

Thinking Backward for Knowledge Acquisition

Ross D. Shachter and David E. Heckerman

This article examines the direction in which knowledge bases are constructed for diagnosis and decision making. When building an expert system, it is traditional to elicit knowledge from an expert in the direction in which the knowledge is to be applied, namely, from observable evidence toward unobservable hypotheses. However, experts usually find it simpler to reason in the opposite direction—from hypotheses to unobservable evidence—because this direction reflects causal relationships. Therefore, we argue that a knowledge base be constructed following the expert's natural reasoning direction, and then reverse the direction for use. This choice of representation direction facilitates knowledge acquisition in deterministic domains and is essential when a problem involves uncertainty. We illustrate this concept with influence diagrams, a methodology for graphically representing a joint probability distribution. Influence diagrams provide a practical means by which an expert can characterize the qualitative and quantitative relationships among evidence and hypotheses in the appropriate direction. Once constructed, the relationships can easily be reversed into the less intuitive direction in order to perform inference and diagnosis. In this way, knowledge acquisition is made cognitively simple; the machine carries the burden of translating the representation

A few years ago, we were discussing probabilistic reasoning with a colleague who works in computer vision. He wanted to calculate the likelihood of a tiger being present in a field of view given the digitized image. "OK," we replied, "If the tiger were present, what is the probability that you would see that image? On the other hand, if the tiger were not present, what is the probability you would see it?" Before we could say "what is the probability there is a tiger in the first place?" our colleague threw up his arms in despair. "Why must you probabilists insist on thinking about everything backwards?"

Since then, we have pondered this question. Why is it that we want to look at problems of evidential reasoning backward? After all, the task of evidential reasoning is, by definition, the determination of the validity of unobservable propositions from observable evidence; it seems best to represent knowledge in the direction it will be used. Why then should we represent knowledge in the opposite direction, from hypothesis to evidence?

In this article, we attempt to answer this question by showing how some backward thinking can simplify reasoning with expert knowledge.¹ We believe that the best representation for knowledge acquisition is the simplest representation which captures the essence of an expert's beliefs. We argue that in many cases, this representation will correspond to a direction of reasoning that is opposite the direction in which expert knowledge is used in uncertain reasoning.

This question has relevance to artificial intelligence applications because several popular expert system architectures represent knowledge in

the direction from observables to unobservables. For example, in the MYCIN certainty factor (CF) model (Shortliffe and Buchanan 1975), knowledge is represented as rules of the form

IF <evidence> THEN
<hypothesis>, CF,

where CF is a measure of the change in belief in the hypothesis given the evidence. Observable propositions are most often found in the antecedent of rules, and unobservable propositions are usually found in the consequent. We argue it is often much simpler to represent knowledge in the direction opposite the direction of usage, that is, in the direction IF <hypothesis> THEN <evidence>.²

Our argument is based on three observations. First, many real-world problems involve causal interactions. In this article, we make no attempt to define causality in terms of basic concepts; we take it to be a primitive and subjective notion. Second, the direction of causality is most often opposite the direction of usage; that is, hypotheses tend to cause evidence. Many examples of this statement exist in medicine, where the unobservable hypothesis is the true cause of an illness, and the observable evidence is the illness' effect in the form of symptoms. Of course, there are exceptions. For example, in trauma cases, such as automobile accidents, the cause is observable, and some of the effects are difficult to observe. However, tests to determine the hidden effects of the accident fit the usual model of unobservable cause and observable effect. Third, experts are more comfortable when their beliefs are elicited in the causal direction. It appears that it is cognitively

simpler to construct assessments in this direction (Kuipers and Kassier 1984). Furthermore, representations of expert knowledge are often less complex in the causal direction. Thus, these three observations suggest that there might be advantages to representing knowledge in the direction opposite the direction of usage. This argument is summarized in figure 1

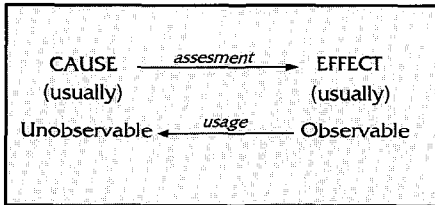


Figure 1.

Two Directions of Representation.

Most real-world problems involve causal interactions in which one event (the cause) affects the outcome of another (the effect). We often use an effect that we can observe to help us learn about a cause which we cannot directly observe. However, we are more comfortable quantifying the relationship between the two events in the causal direction, assessing the likelihood of the effect if only we knew whether the cause were true

Influence Diagrams

In this article, we examine these issues in the context of probability theory. Nonetheless, we believe that the distinction between direction of usage and direction of natural assessment is a fundamental issue, independent of the language in which belief is represented.

Within the theory of probability, several graphical representations exist for uncertainty that feature both directed graphs (Wright 1921; Howard and Matheson 1981; Pearl 1986; Cooper 1984) and undirected graphs (Speed 1978; Spiegelhalter 1985). The different approaches share the basic concept of the factorization of an underlying joint distribution and the explicit representation of conditional independence. We use a directed graph representation method because direction is central to our discussion. In particular, we use the influence diagram representation scheme (Howard and Matheson 1981).

Although the influence diagram has

proven intuitive for knowledge acquisition, the representation is precise and well-grounded in theory. Each of the oval nodes in an influence diagram represents a random variable or uncertain quantity, which can take on two or more possible values. The arcs indicate conditional dependence: a variable's probability distribution depends on the outcomes of its direct predecessors in the graph. For example, three possible influence diagrams exist for the two random variables X and Y shown in figure 2. In the first case, X has no predecessors, so we assess a marginal (unconditional) distribution for X and a conditional distribution for Y given X. In the next case, with the arc reversed, we assess a marginal distribution for Y and a conditional distribution for X given Y.

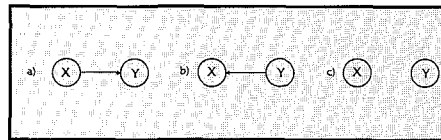


Figure 2.

Influence Diagrams with Two Nodes

The joint probability distribution for two variables can be represented by three different influence diagrams. In diagram a, there is a probability distribution for variable X and a conditional distribution for variable Y given that we observe the value of X (indicated by the arc from X to Y). In diagram b, there is an unconditional distribution for Y and a distribution for X given Y. Diagram c asserts that the two variables are independent because we can obtain their joint distribution from an unconditional distribution for each. Any two variables can be modeled by a or b, but the missing arc in diagram c imposes a strong assumption on the relationship between X and Y, namely, that knowing X will not change our distribution for Y and vice versa.

tribution for X given Y. Both correspond to the same fundamental model at the underlying joint distribution, but the two diagrams represent two different ways of factoring the model. The transformation between them, which involves reversing the direction of conditioning and, hence, the reversal of the arc, is simply Bayes' theorem. Finally, in the third case, neither node has any predecessors, so X and Y are independent. Therefore, we can

obtain the joint by assessing marginal distributions for both X and Y. When the two variables are independent, we are free to use any of the three forms, but we prefer the last one, which explicitly reveals this independence in the graph.

As a detailed example, consider the influence diagram in figure 3. This figure shows the relationship between diabetes and a possible symptom of the disease, blue toe. Associated with the influence diagram are the probability distributions, a marginal (unconditional) distribution for diabetes, and a conditional distribution for blue toe given diabetes.

In figure 4, we see four possible influence diagrams for the uncertain variables X, Y, and Z. In the first case—the general situation—the three variables appear to be completely dependent. We assess a marginal distribution for X and conditional distributions for Y, given X, and Z given X and Y. In general, there are $n!$ factorizations of a joint distribution among n variables. Each possible permutation leads to a different influence diagram. In the second case in the figure, the three variables are completely independent; in the third case, X and Y are dependent, but Z is independent of both of them. In the fourth case, we see conditional independence. The absence of an arc from X to Z indicates that although X and Z are dependent, they are independent given Y. This type of conditional independence is an important simplifying assumption for the construction and assessment of uncertainty models.

Returning to the earlier diabetes example, blue toe is caused by atherosclerosis, which might be caused by diabetes. This relationship is shown in figure 5. The absence of a direct arc from diabetes to blue toe indicates that given the state of atherosclerosis, knowing whether the patient has diabetes tells us nothing about whether the patient has blue toe. The probability distributions needed to completely describe this model are shown below the influence diagram. They are a marginal distribution for diabetes (the same as previously), and conditional distributions for atherosclerosis, given diabetes, and blue toe given atherosclerosis.

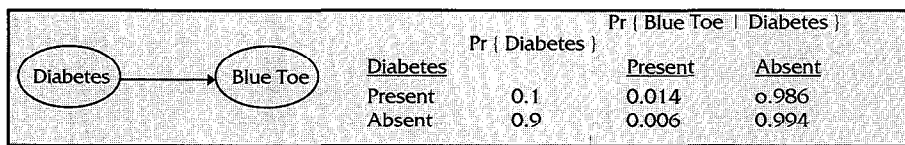


Figure 3
A Simple Influence Diagram for Diabetes.

The influence diagram tells us about the joint behavior of the events diabetes and blue toe. We have an unconditional probability distribution for diabetes: diabetes is present with probability 0.1 and absent with probability 0.9. We also have a conditional distribution for blue toe given that we observe diabetes: if diabetes is present, then blue toe is present with probability 0.014, but if diabetes is absent, then blue toe is present with probability 0.006.

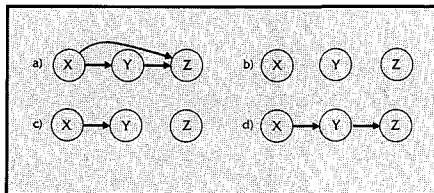


Figure 4
Influence Diagrams with Three Nodes

The joint probability distribution for three variables can be represented by many possible influence diagrams, some of which are shown in figure 4. Diagram a is the general case, and it could apply to any of the three variables: we have an unconditional distribution for X; a conditional distribution for Y, given X; and a conditional distribution for Z given both X and Y. Diagram b is the most restrictive case: it indicates that the three variables are independent because we have an unconditional distribution for all three. Diagram c allows X and Y to be dependent but asserts that Z is independent of X and Y. Finally, diagram d represents conditional independence: X and Z might be dependent, but once we observe Y, we would not learn anything more about Z by seeing X. Z and X are said to be conditionally independent given Y.

In the influence diagram, we always require that there be no directed cycles. By doing so, an ordering always exists among the variables so that we can recover the underlying joint distribution.

One other type of influence diagram node is relevant to our discussion, a deterministic node, drawn as a double oval (as opposed to the probabilistic node, which we have shown as a single oval). The deterministic variable is a function of its predecessors, so its outcome is known with certainty if we can observe the outcomes of these

predecessors. In general, we cannot observe all these predecessors, so there can be uncertainty in the outcome of a deterministic variable.

In an influence diagram, knowledge can be encoded in a comfortable and intuitive direction for the expert and later transformed for use. Such trans-

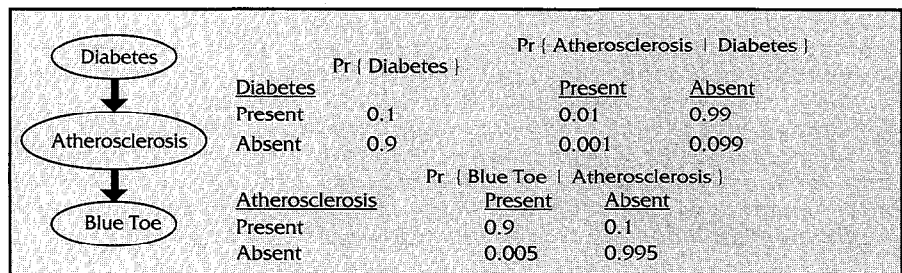


Figure 5
An Influence Diagram Representing Conditional Independence

This influence diagram shows that blue toe is conditionally independent of diabetes given atherosclerosis. Our distribution for blue toe would change if we knew whether diabetes were present. However, once we know whether atherosclerosis is present, finding out about diabetes would tell us nothing new about blue toe. The distribution for blue toe is now conditioned by atherosclerosis. If atherosclerosis is present, then blue toe is present with probability 0.9; otherwise, it is only present with probability 0.005.

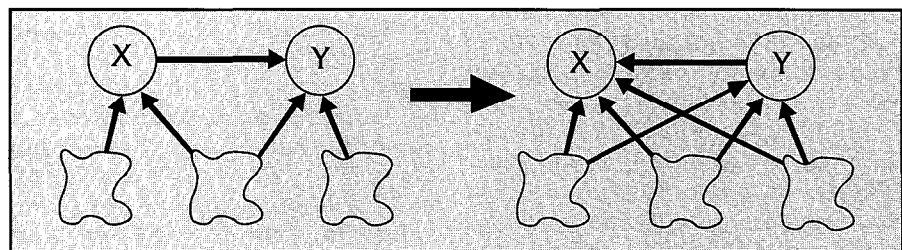


Figure 6
Arc Reversal in an Influence Diagram.

Bayes' theorem is represented in the influence diagram by an arc reversal. In the diagram on the left, Y is conditioned by X. If we want to express the distribution of X conditioned by Y instead, we can reverse the arc between them. In general, both X and Y will have other conditioning variables, indicated by the three sets. (The middle set contains their common conditioning variables.) As a consequence of the arc reversal, the new distributions for X and Y will (in general) be conditioned by all the variables in the three sets. As a rule, the distributions become more complicated after we perform arc reversals.

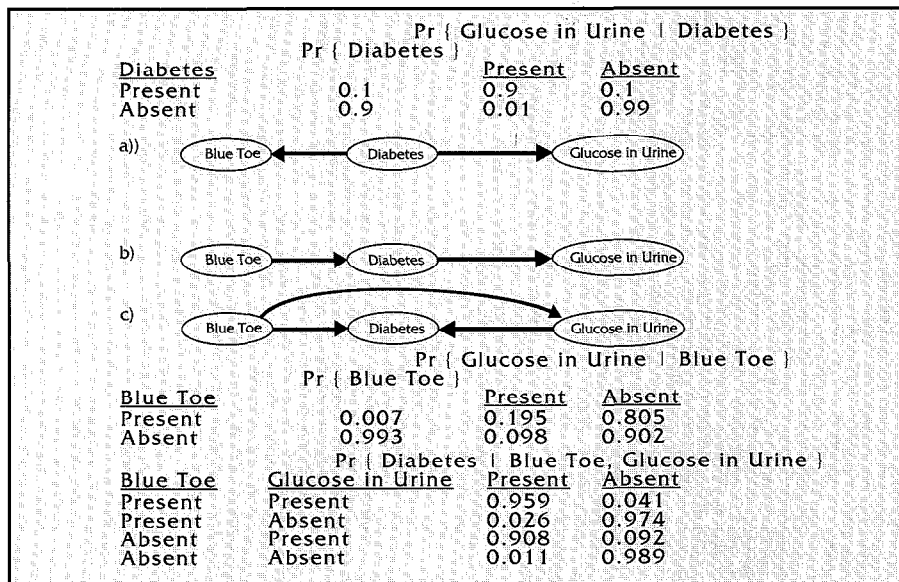


Figure 7
Influence Diagram Transformation for Diagnosis.

Arc reversals permit us to move from an influence diagram that is easy to assess to one which helps with diagnosis. The influence diagram a shows that blue toe and glucose in urine are conditionally independent given diabetes. This fact is also evident in diagram b, which is obtained by reversing the arc from diabetes to blue toe. When the arc from diabetes to glucose in urine is reversed to obtain diagram c, we must also add an arc from blue toe to glucose in urine, and there is no longer any conditional independence indicated in the diagram. The distribution for diabetes is now conditioned by both blue toe and glucose in urine. If they are both present, then diabetes is present with probability .959; if they are both absent, then diabetes is present with probability .001.

sals are needed, then the order in which they are performed can affect this arc structure.³

To continue with the diabetes example, consider the influence diagram shown in figure 7a. We have added a second symptom of diabetes, the presence of glucose in the urine. The conditional probabilities for glucose in the urine, given diabetes, are given below the influence diagram. The two symptoms are shown to be conditionally independent in the diagram given diabetes. If we know whether the patient has diabetes, then the presence of one symptom tells us nothing about the other. Although the diagram is natural for assessment, it cannot be used for diagnosis. We can transform the diagram into one that is useful for diagnosis by first reversing the arc from diabetes to blue toe and then reversing the arc from diabetes to glucose in urine as shown in figures 7b and 7c. When performing the second reversal, an arc is added from blue toe to glucose in urine. The arc is

needed because the symptoms are not independent. When we do not know whether the patient has diabetes, the presence of one symptom tells us much about the other.

Deterministic Models

The importance of the distinction between direction of assessment and direction of usage appears even in the simplest categorical models. Suppose, for example, that we have an error in the output from a small computer program. If we knew the source of the error, then we would know with certainty the type of output error to expect. Thus, we could use the model shown in figure 8, in which the programming error is represented by a probabilistic node conditioning the computer output represented by a deterministic node. When we observe the output and wish to learn about the source of the error, we reverse the arc using Bayes' theorem and find that after the reversal, both nodes have become probabilistic. Given the out-

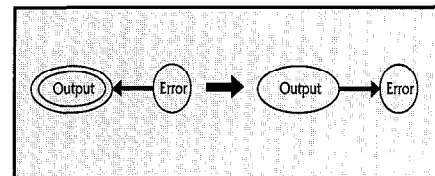


Figure 8.
An Influence Diagram for Diagnosing Computer Programming Errors.

The influence diagram on the left is a simple model of a computer program. Output is a deterministic variable; if we know whether a particular error is present, then we know exactly what output to expect. However, we are uncertain about what the output will be because we are uncertain about the error. Therefore, after we reverse the arc to perform diagnosis, output becomes a probabilistic variable. The error remains probabilistic because the same output could arise from different errors.

put, we do not necessarily know what type of error caused it, but we are able to update our previous beliefs about the possible errors in light of this new evidence. Our direction of usage is clearly in the more complex, reverse direction, but the model is easier to construct in the original direction, which exploits the categorical behavior of our man-made computer system.

Suppose now that we have a much larger computer system. If it were written with a modular design, we might have the influence diagram shown in figure 9. Again, this model is relatively easy to construct because of the categorical nature and the independence among subsystems. If, however, we observe the output and wish to update our knowledge about these subsystems, we find that they are no longer categorical, nor are they independent, in light of the new information.⁴ This newer, more complex model is the correct one to use to update our beliefs as we observe evidence, but it is less convenient for knowledge acquisition.

Probabilistic Models

In most real-world domains, we do not have a categorical model to assess, but there is still considerable advantage to thinking about a problem in the causal direction. Often, basic and

straightforward probabilistic models become complex when viewed in the direction of usage.

Consider the case of two effects with a single common cause. Even when the effects are conditionally independent given the cause, as in figure 10, in general, they are dependent when the problem is reversed for usage. We saw a specific example of this case earlier in figure 7. Similarly, when there are two independent causes and a single common effect, as in figure 11, we see complete dependency when the problem is reversed. Clearly, as the number of causes and effects increases, the problem stays straightforward in the causal direction but becomes complex in the direction of usage.

As an example, consider two disorders—congestive heart failure and nephrotic syndrome (a kidney disease)—that essentially arise independently. Congestive heart failure often results in an enlarged heart (cardiomegaly) and an accumulation of fluid in the ankles (pitting edema), and nephrotic syndrome often leads to protein in the urine and pitting edema as well. A simple test for protein in the urine is whether the urine is frothy; an X ray can detect cardiomegaly. The model corresponding to this problem is shown on the left in figure 12. If we turn the model around to show how the unobservable events of interest—heart failure and nephrotic syndrome—depend on the observables—X ray, pitting edema, and frothy urine—then the model becomes the one shown on the right in figure 12. The original model was not only simpler but more natural to assess, going in the causal direction. The reversed model would be intolerably confusing to assess, but it has all the dependencies one needs for proper usage.

Another major advantage to viewing a problem in different directions for construction and solution is that parts of the assessment might not vary much from case to case. Consider the simple medical example in figure 13, in which the presence or absence of a disorder affects the likelihood that a set of symptoms will manifest. Although the probability distribution for the disorder can vary

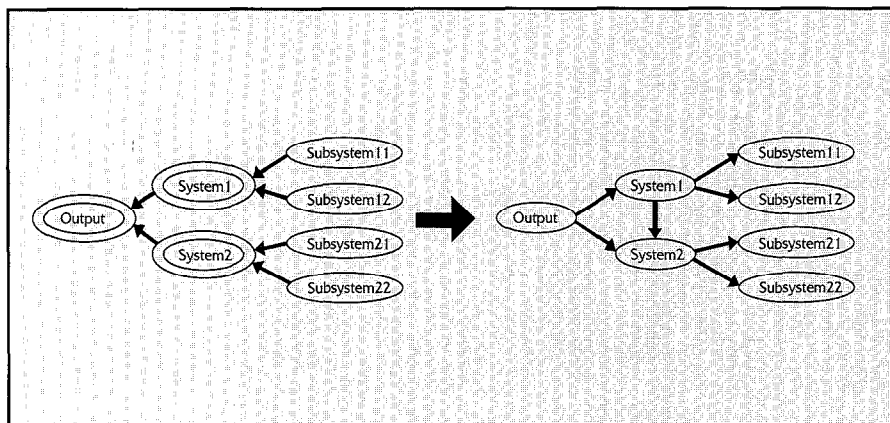


Figure 9
An Influence Diagram for a Modular Computer Program

This influence diagram represents a modular computer system. The diagram on the left shows considerable conditional independence; for example, neither system2 nor subsystem12 will be affected by errors in subsystem11. The new diagnostic diagram obtained after arc reversals shows much less conditional independence. It helps one to understand the relationships among the variables after we observe output. In this case, knowing whether system1 is working could change the distribution for system2; if output reveals an error that is not explained by system1's state, then system2 must be responsible.

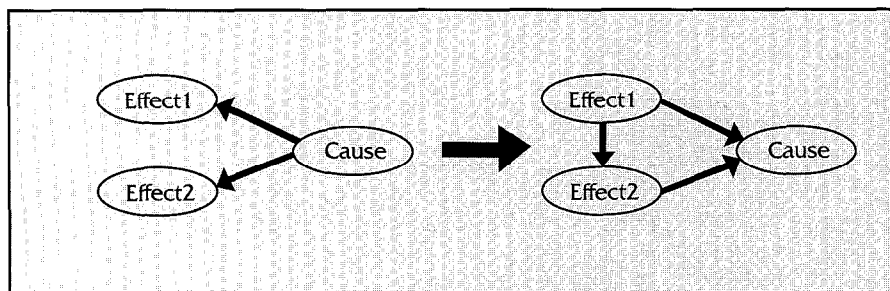


Figure 10.
A Single Cause with Multiple Effects

Figure 10 illustrates a cause for which there are multiple effects. The diagram on the left is easier to assess because the effects are conditionally independent given the true cause. To reason about the cause from the effects, the arcs are reversed to obtain the cause conditioned by the effects. The effects are now dependent because observing one will change the distribution for the cause, which in turn changes the distribution for the other effect.

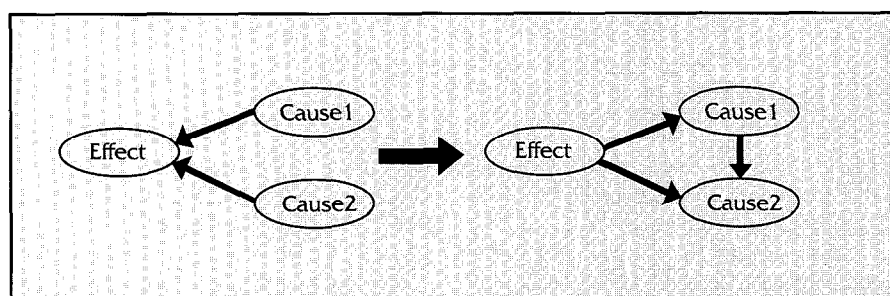


Figure 11.
A Single Effect with Multiple Causes

In this case, multiple causes have a common effect. The diagram on the left is easier to assess because causes are independent. To learn about the causes from the effect requires reversing the arcs. The causes are now dependent because if the effect were observed, then cause2 would be much more likely if cause1 were not true.

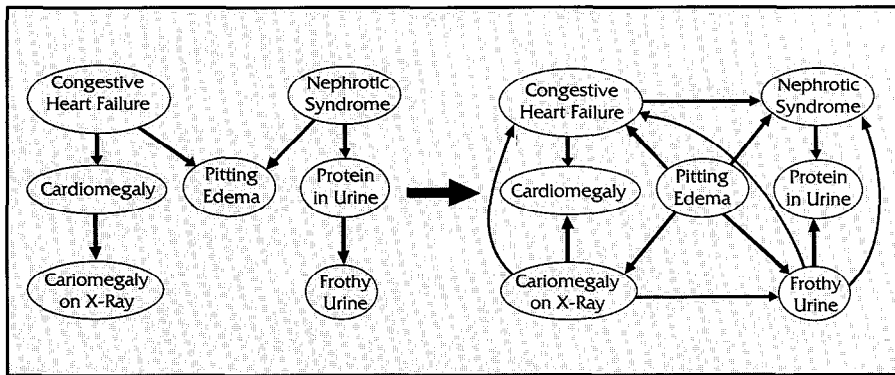


Figure 12

An Influence Diagram for a Medical Diagnosis Problem

Figure 12 illustrates a medical example in which two disorders (causes), congestive heart failure and nephrotic syndrome, are indicated by test results (effects), including a common one, pitting edema. The diagram on the left is fairly simple and one that can be comfortably assessed by the physician. To perform diagnosis, a number of arcs must be reversed to obtain the diagram on the right; this diagram has a distribution for congestive heart failure conditioned by all possible test results. Although the new diagram is useful for diagnosis, it is practically impossible for a physician to assess its conditional distributions.

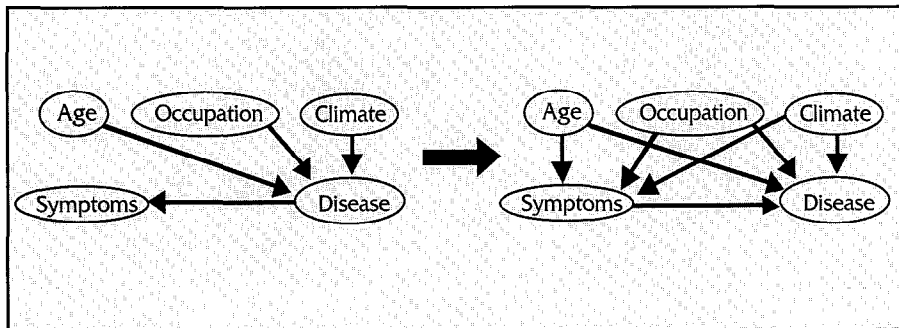


Figure 13

Taking Advantage of Constant Likelihood

In this generic medical example, a number of factors condition the distribution for a disease (cause), but the symptoms (effect) are conditionally independent of those factors given the disease. This diagram is natural to assess and, in fact, corresponds to the organization of most medical texts. To perform a diagnosis, however, the arcs must be reversed to produce the diagram on the right, which is much more difficult to assess. This new diagram corresponds to the physician's clinical experience. If an experienced physician changes jobs, there is often a period of adjustment as the physician learns about the new factors. Similarly, a knowledge base organized like the diagram on the right would be less portable than one based on the diagram on the left.

widely between cases on the basis of patient-specific factors such as age, occupation, and climate; the probability distribution for the symptoms, given the disorder, is often independent of these factors. However, when the model is reversed to the direction of usage, both distributions become patient specific. Thus, a knowledge base organized in the intuitive direction is more portable than one constructed in the direction for use.

Moreover, by building models in the direction of natural assessment, constant likelihoods can be exploited to decrease knowledge acquisition time.

Finally, it is often useful to add new variables that simplify the construction process. Consider the medical example in Figure 12. If we are interested in the probability distribution for pitting edema given congestive heart failure, it is much easier to assess with nephrotic syndrome pre-

sent in the model. We can then take expectation over nephrotic syndrome to arrive at our desired conditional distribution. If we did not explicitly consider nephrotic syndrome, then we would be forced to perform the integration mentally instead. The addition of variables can be of considerable cognitive aid when trying to assess the probability distributions. This process is what Tribus (1969) calls "extending of the conversation." We get little benefit from this technique unless we first build our model in the causal direction.

Conclusions

We believe it is important to distinguish between the direction in which a model is constructed and the direction in which it is applied. Models that capture rich interactions can become impossible to assess in the usage direction even though they might be simple and natural to think about in the opposite direction.

AI researchers have argued that various methods for reasoning with uncertainty are impractical because of the complexity of knowledge acquisition (Shortliffe 1976; Rich 1983). Indeed, many AI researchers have sacrificed self-consistency of the reasoning mechanism in order to facilitate simplicity in the knowledge representation process (Heckerman 1986). We contend that the desired simplicity can often be found by constructing models in the direction opposite that of usage without having to sacrifice fundamentals.

Notes

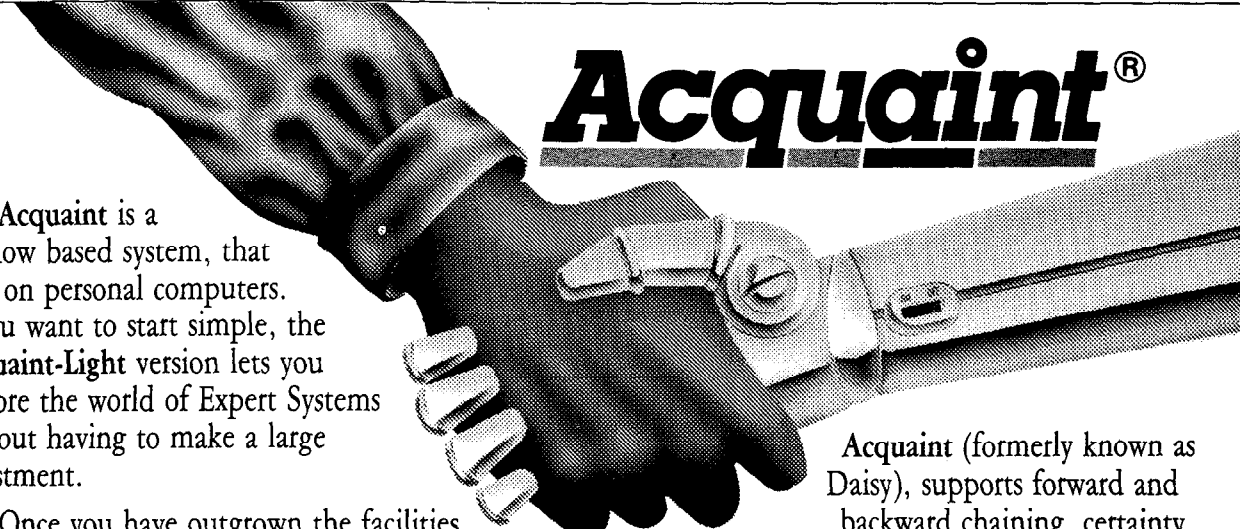
- 1 Many of the concepts discussed in this article were examined previously in Shachter and Heckerman (1986). Also relevant to our discussion is Pearl (1987).
- 2 This is not to be confused with forward versus backward chaining which is an issue of control rather than representation.
- 3 It is interesting to note that the reversal operation is simplified considerably when the predecessor X is a deterministic node. It is not simplified, however, when the successor Y is the deterministic node.
- 4 Notice that the newer model does show some conditional independence which can be exploited at the time of usage.

Acknowledgments

This work was supported in part by Decision Focus, Inc.; the National Library of Medicine under grant R01-LM04529; NASA-Ames Research Center, the Henry J Kaiser Family Foundation; and the Ford Aerospace Corporation. Computing facilities were provided by the SUMEX-AIM resource under NIH grant RR-00785.

References

- Cooper, G. F. 1984. NESTOR: A Computer-Based Medical Diagnostic Aid That Integrates Causal and Probabilistic Knowledge. Ph.D. diss., Technical Report STAN-CS-84-48, HPP-84-48, Dept. of Computer Science, Stanford Univ.
- Heckerman, D. E. 1986. Probabilistic Interpretations for MYCIN's Certainty Factors. In *Uncertainty in Artificial Intelligence*, eds. L. Kanal and J. Lemmer, 167-196. New York: North Holland.
- Howard, R. A., and Matheson, J. E. 1981. Influence Diagrams. In *Readings on the Principles and Applications of Decision Analysis*, eds. R. A. Howard and J. E. Matheson, 721-762. Menlo Park, Calif.: Strategic Decisions Group.
- Kuipers, B., and Kassier, J. 1984. Causal Reasoning in Medicine: Analysis of a Protocol. *Cognitive Science*, 8:363-385.
- Olmsted, S. M. 1983. On Representing and Solving Decision Problems. Ph.D. diss., Dept. of Economics and Engineering Systems, Stanford Univ.
- Pearl, J. 1987. Embracing Causality in Formal Reasoning. Technical Report, CSD 860020, Dept. of Computer Science, Univ. of California at Los Angeles. Also in *Proceedings of the Sixth National Conference on Artificial Intelligence*, 369-373. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Pearl, J. 1986. Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence* 29:241-288.
- Rich, E. 1983. *Artificial Intelligence*. New York: McGraw-Hill.
- Shachter, R. D. 1986. Intelligent Probabilistic Inference. In *Uncertainty in Artificial Intelligence*, eds. L. Kanal and J. Lemmer, 371-382. New York: North Holland.
- Shachter, R. D., and Heckerman, D. E. 1986. A Backwards View for Assessment. In *Proceedings of the Second Workshop on Uncertainty in Artificial Intelligence*, 237-241. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Shortliffe, E. H. 1976. *Computer-Based Medical Consultations. MYCIN*. New York: Elsevier-North Holland.
- Shortliffe, E. H., and Buchanan, B. G. 1975. A Model of Inexact Reasoning in Medicine. *Mathematical Biosciences* 23:351-379.
- Speed, T. P. 1978. Graphical Methods in the Analysis of Data. Paper presented at the University of Copenhagen Institute of Mathematical Statistics, Copenhagen.
- Spiegelhalter, D. J. 1985. A Statistical View of Uncertainty in Expert Systems. Paper presented at the Workshop on AI and Statistics, Bell Laboratories, Princeton, N.J., 11-12 April.
- Tribus, M. 1969. Bayes' Equation and Rational Inference. In *Rational Descriptions, Decisions, and Designs*. New York: Pergamon.
- Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research* 20:557-585.



Acquaint is a window based system, that runs on personal computers. If you want to start simple, the Acquaint-Light version lets you explore the world of Expert Systems without having to make a large investment.

Once you have outgrown the facilities of Acquaint-Light, you may upgrade to the full version, which is intended to be used by knowledge engineers. Because it is written entirely in muLisp, you can get access to most of the systems internal functions and you may add your own functions as well.

The Expert System Development Tool that grows with you.

Acquaint (formerly known as Daisy), supports forward and backward chaining, certainty factors and fuzzy comparisons.

It has a powerful forms facility, that your users will feel at home with.

For more information, please write to:

LITHP SYSTEMS BV

P.O. BOX 553, 1140 AN PURMEREND, THE NETHERLANDS

For free information, circle no. 77