x_0 were chosen arbitrarily, we have finally $S^+(a,c)$. ∎

Similar arguments with the appropriate signs and directions switched would reveal that chains of influences may be combined by sign multiplication.

In the remainder of this section I present some simple desiderata for a qualitative influence definition that entail the *necessity* of FSD for chaining influences. We start by specifying the form such definitions must take. To capture the intent of "higher values of a make higher values of b more likely" in a probabilistic semantics, it seems reasonable to restrict our attention to conditions on the posterior distribution of b for increasing values of a. Therefore, we postulate that a definition of $S^+(a,b)$ must be of the form

$$\forall a_1, a_2, x \quad a_1 \geq a_2 \Rightarrow F(b|a_1 x) \; R \; F(b|a_2 x), \tag{12}$$

where R is some relation on CDFs. This condition is exactly (5) with FSD replaced by the more abstract relation.

There are two basic desiderata that severely restrict the possible Rs. First, the definition for S^+ induced by R in (12) must satisfy Proposition 1. Without the ability to chain inferences, the qualitative influence formalism has little computational value. Second, the condition must be a generalization of the original definition of S^+ for dichotomous events (1). With only two possible values there does not appear to be a weaker monotonicity condition. These criteria lead to a sharp conclusion.

Proposition 2 *Let $S^+(a,b)$ be defined by (12). Given the following conditions:*

1. *Proposition 1*
2. *For binary b, $a_1 \geq a_2$, and x,*

$$F(b|a_1 x) \; R \; F(b|a_2 x) \Leftrightarrow \Pr(B|a_1 x) \geq \Pr(B|a_2 x) \tag{13}$$

the weakest R is FSD.

Proof: First, note that FSD satisfies these conditions. Next, assume that R satisfies them but R does not entail FSD. We will start with an instantiation of Proposition 1 and derive a contradiction. Let a, b, and c be the only variables (so we can safely ignore x) with $S^+(a,b)$, $S^+(b,c)$, and no other direct links. For concreteness, let b range over the unit interval $[0,1]$ and c be binary with $\Pr(C|b) = \phi(b)$, for some $\phi : [0,1] \to [0,1]$ monotonic. The monotonicity of ϕ guarantees $S^+(b,c)$. By assumption, Proposition 1 applies, yielding the conclusion $S^+(a,c)$ and therefore $F(c|a_1) \; R \; F(c|a_2)$. Because c is binary, (13) must hold. Expanding the expression for the posterior probability of c given a, the RHS of (13) becomes

$$\int_0^1 \phi(b)dF(b|a_1) \geq \int_0^1 \phi(b)dF(b|a_2). \qquad (14)$$

Because ϕ may be any monotonic function, FSD is necessary for (14) and is therefore entailed by R. ∎

The force of this result is weakened somewhat by the *a priori* restriction of definitions to those having the form of (12). Many statistical concepts of directional relation (based on correlation or joint expectations, for example) do not fit (12) yet appear to be plausible candidates for a definition of qualitative influence. *Quadrant dependence* [Lehmann, 1966] holds between a and b when[4]

$$\forall a_1, a_2 \quad a_1 \geq a_2 \Rightarrow F(b|a \leq a_1) \text{ FSD } F(b|a \leq a_2). \qquad (15)$$

Lehmann proves that quadrant dependence is necessary but not sufficient for *regression dependence*, which is his terminology for (5) without the quantification over contexts x. As quadrant dependence is weaker yet still exhibits transitivity,[5] it seems to be an attractive alternate to regression dependence. To justify our choice of the latter, we must appeal to the decision-making implications of probabilistic models.

V. Making Decisions

The prime motivation for adopting a probabilistic semantics is so that the behavior of our programs can be justified by Bayesian decision theory [Savage, 1972]. A decision is valid with respect to an influence model if expected utility is maximized. For example, if $U^+(a)$ and there are no *indirect* paths from a to the value node, then a decision of a_1 over a_2 is valid if and only if $a_1 \geq a_2$, by the definition of U^+ (6).[6] Decision-making power is enhanced if we can deduce new influences on utility from chains of influences in the network. Our definition of qualitative influence is necessary as well as sufficient for such inferences.

Proposition 3 *Consider a network where $U^+(b)$ holds and a and u are not connected by any direct links. A necessary and sufficient condition for $U^+(a)$ on removal of b is $S^+(a,b)$ as defined by (5).*

Proof: The expected utility of a_i with any x is given by

$$u(a_i, x) = \int u(b,x)dF(b|a_i x). \qquad (16)$$

[4]This is actually the condition Lehmann proposes as a strengthening of quadrant dependence. The basic quadrant dependence fixes a_1 at a's maximal value.

[5]For transitivity we need to quantify over contexts in (15). The proof parallels that for Proposition 1.

[6]The existence of other paths from a to utility would leave open the possibility that the net influence of a is negative. For example, we could summarize the therapeutic effect of digitalis through conduction and heart rate as a direct positive influence. But this might be outweighed by the indirect negative influence of digitalis via automaticity.

$U^+(a)$ is satisfied in the reduced network if and only if $u(a_i, x)$ is increasing in a_i. From (6) we know that $u(b,x)$ is monotonically increasing in b. In fact, it can be *any* monotonic function. Therefore, (16) is increasing in a_i under the same conditions as (3), which is exactly S^+ as defined by (5). ∎

Proposition 3 demonstrates that while conditions like quadrant dependence which are weaker than S^+ may be sufficient for propagating influences across chains, they are not adequate to justify *decisions* across chains. For choosing among alternatives, the relevant parameter is the utility function evaluated at a point; utilities conditioned on intervals of the decision variable (as in quadrant dependence) do not have the same decision-making import.

VI. Extensions

The basic definitions above can be extended in a variety of ways. Conditional influences—defined for binary events in a previous paper [Wellman, 1987]—simply delimit the range of x in (5). For example, the negative influence of heart rate on utility in the digitalis model is conditional on tachycardia.

Swartout's XPLAIN knowledge base included the "domain principle" that if a state variable acts synergistically with the drug to induce toxicity, then smaller doses should be given for higher observations of the variable [Swartout, 1983]. This fact could be derived by a *domain-independent* inference procedure given a suitable definition for qualitative synergy. We can say that two variables synergistically influence a third if their joint influence is greater (in the sense of FSD) than separate statistically independent influences.[7] In the digitalis example, we need to assert that digitalis acts at least independently with Ca and K deviations in increasing automaticity. In addition, we must specify that the decrease in utility for a given increase in automaticity is larger when automaticity is already high. Such a relationship can be captured by an assertion that automaticity is synergistic with itself in its toxic effects.

VII. Related Work

Philosophers have long attempted to develop mathematical definitions of causality. Motivated by computational rather than philosophical concerns, I have ignored in this treatment temporal properties, mechanisms, spuriousness, and other issues salient to causality. These concerns aside, Suppes [Suppes, 1970] proposes a probabilistic condition equivalent to (1) without the context quantification for binary events. For multi-valued variables, Suppes suggests quadrant dependence (15).

As suggested previously, ordering of random variables has also attracted considerable interest in statistics [Berger, 1985, Lehmann, 1966, Ross, 1983] and decision theory [Whitmore and Findlay, 1978]. Milgrom [Mil-

[7]This type of relationship was exploited by NESTOR [Cooper, 1984], a diagnostic program based on probabilistic inequalities.

grom, 1981] demonstrates the application of MLRP to theoretical problems in informational economics.

The key difference between the S^+ definition proposed here and previous work is that we obtain transitivity by requiring the condition to hold in all contexts. Suppes shows that the causal algebra induced by his condition—defined only at the margin—does not possess the transitive property. As argued above, this is a computationally essential characteristic of qualitative influences.

VIII. Conclusions

Despite the ubiquity of qualitative influence assertions in knowledge representation mechanisms, there has been little study of the semantics of such constructs. Previous work either denies the probabilistic nature of the relationships among variables in the model or takes for granted the ability to draw inferences by chaining influences in the network. I have defined a positive qualitative influence of a on b as an assertion that, in all contexts, the posterior probability distribution for b given a is stochastically increasing (FSD) in a. A series of propositions provided theoretical support for this S^+ definition:

- S^+ supports chaining of influences.
- S^+ is the weakest posterior condition that supports chaining of influences.
- S^+ is necessary and sufficient for chaining *decisions* across influences.

A semantics for qualitative influences should prove valuable for analyzing knowledge bases like the digitalis model of Figure 1, as well as knowledge representation theories that include similar constructs. In particular, the definition of S^+ can help to evaluate the potential of purely qualitative methods like Cohen's endorsement approach [Cohen, 1985] and to characterize the techniques from AI work on qualitative reasoning that are valid in probabilistic domains.

Acknowledgments

Zak Kohane, Bill Long, Ramesh Patil, Elisha Sacks, Kate Unrath, and Alex Yeh positively influenced this paper.

References

[Berger, 1985] James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, second edition, 1985.

[Cohen, 1985] Paul R. Cohen. *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach.* Volume 2 of *Research Notes in Artificial Intelligence*, Pitman, 1985.

[Cooper, 1984] Gregory Floyd Cooper. NESTOR: *A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge.* PhD thesis, Stanford University, November 1984.

[Fishburn and Vickson, 1978] Peter C. Fishburn and Raymond G. Vickson. Theoretical foundations of stochastic dominance. In [Whitmore and Findlay, 1978].

[Gorry *et al.*, 1978] G. Anthony Gorry, Howard Silverman, and Stephen G. Pauker. Capturing clinical expertise: A computer program that considers clinical responses to digitalis. *American Journal of Medicine*, 64:452–460, March 1978.

[Grosof, 1987] Benjamin N. Grosof. Non-monotonicity in probabilistic reasoning. In John F. Lemmer, editor, *Uncertainty in Artificial Intelligence*, North-Holland, 1987.

[Heckerman and Horvitz, 1987] David E. Heckerman and Eric J. Horvitz. The myth of modularity in rule-based systems. In John F. Lemmer, editor, *Uncertainty in Artificial Intelligence*, North-Holland, 1987.

[Krantz *et al.*, 1971] David H. Krantz, R. Duncan Luce, Patrick Suppes, et al. *Foundations of Measurement.* Academic Press, New York, 1971.

[Lehmann, 1966] E. L. Lehmann. Some concepts of dependence. *Annals of Mathematical Statistics*, 37:1137–1153, 1966.

[Milgrom, 1981] Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12:380–391, 1981.

[Pearl, 1986a] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[Pearl, 1986b] Judea Pearl. *Markov and Bayes networks: A comparison of two graphical representations of probabilistic knowledge.* Technical Report R-46, UCLA Computer Science Department, September 1986.

[Ross, 1983] Sheldon M. Ross. *Stochastic Processes.* John Wiley and Sons, 1983.

[Savage, 1972] Leonard J. Savage. *The Foundations of Statistics.* Dover Publications, New York, second edition, 1972.

[Shachter, 1986] Ross D. Shachter. Evaluating influence diagrams. *Operations Research*, 34, 1986.

[Suppes, 1970] Patrick Suppes. *A Probabilistic Theory of Causality.* North-Holland Publishing Co., Amsterdam, 1970.

[Swartout, 1983] William R. Swartout. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21:285–325, 1983.

[Wellman, 1987] Michael P. Wellman. Qualitative probabilistic networks for planning under uncertainty. In John F. Lemmer, editor, *Uncertainty in Artificial Intelligence*, North-Holland, 1987.

[Whitmore and Findlay, 1978] G. A. Whitmore and M. C. Findlay, editors. *Stochastic Dominance: An Approach to Decision Making Under Risk.* D. C. Heath and Company, Lexington, MA, 1978.