On Irrelevance and Partial Assignments to Belief Networks

Technical Report CS-90-14

Solomon E. Shimony

October 1990

On Irrelevance and Partial Assignments to Belief Networks*

Solomon E. Shimony

Computer Science Department
Box 1910, Brown University
Providence, RI 02906

ses@cs.brown.edu

October 16, 1990

Abstract

We evaluate current methods for finding best consistent explanations. The only known general probabilistic scheme for finding a consistent explanation, that of complete MAPs, suffers from the overspecification problem. We propose a partial solution by applying a probabilistic generalization to logical irrelevance. We evaluate cost-based abduction in a probabilistic context by providing a probabilistic semantics for it in the case of partial assignments.

*The author is supported by a Corinna Borden Keen Fellowship. I wish to thank Eugene Charniak for many helpful suggestion, and for critical reading of multiple versions of this paper.

1 Introduction

When attempting to find a best consistent explanation for some evidence in probabilistic reasoning, one usually looks for most probable explanations (see description of MPEs, in [7]), which are complete models of maximum a-posteriori probability (MAP). With this method, the best explanation for evidence $\mathcal{E}$ is the assignment $\mathcal{A}$ to all the variables that maxizes $P(\mathcal{A}|\mathcal{E})$. This method seems reasonable, and a recent paper of ours ([1]) showed that even some rule based systems with costs, which seem non-probabilistic in nature, essentially find the most-probable explanations, assuming a certain mapping between costs and prior probabilities.

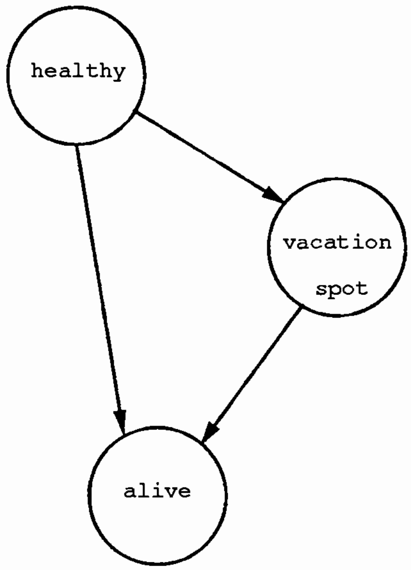
Pearl shows how to compute a most-probable complete assignment for a belief network (MPE), and we proposed another method in [9]. We suspect, however, that complete assignments are not, in general, the way to go. One clear indication of this is the following problem, taken from chapter 5 of Pearl's book (we call it the overspecification problem).

Suppose that I am planning to take some time off on vacation, and take some medical tests beforehand. Now, some medical test gives me a 0.2 probability of having cancer[1] (i.e. probability 0.8 of being healthy). Now, if I am alive and go on vacation whenever I am healthy, then the most-probable state is of me being healthy, alive, and on vacation (ignoring temporal problems). Suppose, however, that now I plan my vacation, and am considering 10 different vacation spots (including one of not going anywhere), and make them all equally likely, given that I'm healthy (i.e. no preference). Also, assume that I will stay at home if I'm not well, and that I will not live if I do indeed have cancer. One way of representing the data is in the form of a belief network, as shown in figure 1. The network has three nodes: alive, healthy, and vacation spot[2].

In this network, however, since we have 10 vacation spots, the probability

[1] We assume here that cancer= $\neg$*healthy*, i.e. we assume that there are no other diseases for the purposes of this example

[2] The terms *nodes* and *variables* are used interchangeably throughout. We use lower case names and letters for variable names, and their respective capitalized words and letters to refer to either a particular state of the variable, or as a shorthand for a particular assignment event to the variable. In addition, a variable name appearing in an equation without explicit assignment, means that we actually have a set of equations, one for each possible assignment to the variable. E.g. $P(x) = P(y)$ where x and y are boolean valued nodes stands for $P(x = T) = P(y = T) \ \wedge \ P(x = F) = P(y = T) \ \wedge \ P(x = T) = P(y = F) \ \wedge \ P(x = F) = P(y = F)$.



P(Healthy) = 0.8

P(Alive|Healthy, vacation-spot) = 1

P(Alive|~Healthy, Stay Home) = 0

P(Alive|~Healthy, Other Location) = 0.1

P(vacation-spot=Stay Home|~Healthy) = 1

P(vacation-spot|Healthy) = 0.1

Figure 1: Example Belief Network

of any scenario where I am alive is 0.08, and the scenario where I die of cancer is the most likely (probability of 0.2)! This is an undesirable property of most-probable explanations, because it is not reasonable to expect to die just because of planning ahead one step.

Pearl proposes "circumscribing explanations", in two ways. First, he suggests not deciding variables with no evidence coming from below (we will call this the "evidential support" criterion). Evidential support seems like a good idea. If we accept that causality flows in the direction of the arrows in the belief network, and explanations are in terms of things causally prior, then an explanation of evidence need not include anything lower in the network. Thus, we adopt this view. A second method is the "principle of disparate risks", from [4]. This latter method, however, makes some independence or correlation assumptions that we cannot accept, which are equivalent to assuming that "multiple risks cancel out".

Even the use of the evidential support is insufficient, however, because in the case of figure 1, there is evidence coming from the alive node[3], and that

[3]We allow evidence to be of the form Alive$\vee\neg$Alive, which does not limit the a-posteriori sample-space, and is thus "null" evidence, in order to express "interest" in the variable without introducing actual evidence (i.e. evidence that modifies the a-posteriori sample-space).

what nodes are allowed to be evidence (even though we do not rule out that possibility). We propose selecting the partial assignment with the highest probability, provided that unassigned variables are irrelevant to any of the assigned variables. Initially, we use statistical independence as a measure of irrelevance. We show that this is insufficient for our purposes, because of the instability of the method relative to any small change in conditional probability distribution.

We also discuss cost-based abduction, our variant of Hobbs and Stickel's least cost proofs. We extend our previous results on probabilistic semantics for cost-based abduction to include partial models. Finally, we show that even though these least cost proof based schemes seem to solve Pearl's problem, they give wrong results for other simple problems.

2 Irrelevance and Independence

Our definition of irrelevance is closely related to statistical independence. We will define probabilistic irrelevance relative to sets of models, where we limit ourselves to the sets of models induced by partial assignments.

Definition 1 *A partial assignment is an assignment of values to variables, where some variables may remain unassigned.*

Thus, a partial assignment is synonymous with a set of models. For example, for the set of variables $\{x, y, z\}$ each with a binary domain, the assignment $\{x = T, y = F\}$ with z unassigned is equivalent to the set of models $\{(x = T, y = F, z = F),\ (x = T, y = F, z = T)\}$. As a notational convenience, we will augment our set of domain values by a designated special value, U (for "unassigned"). Note that this is shorthand for saying that the variable can have any value. Thus for any variable x, we have $P(x = U) = 1$, and variables assigned U can always be ignored in a probability term. We will also allow assignments of U to appear for evidence nodes. That signifies "interest" in the node, without actual evidence being introduced. Thus, in our example from the introduction, the "evidence" is "Alive $= U$".

We say that $In(f, g, \mathcal{A})$ if g is independent of f given $\mathcal{A}$ (where $\mathcal{A}$ is a partial assignment)[4], i.e. if $P(g|f, \mathcal{A}) = P(g|\mathcal{A})$. We allow f and g to be either sets of variables or partial assignments to sets of variables. This

[4]We use the notation $\mathcal{A}^S$ to denote an assignment to the set of variables S whenever the use of $\mathcal{A}$ is ambiguous.

would cause us to assign a value to the vacation spot node, whereas we would rather leave it unassigned. That is because our intuition would suggest that it is *irrelevant* to the "alive" node in the context of being healthy.

We see that not assigning some nodes may provide a solution to the problem. Another way to state that is to say that we are looking for best *partial* models. The question is what criterion to use for leaving nodes unassigned in the models. Pearl's suggestion of not assigning nodes below the evidence is one such criterion. Other researchers essentially divide nodes into those that may be assigned (the primary causes), evidence nodes, and other nodes. Two such schemes are reviewed here.

Cooper, in his PhD thesis (see [2], or [6]), finds most probable sets of diseases for a given set of symptoms. He performs a best-first search for a most probable set of "diseases", or causes, given the evidence. His models are partial in the sense that non-root nodes that are not in the evidence set are never assigned. His work, however, does not handle the general case, as he assumes mutual independence of all causes (i.e. they all have to be root nodes). Because of that, spontaneous occurrences of non-root nodes cannot be explanations of the evidence. This means that non-root nodes can never be causes. The latter is a deficiency of the theory, if it is to be used for general case explanations. For example, look at the case where we are trying to understand a story. We have the taking-a-bus action, which can explain a sentence, and the going-to(x) actions (with x all the possible locations one may take a bus to), which are possible causes for taking-a-bus. If we want the going-to(x) to be in our system at all, then we may not have taking-a-bus as a cause. That, in turn, means that we will always have to make up our mind about the location, even when it is undesirable (such as when there are thousands of locations, with no evidence for preferring any of them over the others).

Peng and Reggia, in [8], have defined a diagnostic problem which they solve using a cover set method. The probabilistic extension of that uses a 2-level belief network. They also designed a best-first algorithm that finds hypotheses in decreasing order of probability. It is not clear, however, how their methods would extend to a general belief network, given that one of their assumption is that all symptoms have causes (thus root nodes cannot be evidence).

In this paper, we present a solution that also relies on partial assignments, but is more general than the previously proposed schemes. Our solution does not depend on a prior decision of what nodes are allowed to be causes and is similar to Pearl's independence notation $I(a, b, c)$, where variable set a is independent of variable set c given variable set b. The difference is that Pearl's notation does not require a certain assignment to b, just that the assignment be known; whereas our notation does require it. $I(f, b, g)$ implies $In(f, g, \mathcal{A}^b)$ and $In(g, f, \mathcal{A}^b)$, but *not* vice-versa.

Our notion of probabilistic irrelevance is syntactically related to the definition of strong irrelevance in [10]. In that paper, $SI(f, g, M)$ is used to signify that f is irrelevant to g in theory M if f is not necessary to prove g in M, and vice versa (see [3] for the precise definition). However, the semantics of the definitions where not directly applicable, because we are interested in irrelevance of f to g even if f and g are not true.

We now define our notion of best explanation for belief networks (which we call Irrelevance-based Partial MAP, acronym IPMAP) as the most probable irrelevance-based partial assignment.

Definition 2 *An irrelevance-based partial assignment $\mathcal{A}$ is a partial assignment where $In(\mathcal{A}^B, \mathcal{U}, \mathcal{A}^T)$, where $\mathcal{U}$ is the set of unassigned variables strictly above the evidence*[5]*, B is the set of variables in $\mathcal{A}$ that have at least one parent in $\mathcal{U}$, and $T = ($variables in $\mathcal{A}$ $) - B$.*

Definition 3 *An assignment $\mathcal{A}$ is an* irrelevance based best partial MAP *if it is an irrelevance-based partial assignment with $P(\mathcal{E}|\mathcal{A}) = 1$,*[6] *and for any other irrelevance-based partial assignment $\mathcal{A}'$ such that $P(\mathcal{E}|\mathcal{A}') = 1$, we have $P(\mathcal{A}) \geq P(\mathcal{A}')$.*

Note that the definition relies on the directionality of belief networks, the "cause and effect" directionality. Thus, $\mathcal{A}^T$ are the causes, and we do not assign (i.e. are not interested in) variables that are irrelevant to the evidence given the causes. We also do not assign nodes that are below the evidence, as per Pearl's suggestion.

Irrelevance-based partial MAP solves our example problem. In the example, (with evidence "Alive=U"), we have an irrelevance-based MAP of (Alive, Healthy, vacation spot undetermined) with a probability of 0.8 as desired. We can avoid assigning a value to vacation location because we have: $\mathcal{A}^T$={Healthy}, $\mathcal{A}^B$={Alive}, $\mathcal{U}$ ={vacation spot}, and P(alive | Healthy) = P(alive | Healthy, vacation spot), i.e. it is independent of the location.

[5]A node v is strictly above the evidence iff v is not evidence and there exists a directed path from v to an evidence node.

[6]This means that either the evidence is implied by $\mathcal{A}$, or the evidence is actually in $\mathcal{A}$.

3 Cost-Based Abduction Semantics

An alternate way of finding a criterion for deciding which variables should remain unassigned in partial models, is to start with a working non-probabilistic system for explanation, find a probabilistic semantics for it and evaluate the resulting model. We went part of the way in that direction in a prior paper ("Probabilistic Semantics for Cost-Based Abduction", [1]), where we started off with Hobbs and Stickel least-cost proof [5], modified that into cost-based abduction, and then proceeded to provide probabilistic semantics for the latter. Our paper provided precise semantics only for finding *complete* models. It showed that finding such a least-cost explanation is equivalent to finding the MAP for the induced belief network. The issue of semantics for *partial* models remained open, however. As we will show presently, the partial model version of cost-based abduction also uses a form of independence to justify non-assignment of certain nodes. Thus, the use of cost-based abduction for solving the overspecification problem appears justified.

We now provide a semantics for the partial model cost-based abduction, for the case where there are no negated literals in rules or in the evidence.

The cost based abduction system has rules of the form:

$$R: \quad (p_1 \wedge p_2 \wedge \ \dots \ \wedge p_k)^{c_R} \ \rightarrow q$$

where q can be proved true with this rule for a cost of c_R if all of p_i are true. If some of the p_i are not known to be true (or false), they may be *assumed true* for a cost c_i. In our semantics, this c_i is negative log probability of p_i being true in the world, *only* if p_i does not appear on the right hand side of any rule. If it does, then the cost is negative log probability of p_i being true given that all its antecedents are false. This works correctly for *complete* assignments, i.e. the minimal cost complete assignment to all literals will in fact be an MAP when the costs are negative log probabilities as above.

In the semantics for *partial* models, we suggest the same probabilities as for the complete model case, except for the case where p_i appears both as an antecedent and as a consequent (of different rules). In the latter case, we just use the relevant prior probability, instead of the conditional probability[7]. If the rules have no negation (i.e. the resulting graph is pure AND/OR), and all the evidence is in the form of positive literals, then the cost of assuming

[7]For example, if the only rule that has q on the right-hand side is $p \rightarrow q$, and there is a rule $q \rightarrow r$, the cost of assuming q will be $-log(P(q))$, not $-log(P(q|\neg p))$ as in the complete model case.

an OR node true is exactly minus log probability of the prior probability of the node. This is true because in our construction, the only internal nodes which may be assumed true are OR nodes. The resulting best partial model will not have any nodes assigned false, since such an assignment would not contribute to an explanation (it may have many unassigned nodes).

In order to represent Pearl's overspecification problem example and other interesting explanation problems, however, we need to generalize cost-based abduction so as to allow negation and multi-valued variables. Instead of actually generalizing cost-based abduction, we will just note the kind of independence assumptions that cost-based abduction makes, and let them carry over to general belief networks.

Note that in the case of AND/OR networks, Cost-based abduction behaves as follows. Nodes that are assigned U values are treated as if not in the graph, i.e. sets of nodes N_1 and N_2 which are not connected after removing nodes assigned U are assumed (perhaps incorrectly) independent, and the probability of the assignment will contain the product of their distribution, $P(N_1)\ P(N_2)$, instead of $P(N_1, N_2)$. Thus, when evaluating the cost (or probability) of an assignment, we essentially compute the product:

$$\prod_{x \in \mathcal{B}} P(\mathcal{A}^x | \mathcal{A}^{parents(x)})$$

where $\mathcal{B}$ is the set of all the nodes in the belief network, and the assignments may be partial assignments. Evaluating these partial conditional probabilities in a belief network may be non-trivial (exponential time), but we are not concerned with efficiency issues at the moment.

Note that in the limiting case of complete assignments we are making no unwarranted independence assumptions, because the above product reduces to exactly the joint probability of the network. It is this property of complete assignments that allowed us to prove the equivalence of complete-model cost-based abduction to finding most probable a-posteriori models. Also, evaluating the terms of the product becomes trivial in the case of complete models.

Partial-model cost-based abduction will do fine on our example[8], but the unwarranted independence assumptions will cause problems, such as selecting clearly incorrect answers in certain cases, as we will show in the discussion.

[8]It will prefer {Healthy, Alive} with Vacation Spot unassigned to {¬Healthy, ¬Alive, vacation spot = stay at home}, but will also prefer not to assign Healthy at all!

4 Implementation Issues

Our best-first search algorithm for computing MAPs, as described in [9], has been modified to compute best explanation where variables are assigned U if the current assignment is independent of them. In order to explain the modification of the algorithm, we will review its operation here. The original algorithm is a variant of A^*, and uses an agenda of states sorted by current probability, which is a product of all conditional probabilities seen in the current expansion. The operation of the algorithm is shown in the following table.

1	insert evidence as a state into the agenda
2	de-queue agenda into current
3	if current state is not complete go to step 5
4	if need only one result, halt
5	extend current state, and go to step 2

The modifications are in checking for completeness in step 3, and in extending the current agenda item. In both cases, the extension consists of finding next states, evaluating them and inserting them into the agenda. The difference is that with the modified algorithm, when extending a node some of the parents may not be assigned, as we will show later. Completeness in the modified algorithm is different in that an agenda item may be complete even if not all variables are assigned. Also, many complete assignment states are subsumed by some single states with unassigned variables.

We make the following observation: if, for each assigned variable v, v is independent of all of its unassigned parents (given the current assignment of the rest of its parents), and all parents of unassigned nodes are either unassigned nodes or top nodes (recall that top nodes are assigned nodes that are not evidence and do not have unassigned parents); then the entire assignment is independent of any subset of unassigned nodes (given the top nodes).

To take advantage of the above observation, we precompute for each node a set of all the cases where such independence occurs. Then, when the algorithm expands that node, it only assigns the parents that *have* to be assigned. Naturally, since a belief net is not always a tree, some nodes may already be assigned, and we take care of this eventuality.

For example, consider the belief network segment of figure 2, where we are at a noisy OR node v, with parents u_1, u_2, u_3, u_4, where v has the value

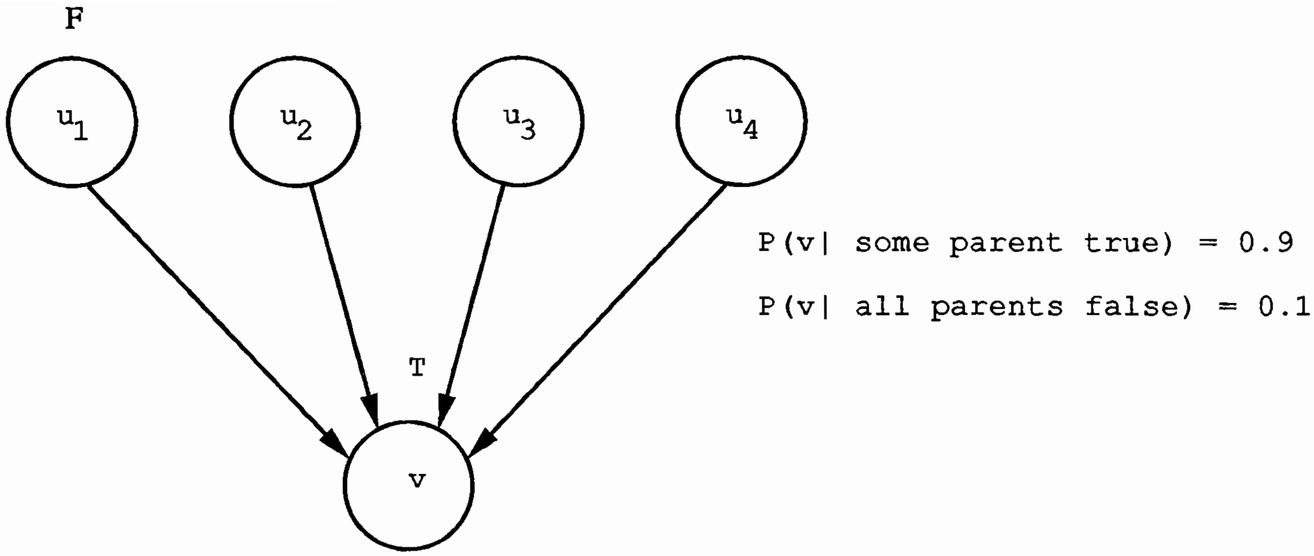


Figure 2: Expanding a Node

T, and u_1 has already been assigned F. We now have to expand all the "interesting" states of the parents of v, i.e. the states of the u_i.

We see that we have only 4 interesting assignments[9], instead of 8 (all possible complete assignments), for the remaining 3 unassigned nodes, (u_2, u_3, u_4):

$$\{(T,U,U),\ (U,T,U,),\ (U,U,T),\ (F,F,F)\}$$

all other assignments are not used, because they would assign values to variables that cannot change the probability of the assignment, that is, they are subsumed by the 4 assignments listed above.

5 Discussion

As far as the partial MAP algorithm is concerned, if the graph has many nodes with conditional independence (for example: AND and OR nodes, both pure and noisy, switching nodes...) the search space is considerably

[9] An "interesting" assignment is one not subsumed by some other assignment. For example, $A_1 = \{T, U, U\}$ subsumes $A_2 = \{T, F, U\}$ because A_1 contains all the models of A_2 and $P(v = T|u_2 = T) = P(v = T|u_2 = T, u_3 = F)$.

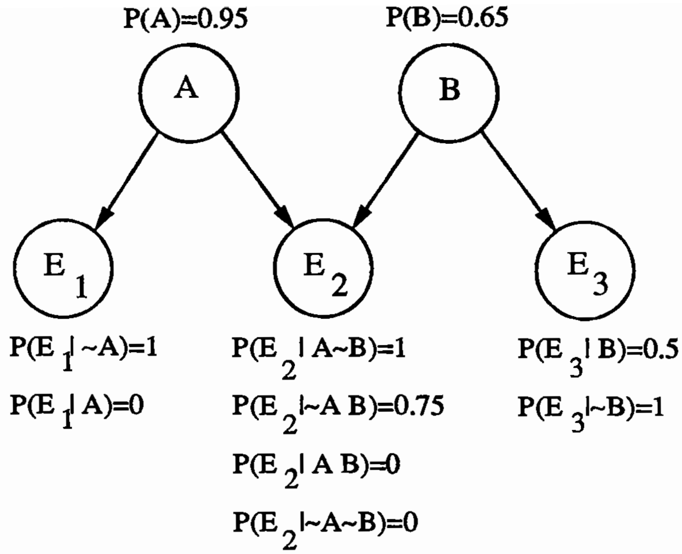


Figure 3: Counterexample for Cost-Based Abduction

reduced, and the algorithm can run much faster than the original MAP computation algorithm.

However, the entire scheme of conditional independence is not sufficiently powerful. We still have a problem of instability, as we show by the following example. Suppose that we change the probability of being alive given the location such that staying at home (when healthy) still gives a probability of 1 of being alive, but going anywhere else only gives a probability of 0.99 of being alive (say an accident is possible during travel). We no longer have independence, and thus are forced into the bad case of finding the "not alive" scenario as the best explanation.

Partial-model cost-based abduction seems to solve the instability problem, because it gives reasonable answers in the example case, and its modification. However, the unwarranted independence assumptions will cause problems, as we may be assuming independence where actually *negative* correlation exists[10]. As a counterexample, consider the binary valued belief network in figure 3, with nodes A, B, E_1, E_2, E_3, and where the evidence is that all the E_i are true.

Clearly, partial model cost-based abduction will prefer the wrong choice

[10]This problem does not arise if we only have *positive* correlations, i.e. the network is essentially an AND/OR DAG).

here ($A = U,\ B = F$) over the correct choices ($A = F,\ B = T$ or $A = F,\ B = U$). The assignment $A = U,\ B = F$ is wrong because it contains no models consistent with the evidence, i.e. both models $A = F,\ B = F$ and $A = T,\ B = F$ have probability 0 given the evidence.

6 Conclusion

We followed Pearl's example showing that MAP calculation is not necessarily the best way to find explanations; and argued that partial assignments may provide better results. We proposed some explanation schemes that give meaning to partial MAPs. Our first scheme, where independence is the deciding factor on whether a variable has to be assigned a value, solves some of the problems of MAPs, but further research, possibly on defining a weaker notion than independence, is necessary to solve the instability problem.

Extending least-cost proofs and related methods make some of the problems go away (the instability problem, for example), but they are still lacking, mainly because they seem to make unwarranted independence assumptions.

We have implemented a comparatively efficient algorithm for computing best explanation with independence, that is a simple modification of our algorithm for finding MAPs.

Finally, we believe that viewing explanations from a decision-theoretic point of view will be the correct way to eliminate the remaining problems (rather than the strict probabilistic view, which is a degenerate case of the decision-theoretic view).

References

[1] Eugene Charniak and Solomon E. Shimony. Probabilistic semantics for rule based systems. Technical Report CS-90-02, Computer Science Department, Brown University, February 1990.

[2] Gregory Floyd Cooper. *NESTOR: A Acomputer-Based Medical Diagnosis Aid that Integrates Causal and Probabilistic Knowledge.* PhD thesis, Stanford University, 1984.

[3] Michael R. Genesereth. The use of design descriptions in automated diagnosis. *Artificial Intelligence*, pages 411–436, 1984.

[4] J. C. Harsanyi. Acceptance of empirical statements: A bayesian theory without cognitive utilities. *Theory and Decision*, 18:1–30, 1985.

[5] Jerry R. Hobbs and Mark Stickel et. al. Interpretation as abduction. In *Proceedings of the 26th Conference of the ACL*, 1988.

[6] Richard E. Neapolitan. *Probabilistic Reasoning in Expert Systems*, chapter 8. John Wiley and Sons, 1990.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[8] Y. Peng and J. A. Reggia. A probabilistic causal model for diagnostic problem solving (parts i and ii). In *IEEE Transactions on Systems, Man and Cybernetics*, pages 146–162 and 395–406, 1987.

[9] Solomon E. Shimony and Eugene Charniak. A new algorithm for finding map assignments to belief networks. In *Proceedings Conference on Uncertainty in AI, 1990*, 1990.

[10] Devika Subramanian. *A Theory of Justified Reformulations.* PhD thesis, Stanford University, 1989. Technical report STAN-CS-89-1260.