The Knowledge Engineering Review, Vol. 4: 2, 1989, 97–117

Applications of nonmonotonic logic to diagnosis

PETER JACKSON*

McDonnell Douglas Research Laboratories, Dept 225, Bldg 105/2, Mail Code 1065165, PO Box 516, St Louis, MO 63166, USA

Abstract

This paper attempts to assess the practical utility of nonmonotonic logic in diagnostic problem solving. We begin with a brief review of the main assumptions which motivate work in this area, and discuss two logic-based approaches which involve nonmonotonic arguments. Then we consider two recent proposals for the application of default logic to diagnosis, as well as a proposal based on counterfactual logic. In conclusion, we briefly compare these methods with other diagnostic reasoning paradigms found in the Artificial Intelligence literature.

1 Introduction

Classical logics, such as the propositional and predicate calculi, are *monotonic*, in the sense that if a given set of sentences S sanctions a set of inferences $\mathrm{Th}(S)$, then if $T \supseteq S$ then $\mathrm{Th}(T) \supseteq \mathrm{Th}(S)$. In plain English, if T extends S in some way, by containing more information, then the inferences we can draw from T contain the inferences we can draw from S. In classical logic, we never have to reexamine the consequence (or theorems) of a theory when we extend it; we just accumulate more theorems.

The term *nonmonotonic* is used to describe reasoning systems in which the set of theorems does not always increase with theory extension. This characteristic arises because some of the premises that we employ in our reasoning are assumptions which may turn out to be false. Such assumptions are not the solid premises of classical arguments, which are either axioms (true by definition) or deductive consequences of the axioms; we are led to hold them by the *absence* of information to the contrary, rather than by the presence of information which establishes them as true. Hence the notion of *consistency* is as important in nonmonotonic reasoning as the notion of logical consequence, and where assumptions lead to inconsistency they must be retracted.

There are a number of reviews and critiques of the theoretical literature on nonmonotonic logic which can be recommended to the reader (e.g., Reiter, 1987a; papers in Ginsberg, 1987; McDermott, 1987 and commentaries). In this paper, we focus upon proposals for the practical application of nonmonotonic logic to the domain of diagnosis. The main emphasis will be upon reasoning systems which have some formal basis in either *default logic* or *counterfactual logic.*

The plan of the paper is as follows. Section 2 briefly reviews Reiter's (1980) default logic, and shows how defaults may figure in diagnostic reasoning. Sections 3 and 4 compare Reiter's (1987b) work on diagnosis from first principles with Poole's (1988a,b) work on constructing explanations. Each of these proposals can be given a formal basis in default logic; their use of defaults is compared and contrasted. Section 5 considers an account of diagnosis based upon counterfactual logic (Ginsberg, 1986); its relationship to Reiter's work is explained. Finally, section 6 attempts a critical evaluation of these nonmonotonic methods in the context of other diagnostic reasoning paradigms found in the Artificial Intelligence literature.

The rest of this introduction contains: (i) an analysis of the main assumptions underlying the application of nonmonotonic reasoning to diagnosis, and (ii) a review of earlier logic-based approaches involving nonmonotonic arguments.

* This work was supported by the McDonnell Douglas Independent Research and Development program.

1.2 Basic assumptions of the approach

Many extant expert systems perform diagnosis by the method of *heuristic classification*, as described by Clancey (1985). In this method, diagnostic knowledge is represented mainly in terms of heuristic rules, which perform a mapping between data abstractions (typical symptoms) and solution abstractions (typical disorders). Such a representation of knowledge is sometimes called "shallow" because it does not contain much information about the causal mechanisms underlying the relationship between symptoms and disorders. The rules typically reflect empirical associations derived from experience, rather than a theory of how the device under diagnosis actually works. The latter is sometimes called "deep" knowledge, because it involves understanding the structure of the device and the way its components function.

It is not that deep knowledge is entirely absent from rule-based expert systems. It is rather that such knowledge is either compiled into the rules in a manner that renders it difficult to differentiate from the shallower stuff, or else hidden in other data structures which serve to control rule-based inference. In any event, it is the shallow knowledge that determines the underlying search space, since it is this knowledge that characterizes the mapping between data and solutions. Empirical knowledge of this sort is largely device-dependent, and tends to reside in someone's head, as opposed to being written down or (even less likely) formalized. The only way to gain such knowledge is via the repeated experience of attempting to perform diagnosis on the device in question and then determining the correctness or otherwise of one's hypothesis, or via an extended knowledge elicitation exercise.

Applications of nonmonotonic reasoning to diagnosis tend to approach the problem from another angle. Rather than assume the existence of an expert experienced in diagnosing the device in question, these applications assume the existence of either:

1 a complete *structural description* of the device, detailing its components and their connections; or
2 a *causal model* of the device, identifying the consequences of component (or connection) failure.

Approaches associated with assumption 1 reason "from first principles" in the sense that they use little or no experiential knowledge in generating hypotheses which account for data. The device description provides all the information that is required for the conjecture of component faults which would produce the observed symptoms. The apparent advantages of first principles over expert systems approaches are as follows:

- Given a device description, the program designer is able to short-cut the laborious process of eliciting empirical associations from a human expert.
- The reasoning method employed is device-independent, so it is not necessary to tailor the inference machinery for different applications.
- Since only knowledge of *correct* device behaviour is required, the method ought to be able to diagnose faults which have never occurred before.

Approaches associated with assumption 2 use a representation of knowledge that maps disorders to symptoms, rather than vice versa (as in heuristic classification). The advantage over expert system approaches is that eliciting the necessary information from a human expert is much easier than uncovering the inverse mapping from symptoms to disorders; such information can also be obtained (or verified) by simulation techniques. The advantage over reasoning from first principles is that a complete description may not be available for the device in question, in which case a causal model will have to do.

We shall outline additional assumptions relevant to the application of particular approaches as we describe and discuss the corresponding systems. But first we consider two systems which, although they are not based on any particular nonmonotonic logic, use nonmonotonic arguments in generating hypotheses to account for symptoms. We describe them here because the proposals of Reiter and Ginsberg can be seen as an attempt to "rationally reconstruct" such systems in a nonmonotonic logic.

1.b Some background: DART and GDE

The DART program (Genesereth, 1984) was a prototype for a diagnostic system in the domain of digital circuits. A distinguishing feature of this system was that both the representation language and the inference mechanism were more or less device independent. Genesereth used the predicate calculus to encode design descriptions of devices under diagnosis, and a form of theorem proving called *resolution residue* to generate both sets of suspect components and tests to confirm or disconfirm fault hypotheses.

Although DART's inference method was not based on a particular nonmonotonic logic, it incorporated a form of nonmonotonic reasoning within a monotonic logistic system. A hypothetical fault was represented by a proposition which was the negation of a proposition in the design description. Such hypotheses were only allowed to stand as long as they were consistent with the current state of knowledge; thus further information might lead to their retraction.

The basic flowchart for the DART program is reproduced in Figure 1 from Genesereth's paper. Although we need not be concerned with the details of DART's implementation, the top-level algorithm is typical of many other diagnostic programs. Suspected faults are generated by theorem proving, and then tests are generated to narrow the space of alternatives. Tests embody knowledge of how components work, e.g., functions which compute the outputs of a component given its inputs. More recent proposals for logic-based diagnosis differ in that they are more firmly based in the context of a particular nonmonotonic logic.

DART's problem solving method relied on a number of simplifying assumptions:

(1) The connections in the device are assumed to be working properly, so we seek faulty *components* to account for symptom data.
(2) Faults are *non-intermittent*, i.e., components behave consistently for the duration of the diagnosis.
(3) There is only a *single fault* in the device.

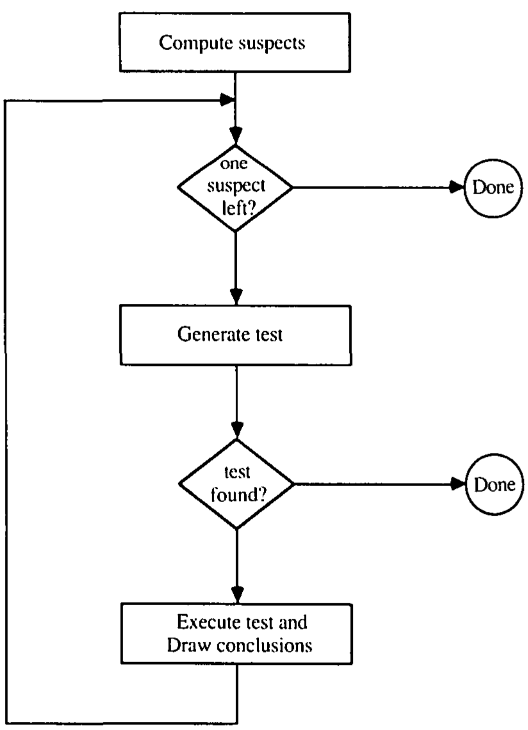

Figure 1 Flowchart for the DART program

Singh (1987) has since extended DART, by improving the test generation procedure and making the resolution theorem prover more efficient, but has retained the original assumptions.

GDE[1] (de Kleer and Williams, 1987) is a system which retains assumptions (1) and (2) but concentrates upon relaxing (3). The trouble is that the hypothesis space now grows exponentially with the number of faults, since we now have to consider sets of hypothetical faults. It is therefore important to generate hypotheses in rough order of "goodness" and then generate tests that will differentiate between competing hypothesis sets in a minimum number of measurements. The inherent complexity of the problem is addressed by a combination of assumption based truth maintenance (ATMS) and probabilistic inference.

Briefly, ATMS (de Kleer, 1986) is a method of maintaining a lattice of *environments*: alternative states of the world which are consistent with some theory but in which different *assumptions* hold. In diagnostic applications, this (implicit) theory is the device description, and the alternative states of the world are characterized by the 2^n different ways in which n components can fail. The environments are arranged in a lattice by set inclusion: the least element is the empty set of faulty components, while the greatest element is the set of all components. Each such environment is called a *candidate*: a hypothesis about the way in which a faulty device actually deviates from its desciption. The environment lattice is called the *candidate space.*

If the device is working perfectly, then the least element (representing no faulty components) "explains" the observation of the correct outputs. Otherwise, faulty outputs are seen as evidence for some component failure in the candidate space. Since the size of this space is exponential in the number of components, we need some way of focusing the search for candidates.

The key notion is that a *conflict set*: a set of components that cannot all be functioning correctly, given the symptom data. The conflict sets are just those assumption sets in the ATMS which generate environments inconsistent with the data. They can be determined by forming the deductive closure of an environment with the data, and then looking for a contradiction. (This is obviously a form of constraint propagation.) The main goal of the truth maintenance system is to identify the complete set of *minimal conflicts*, i.e., to detect those conflict sets which subsume all others by set inclusion. From these, one can compute the minimal sets of faulty components that account for all the symptoms.

As a simple example, let C1, C2 and C3 be system components, and let s be an observation that is a symptom of system failure (see Figure 2). Suppose we know that if s occurs then C1 and C2 cannot be working, and neither can C1 and C3. Thus we can rule out the environments labelled "Ø", "C2" and "C3" (shaded in Figure 2), because each has some forbidden conjunction of components working. All other environments are candidate diagnoses. However, the minimal candidates are obviously {C1} and {C2, C3}, shown in heavy outline.

The ATMS uses a number of strategies to control the complexity of the candidate space, e.g.,

- exploiting a preference for minimal explanations by beginning the search for solutions from the smallest conflict sets;
- recording inferences (such as predictions of device behaviour based on assumption sets) in the ATMS, so that they are never drawn twice.

GDE uses an information-theoretic measure of uncertainty to determine the best next measurement to make. The best measurement is that which minimizes *entropy*, i.e., which causes the greatest diversity among the probabilities of the candidates. They assume that the prior probability of failure is known for each component and that components fail independently. These probabilities are also used to guide the generation of candidates. The search of the environment lattice is conducted best-first on the joint probabilities of multiple faults.

Both DART and GDE are implemented systems that have exerted an influence over more recent proposals. However, each of these programs relies on a particular inference method for the generation of diagnoses: resolution residue for DART and a kind of constraint propagation in GDE.

[1] GDE stands for General Diagnostic Engine.

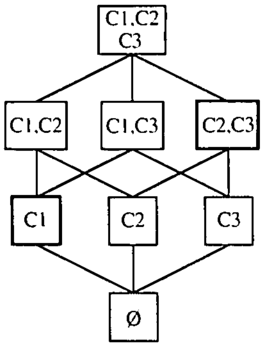

Figure 2 An environment lattice of candidates (see text)

It is arguable that later treatments (such as Reiter's) are more general in that they require nothing more than a general-purpose theorem prover, at least as far as the formulation of the method is concerned. Implementation is typically another story; the useful notion of a minimal conflict reappears in the work of both Reiter and Ginsberg. The next section considers the question of minimality in more detail, and illustrates the connection between diagnosis and default logic.

2 Default logic and minimizing abnormality

Commonsense reasoning often proceeds by making assumptions. Many of these assumptions are "conservative", in the sense that they reflect the belief that "normal" conditions prevail. There are various ways in which one can give a technical account of this kind of reasoning (see papers in Ginsberg, 1987). Here we shall concentrate on Reiters (1980) default logic, which can be employed to construct theories which minimize abnormality. In other words, given a set of propostions which we deem to be true, we can use default logic to extend this set in "likely" ways for problem solving purposes.

Reiter defines a *default theory* to be a pair (D, W). D is a set of default inference rules of the form

$$\frac{\alpha:\beta_1,\ldots,\beta_m}{\gamma},$$

where $\alpha, \beta_1, \ldots, \beta_m, \gamma$ are well-formed formulas of the predicate calculus which may contain free variables[2], and W is a set of closed formulas of the predicate calculus which can be thought of as incompletely specifying the state of the world.

α is called the *prerequisite* of the default: it consists of a set of *monotonic antecedents* that must be satisfied by the theory before the default is considered to be applicable to a given case. $\beta_1, \ldots, \beta_m$ are *nonmonotonic antecedents* that must be consistent with the current theory. γ is the *consequent* which is detached so long as the monotonic antecedents are satisfied and the nonmonotonic antecedents are consistent. We shall write defaults on a single line as $[\alpha: \beta_1, \ldots, \beta_m/\gamma]$.

The main idea behind default logic is that we want to be able to characterize precisely what it means to extend an incomplete theory of the world, like W, under the guidance of a set of defaults. The best way to think of theory extension is in terms of successive approximations to a stable state (or states). Roughly speaking, we let our initial theory W be the first approximation E_0, and then compute further approximations in a way that renders the resultant stable state(s) both consistent and deductively closed, while making as many assumptions as the defaults allow. That is more or less all the general reader needs to understand by the following fixed point equation.

$$E_0 = W$$

and, for $i \geqslant 0$,

$$E_{i+1} = \mathrm{Th}(E_i) \cup \{\gamma \mid [\alpha:\beta_1,\ldots,\beta_m/\gamma] \in D, \alpha \in E_i \text{ and } \neg\beta_1,\ldots,\neg\beta_m \notin E\}.$$

E is then an *extention* for (D, W) if and only if $E = \cup E_i$ from $i = 1$ to $i = \infty$.

[2] *Free variables* are variables not bound by a quantifer. Formulas containing no free variables are called *closed formulas.*

Defaults relate to abnormality in the following way: they allow us to assume the truth of certain propositions so long as we have no information to the contrary. If the contrary information is designated "abnormal", then we can use defaults of the form

$$\frac{\alpha : \neg AB(X_1) \ldots \neg AB(X_n)}{\gamma}$$

to establish a conclusion γ in the absence of positive proof. AB is the *abnormality predicate*, such that $AB(X)$ denotes that X is abnormal. Such a rule allows us to conclude γ, so long as certain background conditions, α, hold and we have no evidence of any abnormality cited in the nonmonotonic antecedents. Thus the default $[CAR(X){:}\ \neg AB(X)/START(X)]$ represents the belief that a car will start so long as it is not abnormal.

Another use for defaults in diagnostic reasoning is to represent hypotheses which, if true, would account for symptom data. We might wish to assume such hypotheses, for the sake of argument, if we cannot already rule them out. Thus defaults of the form

$$\frac{s_1 \wedge \ldots \wedge s_n : \gamma}{\gamma}$$

are *hypothesis schemas* which allow us to assume hypothesis γ in the presence of a set of symptoms $\{s_1, \ldots s_n\}$, as long as it is consistent to do so.

It is easy to see the default theories can have more than one extension. Consider the following example.

Example 1. Suppose that we have a set of defaults

$$D = \{\neg START(X){:}OUT\text{-}OF\text{-}GAS(X)/OUT\text{-}OF\text{-}GAS(X),\\ \neg START(X){:}\,CARB\text{-}FLOODED(X)/CARB\text{-}FLOODED(X)\}$$

and a theory

$$W = \{(\forall X)(CARB\text{-}FLOODED(X) \supset \neg OUT\text{-}OF\text{-}GAS(X)), \neg START(c)\},$$

where c is a particular car. Then (S, W) has the two extensions

$$\mathrm{Th}(W \cup \{CARB\text{-}FLOODED(c), \neg OUT\text{-}OF\text{-}GAS(c)\})$$

and

$$\mathrm{Th}(W \cup \{\neg CARB\text{-}FLOODED(c), OUT\text{-}OF\text{-}GAS(c)\}).$$

Thus there is an extension in which we assume that the car is out of gas, and an extension in which we assume that it has a flooded carburettor. These two assumptions are contradictory and therefore cannot coexit in the same extension.

The literature contains two main proposals for the application of default reasoning to diagnostic tasks. One approach uses defaults to minimize abnormality in a device with respect to a system description; the other uses defaults as hypothesis schemas in reasoning about a causal model. We shall see that each approach has its strengths and weaknesses, but that the former seems to be closer to the original notion of a default.

3 Diagnosis from first principles

Reiter's (1987a,b) theory of diagnosis from first principles involves working from a *system description* (SD) featuring a finite set of *system components*, and a set of observations (OBS). Both SD and OBS are finite sets of first-order sentences.

3.a Computing all diagnoses

A *diagnosis* for (SD, COMPONENTS, OBS) is a minimal set $\Delta \subseteq$ COMPONENTS such that

$$\mathrm{SD} \cup \mathrm{OBS} \cup \{\neg AB(c) \mid c \in \mathrm{COMPONENTS} - \Delta\}$$

is consistent, where AB is the abnormality predicate, as before. The idea is that a diagnosis keeps the set of faulty components as small as possible. We deem a component to be in the set only if its presence is required by a symptom in OBS.

Of course, there may be more than one way of constructing Δ, leading to multiple diagnoses. Although one could compute these sets by a simple generate and test algorithm, it is more efficient to use de Kleer and Williams' notion of a conflict set: a set of components $\{c_1, \ldots, c_n\}$ such that

$$\mathrm{SD} \cup \mathrm{OBS} \cup \{\neg AB(c_1), \ldots, \neg AB(c_n)\}$$

is inconsistent. Given a theorem prover that will generate conflict sets for (SD, COMPONENTS, OBS), Reiter shows how to compute all diagnoses of (SD, COMPONENTS, OBS). The key idea is that of a *hitting* set: a set formed from the collection of conflict sets C by taking at least one element from each member of C. *Minimal hitting sets*[3] correspond exactly to diagnoses as defined above. Reiter provides an algorithm for computing minimal hitting sets using a data structure called a *hitting set tree* (HS-tree). Heuristics are provided for pruning this tree, which are intended to reduce the cost of obtaining the minimal hitting sets.

These ideas are best illustrated with an example. We shall follow the literature by taking as our device the full adder: a one-bit adder with carry in and carry out. A logic diagram of the device is shown in Figure 3.

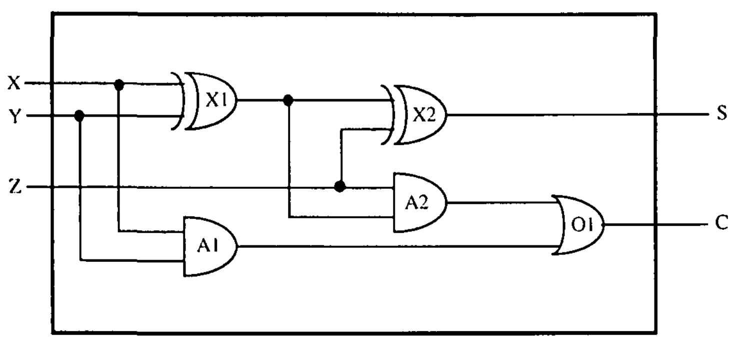

Figure 3 Diagram of a full added. X1 and X2 are exclusive-or-gates, A1 and A2 are and-gates, and O1 is an or-gate

This adder can be represented by the set

$$\mathrm{COMPONENTS} = \{\mathrm{A1}, \mathrm{A2}, \mathrm{O1}, \mathrm{X1}, \mathrm{X2}\}$$

plus a system description that describes the components in terms of their connectivity and their behaviour. We will not reproduce the system description here, but it contains Boolean equations, such as

$$XORG(X) \wedge \neg AB(X) \supset OUT(X) = xor(IN1(X), IN2(X))$$

which states that a normal exclusive-or-gate's output is the usual Boolean function of its two inputs, together with formulas describing circuit components and their connectivities.

[3] For a set H to be a minimal hitting set of a collection of sets C, we require that

1 $H \subseteq \bigcup_{S \in C} S$;
2 $H \cap S \neq \emptyset$ for each $S \in C$; and
3 there is no $H' \subset H$ for which both 1 and 2 hold.

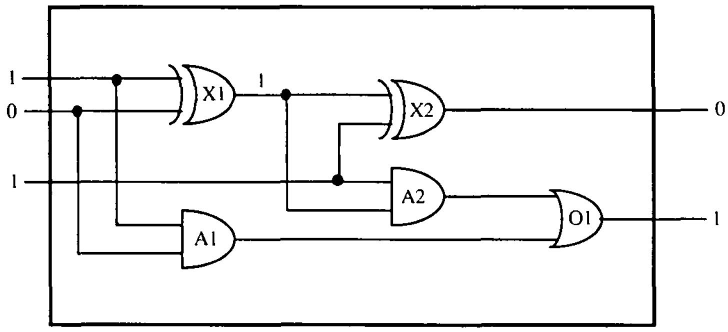

Figure 4 Correct outputs for inputs X = 1, Y = 0, Z = 1

Example 2. Given the inputs X = 1, Y = 0 and Z = 1, we would expect the outputs S = 0 and C = 1, as shown in Figure 4.

Suppose instead that the sum turns out to be 1 and the carry is 0 for the same inputs. One possible cause of these observations is that X1 is faulty. If X1 outputs 0 instead of 1, then this would produce the wrong behaviour, as shown in Figure 5.

Given that SD contains

$$XORG(X) \wedge \neg AB(X) \supset OUT(X) = xor(IN1(X), IN2(X))$$

and OBS contains

$$OUT(\mathrm{X2}) = 1$$

it is easy to see that

$$\mathrm{SD} \cup \mathrm{OBS} \cup \{\neg AB(\mathrm{X1}), \neg AB(\mathrm{X2})\}$$

is inconsistent, given the inputs. Either X1 passed on the correct outputs to X2, in which case X2 is abnormal, because

$$OUT(\mathrm{X2}) \neq xor(IN1(\mathrm{X2}), IN2(\mathrm{X2})),$$

or X1 computed the wrong outputs and is therefore abnormal, because

$$OUT(\mathrm{X1}) \neq xor(IN1(\mathrm{X1}), IN2(\mathrm{X1})).$$

Thus {X1, X2} is a conflict set for (SD, COMPONENTS, OBS). The other conflict set is {X1, A2, O1}, because one of X1, A2, O1 must be abnormal to explain the wrong carry, OUT(O1) = 0. In other words,

$$\mathrm{SD} \cup \mathrm{OBS} \cup \{\neg AB(\mathrm{X1}), \neg AB(\mathrm{A2}), \neg AB(\mathrm{O1})\}$$

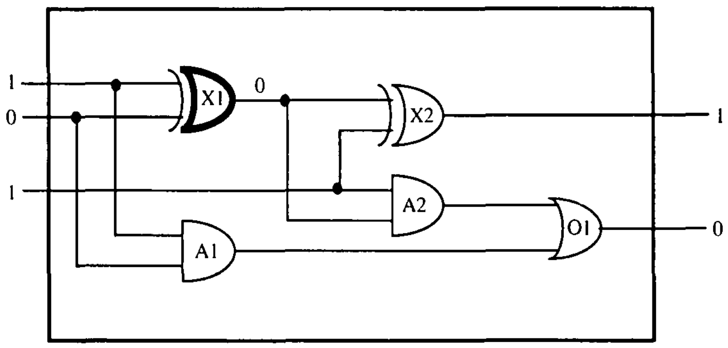

Figure 5 A possible diagnosis for the faulty outputs

is inconsistent. The minimal hitting sets of the set of conflict sets

$$C = \{\{X1, X2\}, \{X1, A2, O1\}\}$$

are $\{X1\}$, $\{X2, A2\}$ and $\{X2, O1\}$.

Figure 6 shows a hitting set tree for this example. The minimal conflicts form the nodes (shown in bold), while a labelled branch *B* below a given node *N* corresponds to the selection of component *B* from the set at node *N*. The hitting sets are just paths through the tree, and the minimal hitting sets are paths that are not subsumed by any other path. Thus the path X1 subsumes the path X1–X2, so X1–X2 is not minimal. Clearly the minimal paths are X1, X2–A2 and X2–O1.

The reader can easily verify that the minimal hitting sets derived in Figure 6 are those sets *S* with the following properties: (i) the failure of components contained in *S* would explain the wrong outputs, and (ii) there is no subset of *S* that satisfies (i). Such sets therefore correspond to the diagnosis of (SD, COMPONENTS, OBS), as defined by Reiter's theory.

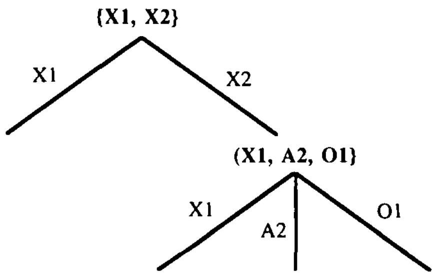

Figure 6 A hitting set tree for Example 2

Once we have generated alternative diagnoses, we are interested in ruling some of them out. Different diagnoses make different predictions concerning what additional observations we might expect. If we fail to observe phenomena predicted by a diagnosis, then we may eliminate that hypothesis from consideration. A diagnosis Δ *predicts* some first-order sentence *P* iff

$$\mathrm{SD} \cup \mathrm{OBS} \cup \{\neg AB(c) \mid c \in \mathrm{COMPONENTS} - \Delta\} \models P.$$

In other words, *P* follows as a logical consequence from the facts of the case, SD and OBS, and the assumption that all components not in the diagnosis are working properly.

The connection with default logic is as follows. Reiter notes that we can use default rules to minimize abnormality in the device in such a way that diagnoses correspond to extensions for a default theory. Notably, *E* is an extension for the default theory

$$(\{:\neg AB(c)/\neg AB(c \in \mathrm{COMPONENTS}\}, \mathrm{SD} \cup \mathrm{OBS})$$

iff there is some diagnosis Δ for (SD, COMPONENTS, OBS) such that

$$\begin{aligned} E &= \{P \mid \Delta \textit{ predicts } P\} \\ &= \mathrm{Th}(\mathrm{SD} \cup \mathrm{OBS} \cup \{\neg AB(c) \mid c \in \mathrm{COMPONENTS} - \Delta\}). \end{aligned}$$

Roughly speaking, the extension corresponding to a diagnosis Δ completely describes the way we would expect the world to be, given that all and only the faults hypothesized by Δ actually occurred. In other words, extension contains all propositions derivable from both the facts of the case and the assumption that all other components are functioning correctly.

These equivalences are not too surprising (in retrospect). As Reiter points out, diagnostic reasoning is clearly nonmonotonic, in the sense that none of our current hypotheses may survive a new observation. Thus the correspondence between diagnoses and extensions captures perfectly the tentative nature of diagnostic hypotheses.

Finally, it is worth noting that the first principles approach does not rule out the use of device-specific information about how faults may manifest themselves. If such information is available, then it can be included in the system description, using the abnormality predicate:

$$COMPONENT(X) \wedge AB(X) \supset S_1(X) \vee \ldots \vee S_n(X)$$

where $COMPONENT(X)$ denotes that X is a particular type of component, and the $S_i(X)$ exhaust the possible symptoms of X's abnormality.

3.b Applicability of Reiter's approach

Reiter's approach appears to be both elegant and formally correct, but it places three important constraints upon the formulation of the diagnostic problem.

- In practice, we can only deal with faulty components. If we try to extend the theory to deal with the connections between components, then the number of possible diagnoses becomes much too large. (For nontrivial circuits, the component-based diagnoses are quite numerous enough, and the computational cost of generating them is high.)
- We need a complete system description; not a trivial requirement for an arbitrary system. Determining the completeness and consistency of the system description need not be straightforward, yet if these properties do not hold then the method will not be sound or complete. Such descriptions are seldom available for complex devices, although it is possible that they could be generated as a by-product of computer-aided design.
- The method assumes the availability of a complete theorem prover for the underlying logic. Again, this is no problem for the full adder example, because the theory of Boolean equations is decidable. But reasoning about consistency is impossible in the general first-order case in terms of which the theory of diagnosis is formulated. Furthermore, the extension membership problem is not even semi-decidable (Reiter, 1980, Theorem 4.9). Thus any automatic diagnostic system using defaults in the full first-order case must have recourse to heuristics that sacrifice completeness, and may even lead to erroneous diagnoses.

Of course, in many domains, we are interested in doing more than merely identifying faulty components. "There is something wrong with the patient's heart" is not very illuminating as a medical diagnosis. Therapy and prognosis ultimately depend upon having more precise hypotheses. However, in many electronic devices, where components are relatively cheap and easy to replace, identifying sets of components that may be faulty is all we need to effect repair. Often, systematic "board swapping" replaces the taking of additional measurements; sometimes it replaces the diagnostic process altogether.

3.c Comparison with GDE

To lend concreteness to the discussion, let us redo the diagnosis of the adder in Example 2, using GDE's approach. Figure 7 shows the lower levels of an ATMS environment lattice for this problem. Each node other than Ø denotes a fault state, e.g., the node labelled "A1, A2" represents the state of affairs where only the and-gates are faulty.

As before, the minimal conflicts are {X1, X2} and {X1, A2, O1}. Given the observations, we can say that one of X1 and X2 must be faulty, and that components X1, A2 and O1 cannot all be working correctly. The conflicts allow us to rule out certain nodes, which are shaded in Figure 7. Thus, we can rule out the node labelled "A1", since both X1 and X2 are deemed to be functioning correctly at this node. The nodes that are left in the lattice fragment are all candidate explanations of the data.

All that remains is to read off the minimal candidates {X1}, {X2, A2}, {X2, O1}: these are clearly the minimal sets of components whose failure is consistent with the observations. Candidates such as {A2, X1} are certainly consistent with the data, but they are subsumed by other candidates, in this case {X1}. The upper levels of the lattice contain no minimal candidates, and are therefore not included in Figure 7.

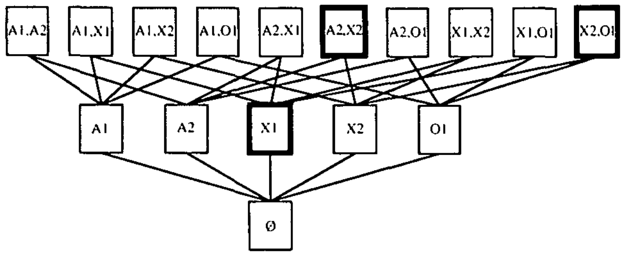

Figure 7 GDE's candidate space for Example 2

It is relatively easy to see that the diagnoses of the two approaches coincide: Reiter's minimal hitting sets are just GDE's minimal candidates (compare Figures 6 and 7). Both methods rely on an inference engine to compute the conflict sets; Reiter assumes that a general purpose theorem prover is available for this. GDE uses the ATMS to calculate the consequences of OBS in each environment, starting with the empty one, and looks for an inconsistency signalling that the environment is a conflict. Needless to say, the general problem of determining these sets is intractable[4] for large numbers of components. However, the search space is normally much sparser than the full lattice, because many joins will contain no consequences not found in their subsets.

The main differences between the two methods are as follows. Firstly, Reiter's algorithm interleaves the generation of conflict sets with the generation of minimal hitting sets, while GDE generates all the minimal conflict sets before computing minimal candidates. Moreover, the algorithm is incremental, in the sense that it generates diagnoses in order of increasing cardinality. Secondly, GDE exploits useful properties of the ATMS, e.g., the ability to record the drawing of inferences while computing conflicts, as well as using probabilities to guide the search for minimal candidates and cost functions to inform the selection of measurements. Although Reiter proves a number of results about measurements, he provides no guidelines for selecting the next measurement to make.

4 Explanations and scenarios

Poole (1988a) has a rather different proposal for using defaults in diagnosis. Given a set of facts F and a set of defaults D, a *scenario* of (D, F) is defined to be a set $D^* \cup F$, where D^* is a set of ground instances of elements of D such that $D^* \cup F$ is consistent. The facts are closed first-order formulas, while the defaults are intended to represent "possible hypotheses", i.e., hypothesis schemas which associates faults with their symptoms, and can be used to construct scenarios in which symptoms are explained, and can be used to construct scenarios in which symptoms are explained. An *explanation* of P from (D, F) is a scenario of (D, F) which logically implies P. The deductive closure of a maximal scenario of (D, F) corresponds to an extension of (D, F) in Reiter's logic.

4.a Diagnosis in Theorist

Diagnostic reasoning from a causal model is really a kind of *abduction*. As a first approximation, abduction can be viewed as the inverse of deduction. In deduction, we are interested in the *consequences* C of a set of premises B in the context of some theory or set of axioms A. We can represent this as the familiar pattern of inference: $A, B \vdash C$, where "$\vdash$" denotes derivability. In abduction, we begin with the consequences C, and we are most interested in what *premises B* we have to add to our *axioms A* such that $A, B \vdash C$. Normally, we want B to be *parsimonious*, i.e., to contain no propositions that are not needed to derive C, and normally there is more than one set of premises that will satisfy these requirements.

[4] For GDE, the problem is NP-complete (i.e., the time complexity of the computation on a serial processor is exponential in the number of components), since it is isomorphic to propositional satisfiability.

The application of Poole's framework to diagnosis is briefly illustrated in Poole et al. (1987), which describes the Theorist program. As the simplest possible example, let

$$F = \{(\forall X)(FLAT\text{-}BATTERY(X) \supset \neg START(X))\};$$

$$D = \{: FLAT\text{-}BATTERY(X) / FLAT\text{-}BATTERY(X)\}.$$

Then

$$S = F \cup \{: FLAT\text{-}BATTERY(c) / FLAT\text{-}BATTERY(c)\}$$

is a scenario which explains $\neg START(c)$, for some car c. It explains $\neg START(c)$ because the nonmonotonic antecedent $FLAT\text{-}BATTERY(c)$ is consistent with F and can therefore be added to S. Once this has been done, $\neg START(c)$ follows from S as an ordinary tautological consequence. S is a maximal scenario, because it contains all instances of all applicable defaults (assuming that c is the only object in the universe). In fact, its deductive closure is the sole extension of (D, F).

Here we do not have a complete system description of an automobile, which would require a substantial knowledge base representing many man-years of knowledge engineering effort. Instead, F contains an incomplete causal theory, detailing the consequences of known faults. Next, we look at the way in which Poole's theory applies to multiple faults and multiple symptoms, using a slightly more complicated example.

Example 3. Suppose that

$$\begin{aligned} F = \{&(\forall X)(FLAT\text{-}BATTERY(X) \supset \neg START(X)),\\ &(\forall X)(FLAT\text{-}BATTERY(X) \supset DEAD(RADIO\text{-}OF(X))),\\ &(\forall X)(DISCONNECTED(RADIO\text{-}OF(X)) \supset DEAD(RADIO\text{-}OF(X))),\\ &(\forall X)(OUT\text{-}OF\text{-}GAS(X) \supset \neg START(X))\}; \end{aligned}$$

$$\begin{aligned} D = \{&: FLAT\text{-}BATTERY(X) / FLAT\text{-}BATTERY(X)\\ &: OUT\text{-}OF\text{-}GAS(X) / OOUT\text{-}OF\text{-}GAS(X),\\ &: DISCONNECTED(RADIO\text{-}OF(X)) / DISCONNECTED(RADIO\text{-}OF(X))\}. \end{aligned}$$

Let us seek an explanation of the set of symptoms

$$S = \{\neg START(c), DEAD(RADIO\text{-}OF(c))\}.$$

There are now eight scenarios of (D, F) according to Poole's definition, five of which explain the symptoms. However, only two of these explanations are really interesting:

$$F \cup \{: FLAT\text{-}BATTERY(c) / FLAT\text{-}BATTERY(c)\}$$

$$\begin{aligned} F \cup \{&: OUT\text{-}OF\text{-}GAS(c) / OUT\text{-}OF\text{-}GAS(c),\\ &: DISCONNECTED(RADIO\text{-}OF(c)) / DISCONNECTED(RADIO\text{-}OF(c))\}. \end{aligned}$$

The other three are redundant, as they account for data several times over. Although Poole writes that "a commitment to the Theorist framework is not a commitment to any particular control structure", it is clear that one would like to generate explanations in an order that reflects minimality (as in GDE) or cardinality (as in Reiter's algorithm). Console et al. (in press) have provided an account of diagnosis from incomplete causal models that incorporates parsimony criteria in a natural way.

In a recent paper, Bylander et al. (1989) show that abduction problems are only tractable if they satisfy a rather restricted set of conditions:

- *Independence*. An explanation that is a composite hypothesis accounts for a datum if and only if some hypothesis in the set accounts for it. Thus, we cannot allow the presence of one hypothesis in the set to "disable" or "cancel" the explanatory power of another hypothesis. However, hypotheses in a composite are allowed to interact additively, i.e., we allow the conjunction of two or more hypotheses to explain more data than the union of the data they individually explain.

- *Compatibility.* The pool of hypotheses from which we draw to form composites must be consistent, i.e. no combination of hypotheses must rule any other hypothesis out.

The authors show that there can be no tractable (i.e., polynomial time) algorithm for finding an explanation for problems which allow cancellation. For problems which allow incompatible hypotheses, finding a minimal explanation is tractable, but finding a "best" explanation (e.g., based on some plausibility function defined over hypotheses) is not. These results are not surprising, but they demonstrate that the causal approach is in no better shape than the first principles approach with respect to computational complexity, and so control issues are crucial.

4.b Comparison with diagnosis from first principles

Reiter's use of defaults to minimize abnormality seems intuitively right for diagnosis. We assume that components of a device are working unless we have evidence to the contrary, and in the event of a fault we prefer parsinomious explanations of the symptoms. Poole's use of defaults, on the other hand, seems too permissive; it tends to maximize abnormality, and we are supplied with no means of preferring minimal explanations or preventing redundant explanations from being generated.

Furthermore, it is arguable that Poole's use of defaults is contrary to our understanding of what a default is. A default of the form $(: P / P)$ is usually taken to mean that, all other things being equal, "P is preferred to its negation" or "P is normally the case". But this is not what Poole's defaults mean. The informal reading of $(: P / P)$ in a scenario seems to be "if it is consistent to suppose that the fault P is present, then it is present". But it is presumably not intended to mean "we prefer the presence of P to its absence" or "the fault P is normally present".

A related problem arises because Poole is content with a conventional (Tarskian) semantics for his logic, which does not assign any meaning at all to defaults. He argues that it is scenarios (not defaults) which have a semantic link to the world; defaults are just assumptions made in building scenarios. Yet, in the course of diagnostic reasoning, we routinely hypothesize entities that may turn out not to exist, such as a tumour in a patient's lung. As Reiter (1980, p. 91) points out, different extensions of a default theory can support different ontologies, e.g., via defaults of the form $(\alpha{:}(\exists X)F(X)/(\exists X)F(X))$. Tarskian semantics cannot accommodate varying domains in any straightforward manner, and so the denotation of "possible individuals" is left completely unaccounted for.

Reiter's theory was unimplemented at the time of its publication, although it has since been implemented by Smith (1988) using Overbeek and Lusk's (1984) Interactive Theorem Prover. Smith has shown that Reiter's algorithm for generating a pruned hitting set tree is in fact incomplete, because the pruning rule interacts with other rules. In the corrected algorithm, it turns out that pruning does not always result in fewer calls to the theorem prover.

In the implementation of Theorist, it appears that explanatory theories are generated by depth-first search and then tested by asking further questions of the user. No details are supplied about the order in which explanations will be generated, and the only basis put forward for preferring one explanation to another is *specificity*. Thus we prefer the diagnosis "bronchial pneumonia" to the diagnosis "respiratory disorder". However, there may be multiple theories at the same level of specificity which are consistent with the data.

In Poole (1988b), rather large claims are made for the generality and utility of Theorist as a framework for comparing different approaches to diagnosis. An attempt to compare a number of logic-based diagnostic systems on semantic grounds can also be found in Jackson (1989). At the present time, it seems to be an open question as to what constitutes a good or useful framework for such comparisons.

5 Counterfactuals and diagnosis

In addition to the work of Reiter and Poole, there is another approach to diagnosis which uses nonmonotonic reasoning, namely that advocated by Ginsberg (1986), which employs *counterfactual reasoning* rather than reasoning by default.

Counterfactuals are conditional statements in which the antecedent is deemed to be false, e.g., "If Waldo were rich, he'd live in Las Vegas". Considered as material conditionals, all such statements are trivially true; thus "If Waldo were rich, he'd live in Milwaukee" would also be a true statement, if Waldo were not rich. Yet the intention seems to be something like: "If things were more or less as they are, except that Waldo were rich, he'd be living in Las Vegas", which rules out living in Milwaukee. It is customary to write a counterfactual conditional of the form "if P then Q" as "$P > Q$" rather than "$P \supset Q$", in order to preserve this distinction.

Counterfactuals are not truth functional, because they can only be evaluated relative to: (i) some theory of what the world is like, and (ii) some notions about what the world *might* be like if certain things were to change. Informally, we want $P > Q$ to be true if and only if Q is true in all those plausible worlds where P holds which are most similar to the real world. From a formal point of view, the problem is to specify what we mean by "plausible" and "similar". Counterfactual logic is nonmonotonic, because it is possible that P counterfactually implies Q but that $P \wedge R$ does not. (For example, R might be $\neg Q$.)

5.a Ginsberg's construction

As counterfactual statements are not truth-functional, we require some kind of *construction*, consisting of *possible worlds* other than the current world, which we can inspect in order to decide whether or not a counterfactual holds in the current world. Ginsberg's construction for counterfactuals has the following components:

- an initial set S of sentences;
- a predicate B on 2^S, the power set of S;
- a partial order $<$ on 2^S which extends set inclusion and respects B.

S is a theory of the world, B is a "bad world" predicate which declares some worlds to be implausible, and $<$ is a comparator of possible worlds along the plausible dimension.

Ginsberg's definition of plausible, similar worlds is as follows. A *possible world for P in S* is any subset T of S such that

$$T \nvDash \neg P;$$

$$\neg B(T);$$

and T is maximal with respect to $<$, given these restrictions.

The set of such worlds, $W(P, S)$, is given by

$$\{T \subseteq S \mid T \nvDash \neg P \;\&\; \neg B(T) \;\&\; \{U \subseteq S \mid T < U \;\&\; U \nvDash \neg P \;\&\; \neg B(U)\} = \varnothing\}.$$

A counterfactual $P > Q$ is now true with respect to S iff

$$T \cup \{P\} \nvDash Q \quad \text{for all } T \in W(P, S).$$

The "worlds" in $W(P, S)$ are partial; they are really theories that can be taken as representing sets of worlds. The basic idea is that if $T \in W(P, S)$ then T represents a minimum revision to S such that $T \cup \{P\}$ is consistent, T is not implausible, and no other revision of S is more plausible than T.

5.b Bad worlds and preferences

Ginsberg's definition is best understood with the aid of an example, adapted from his paper.

Example 4. Let b denote "I have a boat", o "I have oars", and r "I can row across the river", and let $S = \{b, \neg o, \neg r, b \wedge o \supset r\}$. We want to show that $o > r$, i.e., that if I had some oars I could row across the river. To demonstrate this, we must show that r holds in all possible worlds most similar to S in which o holds.

Let us begin by assuming that B, the bad world predicate, is empty, and that the world comparator $<$ is just set inclusion. Then $W(o, S)$ is the set of all subsets T of S that do not entail $\neg o$ and which are maximal with respect to $<$:

$$W(o, S) = \{\{\neg r, b \wedge o \supset r\}, \{b, \neg r\}, \{b, b \wedge o \supset r\}\}.$$

Note that it is not the case that $T \cup \{o\} \models r$ for all $T \in W(o, S)$, so if B is empty and $<$ is set inclusion, it is not the case that $o > r$. The problem is that the three elements of $W(o, S)$ weaken S in different ways to admit the addition of o. The first removes b, the second removes the conditional $b \wedge o \supset r$, while the third removes $\neg r$. Only the latter renders r derivable if o is added.

Ginsberg's solution to this problem has two parts. Firstly, the retraction of the conditional is prevented by means of the statement

$$B(S - \{b \wedge o \supset r\}).$$

We say that the conditional is *protected*, because there are no circumstances under which we are prepared to retract it. Secondly, the retraction of b is prevented by means of the preference

$$W - \{b\} < W - \{\neg r\},$$

which states that losing the boat is less plausible than being able to row across the river.

This machinery has the desired effect.[5] The revised $W(o, S)$ only contains $\{b, b \wedge o > r\}$. r follows from the addition of o to this set, and so $o > r$.

5.c Counterfactuals and minimizing abnormality

Given a diagnostic problem (SD, COMPONENTS, OBS), as in section 3, Ginsberg's approach is to view a diagnosis as a counterfactual consequence of the observations. As in Reiter's approach, the basic idea is that we compute those deviations from normality which are forced upon us by the behaviour of the device. We perform this computation by retracting assumptions to the effect that components are not abnormal.

Put another way, we start from the "perfect" world in which the device is functioning correctly, represented by

$$\text{SD}^* = \text{SD} \cup \{\neg AB(c) \mid c \in \text{COMPONENTS}\}.$$

Then we are made aware of observations which suggest that the device is *not* working properly. We are interested in discovering those worlds most similar to the perfect world in which the malfunction can be explained. Here, most "similar" means departing minimally from the perfect world, but nonetheless explaining all the data. Clearly, there is a possible world in which *every* component has failed, but this world will normally overexplain the data, and will therefore be more "dissimilar" than it needs to be.

Let us refer back to the full adder of Example 2. Given SD* as defined above and OBS as in the example, W(OBS, SSD*) contains the following worlds:

$$\text{SD}^* - \{\neg AB(\text{X1})\}$$
$$\text{SD}^* - \{\neg AB(\text{X2}), \neg AB(\text{A2})\}$$
$$\text{SD}^* - \{\neg AB(\text{X2}), \neg AB(\text{O1})\},$$

so long as we protect all the formulas in SD. Note that these retractions corresponds to Reiter's diagnoses of (SD, COMPONENTS, OBS): {X1}, {X2, A2}, and {X2, O1}.

Furthermore, it is easy to see that the observations counterfactually imply $\neg AB(\text{A1})$, because

[5] A critique of Ginsberg's construction can be found in Winslett (1988).

$\neg AB(\mathrm{A1})$ holds in each of the worlds listed above. Nonmonotonicity enters at this point, because it is clear that $\neg AB(\mathrm{A1})$ may need to be retracted in the light of further information.

In summary, if the system description is complete, then the observations logically imply alternative faults in the device, as in Reiter's approach. Furthermore, the normality of the other components follows as a counterfactual consequence of the observations. Thus the notion of a minimal fault, which we encountered in the work of Reiter, also falls naturally out of the counterfactual construction.

The limitations of this approach are similar to the approach of Reiter, and therefore do not need to be considered in detail. We must have a complete and consistent system description and we will normally need additional device assumptions, e.g., to assure us that the wiring is not faulty. Presently, it is not clear whether one of these approaches offers significant advantages over the other, although Reiter's proposal is the more fully worked out.

6 Comparison with other approaches

In the introduction, we contrasted logic-based approaches to diagnosis using system descriptions with approaches more commonly associated with expert systems technology, such as heuristic classification. In this, the concluding section, we shall dig a little further into the AI literature to effect a slightly deeper (but just as brief) comparison between the utility of nonmonotonic approaches to diagnosis and the utility of approaches based on other kinds of causal reasoning. In particular, we shall concentrate upon: (i) rule-based systems which also use an explicit representation of causal knowledge, and (ii) probabilistic models of causal reasoning.

6.a Rules generated from domain models

A number of recent approaches to expert system construction rely upon the construction of a *domain model*: some declarative representation of objects in the domain and the (primarily causal) relationships between them. This model is then used as a basis for a further knowledge engineering effort, usually the generation of heuristic rules. These rules are sometimes generated automatically from the model, or else the model is used to supply a knowledge acquisition program with interview strategies for the elicitation of rules from an expert.

The following systems are prominent examples of this approach: XPLAIN (Swartout, 1983), MORE (Kahn, 1988) and MOLE (Eshelman, 1988). They are worth mentioning here because they attempt to combine deeper causal knowledge with the shallower experiential knowledge we associate with the rule-based approach. In particular, we need to consider whether or not this use of domain models takes some of the force out of the arguments for logic-based approaches to diagnosis.

To give concreteness to the discussion, let us concentrate on a particular system, MORE. A domain model for MORE contains *symptoms*, *attributes* that serve to further discriminate symptoms, and events that are possible causes of symptoms and therefore serve as *hypotheses*. There are also *background conditions* which make symptoms more or less likely given the occurrence of a hypothetical cause, conditions which make hypotheses more or less likely, *tests* that can be used to determine the presence or absence of any such background conditions, and *test conditions* that have a bearing upon the accuracy of the tests.

This representation is used to generate three kinds of production rule:

(1) *diagnostic rules* which perform the heuristic mapping between symptoms and hypotheses;
(2) *symptom confidence rules* which qualify the data abstraction implicit in the symptom space with estimates of test reliability under different background conditions;
(3) *hypothesis expectancy rules* which qualify the solution abstraction implicit in the hypothesis space with estimates of the prior probability of hypotheses under different background conditions.

Here a causal model is used not as a replacement for heuristic rules, but as an intermediate stage of representation from which such rules can be derived. Having such an underlying model is intended to make the process of building and debugging rule sets easier. In some systems, such as XPLAIN, it is also used to produce explanations which go beyond rule tracing. (No work appears to have been done so far on the explanation of diagnoses derived from first principles.)

Although systems like MORE are still in their infancy, they demonstrate that the two approaches of: (i) causal reasoning from faults to symptoms and (ii) heuristic mapping from symptoms to faults can be mutually supportive rather than mutually exclusive strategies. This hybrid approach to diagnosis appears to be more flexible than approaches based solely on a device description, because one does not require a complete domain model in order to generate useful rules. Neither does one need as many simplifying assumptions as seem to be required by logic-based approaches. As Clancey (1985) has pointed out, in the deep models required by the latter are hard to construct, even for relatively simple electronic devices. Even Genesereth (1984) acknowledges that "not all design descriptions are tuned for the task of diagnosis"—although part of the motivation for using design descriptions is surely that (unlike rule sets) they are not supposed to need tuning!

6.b The role of uncertainty in diagnosis

In contrast with approaches from nonmonotonic logic, diagnosis by expert system typically involves some form of inexact reasoning. This is partly because, as we noted in the introduction, one is reasoning from symptoms to faults, rather than from faults to symptoms. (It is only rarely the case that a set of symptoms uniquely identifies a disorder; physicians call such symptom sets *pathognomic.*)

However, we have seen that nonmonotonic approaches still need some mechanism for comparing the large numbers of explanations that they generate. Thus there may be a place here for quantitative methods, such as certainty factors or Bayesian belief updating. The lack of such mechanisms appears as a major weakness of some of the systems reviewed herein.

This section concentrates upon the main problems involved in trying to incorporate quantitative methods into nonmonotonic diagnostic systems. The review will have to be cursory, with an emphasis on trying to ask the right questions rather than providing any answers. Although no pretence is made at completeness (or originality), it is claimed that any proposal must address *at least* the issues raised here.

- *Independence assumptions* continue to cause problems for those who would incorporate Bayesian belief updating into diagnostic systems. Although the algorithms for propagating degrees of belief through causal trees have nice computational properties, and do not require the specification of too many probabilities, the problem is greatly complicated once we are dealing with general networks. Pearl (1988) has proposed some imaginative solutions to this problem, but it still remains to be seen whether Bayesian methods will have practical utility in the diagnosis of complex devices.
- Plausibility is not the only factor we need to to take into account when scheduling hypotheses for further processing. Two factors of equal importance are *risk* and *cost.* Risk complicates matters because a hypothetical fault with low plausibility may nonetheless be critical for health and safety, and therefore takes precedence over more plausible hypotheses. Cost complicates matters because there is considerable merit in performing quick or cheap tests first, regardless of the relative plausibility of hypotheses. Such considerations have figured in many expert systems, e.g., INTERNIST (Pople, 1977).
- It is not obvious that quantitative methods are the only appropriate way of scheduling hypotheses or preferring one explanation to another. Given that we are mostly interested in *ordering* hypotheses, and given that the grounds for ordering hypotheses involve domain-specific knowledge that can be reasoned about, there is at least a case for exploring non-numerical methods, e.g., reasoning with statements of the form "under conditions *C*, I prefer explanation *E*

to explanation F". The apparent precision of quantitative methods may be illusory in cases where we are assessing probabilities of failure in novel devices where there is very little experiential basis for subjective judgements.

The research literature on the application of nonmonotonic reasoning to diagnosis has some way to go towards dealing with these difficult problems. As noted in section 1.2, de Kleer and Williams (1986) determine the next measurement to make using a cost function based on entropy, and use probabilities to search for candidates. Their simplifying assumption that components fail independently is more reasonable in some domains than in others, but relaxing it requires the expert to supply large numbers of additional probabilities.

Other researchers have augmented ATMS with the Dempster–Shafer theory of evidence (e.g. Provan, 1988; Laskey and Lehner, 1988). However, computing the Dempster–Shafer mass function is intractable[6] in the worst case. As in basic ATMS, efficiency depends upon being able to ignore whole areas of the search space due to limited interactions between components. Also worth noting is Peng and Reggia's (1986) account of hypothesis plausibility in terms of Bayesian decision theory. They suggest using relative likelihood as a heuristic function to guide a best-first search for irredundant diagnoses.

As mentioned earlier, Smith (1988) has implemented Reiter's theory of diagnosis, and her thesis contains a number of suggestions concerning ways in which a first principles program might "home in" on the correct diagnosis.

- Collect more than one set of observations, e.g., using different inputs. This makes more information available about the behaviour of the device. Smith presents an algorithm for creating a composite diagnosis from the individual diagnoses derived from different observation sets.
- Use probabilities of component failures (if available) to guide the construction of minimal hitting sets, rather as de Kleer and Williams used them to guide the generation of candidates.
- Smith suggests a heuristic for generating tests based on the notion of *functional dependency* between components. Briefly, component C1 is functionally dependent on component C2 if and only if the value of the output of C2 must be known to determine the value of the output of C1; C2 is therefore in the *dependency set* of C1. Each suspect component is assigned a weight based on the size of its dependency set and the number of minimal conflict sets in which it appears, and the component with the largest weight is selected for measurement.

Although lacking an information-theoretic basis, it is clear that testing components with large weights will tend to reduce the number of competing diagnoses, since such measurements introduce powerful constraints upon the states of other components. An advantage of Smith's heuristic is that it does not require any information which is not already included in the system description (SD) or the observation set (OBS). It is possible to identify the dependency set of a component using simultation techniques, such as constraint propogation.

6.c Summary and conclusions

Proposals for the application of nonmonotonic logic to diagnosis are intrinsically interesting for a number of reasons:

- Proposals based on device descriptions give us an entirely different perspective on the knowledge engineering problem by shifting some of the responsibility for supplying knowledge from the troubleshooter to the designer. This may facilitate the knowledge acquisition process, if the requisite information is readily available in a form that lends itself to declarative representation in some logical language.
- Proposals based on causal models also facilitate knowledge acquisition in domains where the link

[6] Computing the mass function can be represented as a Network Reliability problem and shown to be NP-hard.

between cause and effect is relatively well understood. Expert system architects are also discovering the advantages of having an intermediate causal representation, as we have seen.

- There is little evidence that human experts engage in reasoning that is strictly probabilistic. Experts often agonize over the attachment of numerical weights to rules, and their main concern is often to create a partial order of rule strengths (see Buchanan and Shortliffe, 1984, p. 155). Some non-numerical features of nonmonotonic reasoning, e.g., consistency among assumptions, and similarity among possible worlds, appear to provide alternative mechanisms for diagnostic inference.

There are a number of problems with the present versions of these proposals which suggest that we should not expect results too soon.

- If design descriptions need to be tuned for diagnosis, then we need some methodology for performing this task (and maybe ideas for automating it).
- We need to understand causal reasoning far better than we do at present. It is not clear that present approaches are adequate for complex tasks, or even convincing from a formal point of view. For example, few AI models of causation incorporate temporal considerations (although see Shoham, 1988).
- The preference machinery provided by existing nonmonotonic approaches is not sufficient for the task of ordering hypotheses, and the tractable methods proposed by Bayesians embody quite restrictive assumptions.

It is possible that diagnosis from first principles, in its purest form, will always be confined to certain domains, namely those which are governed by a finite (and preferably small) set of laws that are capable of being precisely stated. Thus it is no coincidence that the examples are almost invariably drawn from the domain of electronic circuits.[7] Certainly more research will be required to justify some of the claims to generality that have been made in the literature. Nevertheless, the link that such work is forging between design and diagnosis is an extremely valuable one. Similarly, the link between causal reasoning and simulation is one that is worth exploring, even though our current models of causation are rather rudimentary. The safest prediction that one can make is that insights from nonmonotonic logic may be successfully incorporated into hybrid systems which use deep knowledge in combination with heuristic rules.

References

Buchanan, BG and Shortliffe, EH eds, 1984. *Rule-based expert systems*, Reading, Massachusetts: Addison-Wesley [A collection of seminal papers on expert systems by researchers at Stanford University]

Bylander, T, Allemang, D, Tanner, MC and Josephson, JR, 1989. "Some results concerning the computational complexity of abduction" In: *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning*, pp 44–54, Morgan Kaufmann [A useful and clearly presented paper that outlines the conditions under which abductive inference is tractable]

Clancey, WJ, 1985. "Heuristic classification" *Artificial Intelligence*, **27** 289–350 [A theoretical paper that attempts to reconstruct the problem solving paradigm employed by many rule-based expert systems]

Console, L, Dupré, DT and Torasso, P, in press, "A theory of diagnosis for incomplete causal models" To appear in: *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Detroit Michigan, August 1989 [A formal treatment of causal reasoning with incomplete knowledge that makes interesting connections with nonmonotonic logic]

Davis, R, 1984. "Diagnostic reasoning based on structure and behavior" *Artificial Intelligence* **24** 347–410 [An influential paper on electronic troubleshooting from first principles. It is not reviewed here because it makes no explicit connections with nonmonotonic logic. It is well worth reading, nonetheless]

de Kleer, J, 1986. "An assumption-based TMS" *Artificial Intelliegence* **28** 127–162 [The main paper on ATMS. It is neither clear not concise, but it is important]

de Kleer, J and Williams, BC, 1987. "Diagnosing multiple faults" *Artificial Intelligence* **32** 97–130 [A difficult

[7] Even in this domain, the method cannot be applied to bridge faults, where the system description no longer fits the actual circuitry of the device; for a treatment of such faults, see Davis (1984).

but important paper that repays study. If you read the version in Ginsberg (1987), note that the columns of text on page 382 are the wrong order!]

Eshelman, L, 1988. "MOLE: A knowledge acquisition tool for cover-and-differentiate systems" Chapter 3 of Marcus, S ed., *Automating knowledge acquisition for expert systems*, Boston, Massachusetts: Kluwer Academic [Describes a knowledge acquisition tool for systems that use a form of heuristic classification. MOLE is interesting because it reasons explicitly about the space of possible explanations]

Genesereth, MR, 1984. "The use of design descriptions in automated diagnosis" *Artificial Intelligence* **24** 411–436 [Describes one of the first logic-based diagnosis programs]

Ginsberg, M, 1986. "Counterfactuals" *Artificial Intelligence* **30** 35–80 [A wide-ranging treatment of counterfactuals and their relevance to AI applications, such as diagnosis and planning]

Ginsberg, M, 1987. *Readings in nonmonotonic reasoning*, Los Altos, California: Morgan Kaufmann [A good collection of important papers on nonmonotonic logic at a price you can afford]

Jackson, P, 1989. "Propositional abductive logic" In: *Proceedings of the 7th Conference on Artificial Intelligence and the Simulation of Behaviour*, pp 89–94, London: Pitman [An attempt to provide a proof theory and semantics for abductive inference]

Kahn, G, 1988. "MORE: From observing knowledge engineers to automating knowledge acquisition" Chapter 2 of Marcus, S, ed., *Automating knowledge acquisition for expert systems*, Boston Masschusetts: Kluwer Academic [An attempt to automate the acquisition of diagnostic knowledge using causal models]

Laskey, K and Lehner, PE, 1988. "Belief maintenance: An integrated approach to uncertainty management" In: *Proceedings of the 7th National Conference on Artificial Intelligence*, pp. 210–214, American Association for Artificial Intelligence [A convincing theoretical account of the combination of ATMS and Dempster–Shafer theory]

McDermott, D, 1987. "A critique of pure reason" *Computational Intelligence* **3** 151–160 [McDermott's now (in)famous attack on logic-based problem solving: still food for thought]

Overbeek, R and Lusk, E, 1984. "The automated reasoning system ITP—user's manual" Technical Report ANL-84-27, Argonne National Laboratory [The theorem prover upon which Smith's implementation of Reiter's theory was based, chosen partly for its effective treatment of equational theories]

Pearl, J, 188, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Los Altos, California: Morgan Kaufmann [An extended account of Bayesian belief updating, including its relation to nonmonotonic logic]

Peng, Y and Reggia, JA, 1986. "Plausibility of diagnostic hypotheses: The nature of simplicity" In: *Proceedings of the 6th National Conference on Artificial Intelligence*, 140–145, American Association for Artificial Intelligence [Describes the incorporation of probabilistic reasoning into the Generalized Set Covering model of diagnosis]

Poole, D, 1988a. "A logical framework for default reasoning" *Artificial Intelligence* **36** 27–47 [Proposes an account of nonmonotonic reasoning in terms of scenarios, and relates it to default logic. This forms the basis of the Theorist framework]

Poole, D, 1988b. "Representing knowledge for logic-based diagnosis" In: *Proceedings of the International Conference on Fifth Generation Computer Systems*, Tokyo, Japan [Argues that the Theorist framework can be used effectively to compare abduction, diagnosis from first principles, and rule-based diagnosis]

Poole, D, Goebel, R and Aleliunas, R, 1987. "Theorist: A logical reasoning system for defaults and diagnosis" Chapter 13 of Cercone, N and McCalla, G, eds, *The knowledge frontier*, New York: Springer-Verlag [A somewhat sketchy account of the Theorist framework]

Pople, HE, Jr, 1977. "The formation of composite hypotheses in diagnostic problem solving: An exercise in synthetic reasoning" In: *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pp 1030–1037, American Association for Artificial Intelligence [An early paper on abductive inference in AI; still worth reading today]

Provan, GM, 1988. "Solving diagnostic problems using extended truth maintenance systems" In: *Proceedings of the 8th European Conference on Artificial Intelligence*, pp 547–552, London: Pitman [An excellent paper that addresses the difficult problem of integrating truth maintenance and Dempster–Shafer belief functions. A more computational treatment than Laskey and Lehner]

Reiter, R, 1980. "A logic for default reasoning" *Artificial Intelligence* **31** 81–132 [A thorough and well-written account of default logic that has required little or no revision over the years]

Reiter, R, 1987a. "Nonmonotonic reasoning" *Annual Reviews of Computer Science* **2** 147–186 [A useful overview of nonmonotonic logic]

Reiter, R, 1987b. "A theory of diagnosis from first principles" *Artificial Intelligence* **32** 57–95 [A clear theoretical account of diagnoses based on the minimization of abnormality in a system description]

Shoham, Y, 1988. *Reasoning about change: Time and causation from the standpoint of artificial intelligence*, Cambridge, Massachusetts: MIT Press [An attempt to integrate nonmonotonic logic and temporal logic for reasoning about change]

Singh, N, 1987. *An artificial intelligence approach to test generation*, Norwell, Massachusetts: Kluwer Academic [Describes some further work on the DART system]

Smith, BA, 1988. "A system for the diagnosis of faults using a first principles approach" PhD thesis, Department of Computer Science, University of Missouri-Rolla [An interesting account of an implementation of Reiter's theory of diagnosis. Also contains excellent review chapters]

Swartout, WR, 1983. "XPLAIN: a system for creating and explaining expert consulting programs" *Artificial Intelligence* **21** 285–325 [Describes the derivation of an expert system from a domain model by automatic programming]

Winslett, M, 1988. "Reasoning about action using a possible models approach" In: *Proceedings of the 7th National Conference on Artificial Intelligence*, pp. 89–93, American Association for Artificial Intelligence [Contains a convincing critique of Ginsberg's construction for counterfactual reasoning]