# Characterizing Diagnoses

**Johan de Kleer**
Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto CA 94304 USA
**Alan K. Mackworth**[1]
University of British Columbia
Vancouver, B.C. V6T 1W5, Canada
**Raymond Reiter**[1]
University of Toronto
Toronto, Ontario M5S 1A4, Canada

## Abstract

Most approaches to model-based diagnosis describe a diagnosis for a system as a set of failing components that explains the symptoms. In order to characterize the typically very large number of diagnoses, usually only the minimal such sets of failing components are represented. This method of characterizing all diagnoses is inadequate in general, in part because not every superset of the faulty components of a diagnosis necessarily provides a diagnosis. In this paper we analyze the notion of diagnosis in depth exploiting the notions of implicate/implicant and prime implicate/implicant. We use these notions to propose two alternative approaches for addressing the inadequacy of the concept of minimal diagnosis. First, we propose a new concept, that of kernel diagnosis, which is free of the problems of minimal diagnosis. Second, we propose to restrict the axioms used to describe the system to ensure that the concept of minimal diagnosis is adequate.

## 1 Introduction

The diagnostic task is to determine why a correctly designed system is not functioning as it was intended — the explanation for the faulty behavior being that the particular system under consideration is at variance in some way with its design. One of the main subtasks of diagnosis is to determine what could be wrong with a system given the observations that have been made.

Most approaches to model-based diagnosis [4] characterize all the diagnoses for a system as the minimal sets of failing components which explain the symptoms. Although this method of characterizing diagnoses is adequate for diagnostic approaches which model only the correct behavior of components, it does not generalize. For example, it does not necessarily extend to approaches which incorporate models of faulty behavior [24] or which incorporate strategies for exonerating components [19]. In particular, not every superset of the faulty components of a diagnosis necessarily provides a diagnosis. In this paper we analyze the notion of diagnosis in depth and propose two approaches for addressing the inadequacy of minimal diagnoses. First, we propose an alternative notion, that of kernel diagnosis, which is free of the problems of minimal diagnosis. Second, we propose to restrict the axioms used to describe the system to ensure that the concept of minimal diagnosis is adequate.

The extended version of this paper [10] expands on the results, includes proofs for all the theorems, develops restrictions on the system description that allow the use of minimal diagnosis, and uses the approach to analyze current model-based diagnostic systems in more detail.

## 2 Problems with minimal diagnosis

Insofar as possible we follow Reiter's [20] framework.

**Definition 1** *A system is a triple (SD,COMPS,OBS) where:*

1. *SD, the system description, is a set of first-order sentences.*
2. *COMPS, the system components, is a finite set of constants.*
3. *OBS, a set of observations, is a set of first-order sentences.*

Most model-based diagnosis papers [7; 8; 13; 19; 20; 24] define a diagnosis to be a set of failing components with all other components presumed to be behaving normally. We represent a diagnosis as a conjunction which explicitly indicates whether each component is normal or abnormal. This representation of diagnosis captures the same intuitions as the previous definitions but generalizes more naturally.

We adopt Reiter's [20] convention that $AB(c)$ is a literal which holds when component $c$ ∈COMPS is behaving abnormally. (Some of the model-based diagnosis literature uses $\neg OK(c)$ instead of $AB(c)$ but this is just terminology and does not affect the results of this paper.) Depending on the exact definition of fault for the diagnostic task being addressed, abnormality will mean something different. This is reflected in how $AB$ is used in the sentences of SD. For example, in GDE [7], being

[1] Fellow, Canadian Institute for Advanced Research.

abnormal does not restrict the possible behaviors in any way since $AB$ only appears in the form $\neg AB(x) \rightarrow M$ where $M$ is the correct behavior of component $x$. In [19] being abnormal means that component behavior necessarily deviates from correct behavior since AB only appears in the form $\neg AB(x) \equiv M$.

**Definition 2** *Given two sets of components $C1$ and $C2$ define $\mathcal{D}(C1, C2)$ to be the conjunction:*

$$\Big[\bigwedge_{c \in C1} AB(c)\Big] \wedge \Big[\bigwedge_{c \in C2} \neg AB(c)\Big].$$

A diagnosis is a sentence describing one possible state of the system, where this state is an assignment of the status normal or abnormal to each system component.

**Definition 3** *Let $\Delta \subseteq COMPS$. A diagnosis for (SD,COMPS,OBS) is $\mathcal{D}(\Delta, COMPS - \Delta)$ such that $SD \cup OBS \cup \{\mathcal{D}(\Delta, COMPS - \Delta)\}$ is satisfiable.*

The following important observation follows directly from the definition (similar to proposition 3.1 of [20]):

**Remark 1** *A diagnosis exists for (SD,COMPS,OBS) iff $SD \cup OBS$ is satisfiable.*

Unfortunately, there may be $2^{|COMPS|}$ diagnoses. Therefore we seek a parsimonious characterization of the diagnoses of a system.

**Definition 4** *A diagnosis $\mathcal{D}(\Delta, COMPS - \Delta)$ is a minimal diagnosis iff for no proper subset $\Delta'$ of $\Delta$ is $\mathcal{D}(\Delta', COMPS - \Delta')$ a diagnosis.*

Thus a minimal diagnosis is determined by a minimal set of components which can be assumed to be faulty, while assuming the remaining components are functioning normally.

Note that these definitions subsume Reiter's [20]. Reiter's definition of the concept of diagnosis corresponds to our notion of *minimal* diagnosis. Reiter provides no definition corresponding to our notion of a diagnosis. All the results of [20] therefore apply to our concept of a minimal diagnosis.

The following is an easy consequence of the above definitions:

**Remark 2** *If $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis, then there is a minimal diagnosis $\mathcal{D}(\Delta', COMPS - \Delta')$ such that $\Delta' \subseteq \Delta$.*

Most previous approaches to model-based diagnosis have assumed that the converse holds, i.e., if $\mathcal{D}(\Delta', COMPS - \Delta')$ is a minimal diagnosis and if $\Delta' \subseteq \Delta$, then $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis. However, as we relax the commonly made assumptions, for example by allowing fault models or exoneration axioms, the converse fails to hold and we must explore alternative means for parsimoniously characterizing all diagnoses.

**Remark 3** *If $\mathcal{D}(\Delta', COMPS - \Delta')$ is a minimal diagnosis and $\Delta' \subset \Delta$, then $\mathcal{D}(\Delta, COMPS - \Delta)$ need not be a diagnosis.*

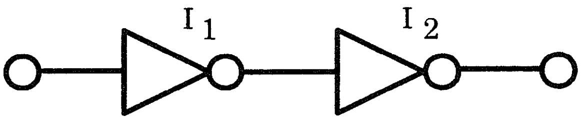


Figure 1: Two inverters

Thus, not every superset of the faulty components of a minimal diagnosis need provide a diagnosis. To see why, consider the following two simple examples. The first example arises if we presume we know all the possible ways a component can fail such as in [24].

**Example 1** Consider the simple two inverter circuit of Fig. 1. If we are considering making observations at different times, then we must represent this in SD in some way. One scheme is to introduce observation time $t$ as a parameter. Thus the model for an inverter is:
$INVERTER(x) \rightarrow$
$\neg AB(x) \rightarrow [in(x,t) = 0 \equiv out(x,t) = 1]$.
We assume that SD is extended with the appropriate axioms for binary arithmetic, etc. Suppose the input is 0 and the output is 1: $in(I_1, T_0) = 0, out(I_2, T_0) = 1$. There are three possible diagnoses: $AB(I_1) \wedge \neg AB(I_2)$, $AB(I_2) \wedge \neg AB(I_1)$ and $AB(I_1) \wedge AB(I_2)$; these are characterized by the first two diagnoses, which are minimal. Suppose we know that the inverters we are using have only two failure modes: they short their output to their inputs or their output becomes stuck at 0. We model this as:

$$INVERTER(x) \wedge AB(x) \rightarrow [SA0(x) \vee SHORT(x)],$$

$$SA0(x) \rightarrow out(x,t) = 0,$$

$$SHORT(x) \rightarrow out(x,t) = in(x,t).$$

From these models we can infer that it is no longer possible that both $I_1$ and $I_2$ are faulted. Intuitively, if $I_2$ is faulted and producing the observed 1, then it cannot be stuck at 0, and must have its input shorted to its output. But then $I_1$ must be outputting a 1 and there is no faulty behavior of $I_1$ which produces a 1 for an input of 0. Thus, $AB(I_1) \wedge AB(I_2)$ is no longer a diagnosis, but the minimal diagnoses (remain) unchanged.

The only way to determine which of $I_1$ or $I_2$ is actually faulted is to make additional observations. For example, if we observed $out(I_1, T_0)$, we could distinguish whether $I_1$ or $I_2$ is faulted. Suppose $I_1$ is faulted such that $out(I_1, T_0) = 0$. To identify the actual failure mode of $I_1$ we have to observe $out(I_1, T_1)$ or $out(I_2, T_1)$ given $in(I_1, T_1) = 1$.

This example shows that the use of exhaustive fault models such as in [24] leads to difficulties with the usual definition of diagnosis. One way to avoid this difficulty is not to presume all the faulty behaviors are known as in [8]. However, if we do not know all the faulty behaviors, then nothing useful can ever be inferred from a component being abnormal which defeats the purpose of fault modes in the first place (this is addressed in [8]) by introducing probabilities).

**Example 2** The usual definition of diagnosis encounters similar difficulties with the TRIAL framework of [19]. In this framework a component is considered faulty if it is actually manifesting a faulty behavior given the current set of inputs. If we are only concerned with one set of inputs, then every component is modeled as a biconditional. Thus, the inverters of Fig. 1 are instead described by:

$$INVERTER(x) \rightarrow \neg AB(x) \equiv [in(x) = 0 \equiv out(x) = 1].$$

Suppose the input and output are measured to be 0. There are only two diagnoses (the second of which is minimal):

$$AB(I_1) \wedge AB(I_2), \quad \neg AB(I_1) \wedge \neg AB(I_2).$$

It is not possible that one inverter is faulted and the other not. Each inverter exonerates the other. In terms of [19], each inverter is an alibi for the other. Thus, although $\neg AB(I_1) \wedge \neg AB(I_2)$ is a minimal diagnosis, neither $\neg AB(I_1) \wedge AB(I_2)$ nor $AB(I_1) \wedge \neg AB(I_2)$ are diagnoses. Again, we see that by including axioms which restrict faulty behavior in any way, the usual definition of diagnosis is inadequate to characterize all diagnoses.

In the remainder of this paper we explore two approaches to address this problem: (1) find an alternative means to characterize all diagnoses, and (2) restrict the form of $SD \cup OBS$ such that the notion of minimal diagnosis does characterize all diagnoses. We first require some preliminaries.

## 3 Minimal diagnoses

The minimal diagnoses are conveniently defined in terms of the familiar [17] notions of implicates and implicants (see [16; 21] for similar uses of these notions).

**Definition 5** *An AB-literal is $AB(c)$ or $\neg AB(c)$ for some $c \in COMPS$.*

**Definition 6** *An AB-clause is a disjunction of AB-literals containing no complementary pair of AB-literals. A positive AB-clause is an AB-clause all of whose literals are positive.*

Note that the empty clause is considered a positive AB-clause.

**Definition 7** *A conflict of (SD,COMPS,OBS) is an AB-clause entailed by $SD \cup OBS$. A positive conflict is a conflict all of whose literals are positive.*

If $SD \cup OBS$ is propositional, then a conflict is any AB-clause which is an implicate of $SD \cup OBS$.

The conflicts provide an intermediate step in determining the diagnoses and are central to many diagnostic frameworks. The reason for this can be understood intuitively as follows. The diagnostic task is to determine malfunctions, and therefore the primary source of diagnostic information about a system are the discrepancies between expectations and observations. A conflict represents such a fragment of diagnostic information. For example, the conflict $AB(A) \vee AB(B)$ might result from the discrepancy between observing $x = 1$ while expecting it to be 2, if components $A$ and $B$ were normal. As a consequence, we infer that at least one of $A$ or $B$ is abnormal, i.e., the conflict $AB(A) \vee AB(B)$. Most researchers have focused only on positive conflicts. (As most previous research has focused on the positive conflicts, they usually represented conflicts as sets of abnormal components.) However, as we see in Section 4, the non-positive conflicts are important when modeling faults and doing exoneration.

**Remark 4** *A diagnosis exists for (SD,COMPS,OBS) iff the empty clause is not a conflict of (SD,COMPS,OBS).*

**Theorem 1** *Suppose (SD,COMPS,OBS) is a system, $\Pi$ is its set of conflicts, and $\Delta \subseteq COMPS$. Then $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis iff $\Pi \cup \{\mathcal{D}(\Delta, COMPS - \Delta)\}$ is satisfiable.*

**Definition 8** *A minimal conflict of (SD,COMPS,OBS) is a conflict no proper subclause of which is a conflict of (SD,COMPS,OBS).*

Thus, if $SD \cup OBS$ is propositional, then a minimal conflict is any AB-clause which is a prime implicate of $SD \cup OBS$.

**Theorem 2** *Suppose (SD,COMPS,OBS) is a system, $\Pi$ is its set of minimal conflicts, and $\Delta \subseteq COMPS$. Then $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis iff $\Pi \cup \{\mathcal{D}(\Delta, COMPS - \Delta)\}$ is satisfiable.*

**Remark 5** *If all the minimal conflicts of (SD,COMPS, OBS)are non-empty and positive, then $\mathcal{D}(COMPS, \{\})$ is a diagnosis.*

As the minimal conflicts determine the diagnoses, they play a central role in most diagnostic frameworks.

**Example 3** Consider the familiar circuit of Fig. 2. Suppose the component models are:

$$ADDER(x) \rightarrow [\neg AB(x) \rightarrow out(x) = in1(x) + in2(x)]$$

$$MULTIPLIER(x) \rightarrow [\neg AB(x) \rightarrow out(x) = in1(x) \times in2(x)].$$

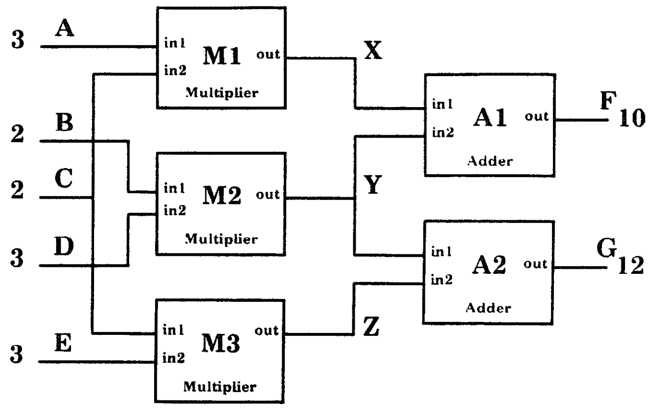
As before we assume that SD is extended with the appropriate axioms for arithmetic, etc. With the given inputs, there are two minimal conflicts:

$$AB(A_1) \vee AB(M_1) \vee AB(M_2),$$

$$AB(A_1) \vee AB(M_1) \vee AB(M_3) \vee AB(A_2),$$

and four familiar minimal diagnoses:

$\mathcal{D}(\{A_1\}, \{A_2, M_1, M_2, M_3\})$ :
$AB(A_1) \wedge \neg AB(A_2) \wedge \neg AB(M_1) \wedge \neg AB(M_2) \wedge \neg AB(M_3)$



Char-Diag > Polybox

Figure 2: $F = AC + BD, G = CE + BD$

$\mathcal{D}(\{M_1\}, \{A_1, A_2, M_2, M_3\})$ :
$AB(M_1) \wedge \neg AB(A_1) \wedge \neg AB(A_2) \wedge \neg AB(M_2) \wedge \neg AB(M_3)$

$\mathcal{D}(\{M_2, M_3\}, \{A_1, A_2, M_1\})$ :
$AB(M_2) \wedge AB(M_3) \wedge \neg AB(A_1) \wedge \neg AB(A_2) \wedge \neg AB(M_1)$

$\mathcal{D}(\{A_2, M_2\}, \{A_1, M_1, M_3\})$ :
$AB(A_2) \wedge AB(M_2) \wedge \neg AB(A_1) \wedge \neg AB(M_1) \wedge \neg AB(M_3)$.

**Definition 9** *A conjunction C of literals covers a conjunction D of literals iff every literal of C occurs in D.*

**Definition 10** *Suppose $\Sigma$ is a set of propositional formulas. A conjunction of literals $\pi$ containing no pair of complementary literals is an implicant of $\Sigma$ iff $\pi$ entails each formula in $\Sigma$. $\pi$ is a prime implicant of $\Sigma$ iff the only implicant of $\Sigma$ covering $\pi$ is $\pi$ itself.*

**Theorem 3** *(Characterization of minimal diagnoses) $\mathcal{D}(\Delta, COMPS - \Delta)$ is a minimal diagnosis of (SD,COMPS,OBS) iff $\bigwedge_{c \in \Delta} AB(c)$ is a prime implicant of the set of positive minimal conflicts of (SD,COMPS,OBS).*

This theorem underlies many model-based diagnostic algorithms. The first step, conflict recognition, finds positive minimal conflicts, and the second step, candidate generation, finds prime implicants. Clearly, if we were only interested in minimal diagnoses, then we would only be interested in identifying the positive minimal conflicts, but, in general, we must consider the nonpositive minimal conflicts as well.

We now have the machinery to state precisely when the minimal diagnoses characterize all diagnoses.

**Theorem 4** *The following are equivalent:*

1. *If $\mathcal{D}(\Delta', COMPS - \Delta')$ is a minimal diagnosis for (SD,COMPS,OBS), then $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis for (SD,COMPS,OBS) for every $\Delta$ such that $COMPS \supseteq \Delta \supseteq \Delta'$ (i.e., every superset of the faulty components of a minimal diagnosis provides a diagnosis).*
2. *All minimal conflicts of (SD,COMPS,OBS) are positive.*

In Example 1, $AB(I_1) \wedge \neg AB(I_2)$ was a diagnosis, but $AB(I_1) \wedge AB(I_2)$, which has more faulty components, was not. By theorem 4 this must arise because one of the minimal conflicts is not positive. In this example, the negative clause, $\neg AB(I_1) \vee \neg AB(I_2)$, is a minimal conflict, which follows directly from the fault models of $I_1$ and $I_2$.

## 4 Partial diagnoses

Suppose we have the following two diagnoses for a three component system: $AB(c_1) \wedge AB(c_2) \wedge AB(c_3)$ and $AB(c_1) \wedge AB(c_2) \wedge \neg AB(c_3)$. We can interpret this as saying that $c_1$ and $c_2$ are faulty, and that $c_3$ may or may not be faulty. Thus, the two diagnoses may be represented more compactly by $AB(c_1) \wedge AB(c_2)$. In fact, we can view this as a 'partial' diagnosis in which we are uncommitted to the status of $c_3$; no matter what that status is, it leads to a diagnosis. This is the basis for Poole's observation [18] that a diagnosis need not commit to a status for each component whenever that status is a 'don't care'. Accordingly, we introduce the concept of a partial diagnosis. This concept also has the nice side effect of providing a convenient representation characterizing the set of all diagnoses.

**Definition 11** *A partial diagnosis for (SD,COMPS, OBS) is a satisfiable conjunction P of AB-literals such that for every satisfiable conjunction of AB-literals $\phi$ covered by P, $SD \cup OBS \cup \phi$ is satisfiable.*

The following is an easy consequence of this definition:

**Remark 6** *If P is a partial diagnosis of (SD,COMPS, OBS) and C is the set of all components mentioned in P, then $P \wedge \bigwedge_{c \in COMPS - C} A(c)$ is a diagnosis, where each $A(c)$ is $AB(c)$ or $\neg AB(c)$.*

Thus, a partial diagnosis $P$ represents the set of all diagnoses which contain $P$ as.a subconjunct. It is natural then to consider the minimal such $P$'s, which we call kernel diagnoses.

**Definition 12** *A kernel diagnosis is a partial diagnosis with the property that the only partial diagnosis which covers it is itself.*

The following easy result provides exactly the characterizing property we have been looking for:

**Theorem 5** *(Characterization of diagnoses) $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis iff there is a kernel diagnosis which covers it.*

Consider the example of Fig. 1. Without the introduction of fault models there were three diagnoses: $AB(I_1) \wedge \neg AB(I_2)$, $\neg AB(I_1) \wedge AB(I_2)$, $AB(I_1) \wedge AB(I_2)$ which are characterized by the two kernel diagnoses: $AB(I_1)$ and $AB(I_2)$. With the addition of the fault models, the kernel diagnoses become: $AB(I_1) \wedge \neg AB(I_2)$ and $\neg AB(I_1) \wedge AB(I_2)$.

Partial and kernel diagnoses can be particularly easily characterized in terms of prime implicants and minimal conflicts. Recall that a conjunction of literals $\pi$ containing no pair of complementary literals is an implicant of $\Sigma$ iff $\pi$ entails each formula in $\Sigma$.

**Theorem 6** *The partial diagnoses of (SD,COMPS, OBS) are the implicants of the minimal conflicts of (SD,COMPS,OBS).*

**Corollary 1** *(Characterization of kernel diagnoses) The kernel diagnoses of (SD,COMPS,OBS) are the prime implicants of the minimal conflicts of SD $\cup$ OBS.*

As a consequence of this corollary and theorem 3, if all minimal conflicts are positive, then there is a simple one-to-one correspondence between minimal diagnoses and kernel diagnoses.

Corollary 1 provides a direct way of computing the kernel diagnoses. One way of doing this is to convert the CNF-form of the minimal conflicts to DNF and simplify as follows (we omit the proof):

1. 'Multiply' the minimal conflicts to give a disjunction of conjunctions.
2. Delete any conjunction containing a complementary pair of literals.
3. Delete any conjunction covered by some other conjunction.
4. The remaining conjunctions are the prime implicants of the original minimal conflicts, and hence the kernel diagnoses.

**Example 4a** Consider Example 3. There are two minimal conflicts:

$$AB(A_1) \vee AB(M_1) \vee AB(M_2),$$

$$AB(A_1) \vee AB(M_1) \vee AB(M_3) \vee AB(A_2),$$

and four kernel diagnoses:

$$AB(A_1), \quad AB(M_1),$$

$$AB(M_2) \wedge AB(M_3), \quad AB(M_2) \wedge AB(A_2).$$

As all minimal conflicts are positive, these diagnoses correspond one-to-one to the familiar minimal diagnoses.

**Example 4b** If we considered a component to be faulted only if it manifested a faulty behavior under the current set of inputs (such as in [19]), then we would use slightly different component models:

$$ADDER(x) \rightarrow \left[\neg AB(x) \equiv [out(x) = in1(x) + in2(x)]\right]$$

$$MULTIPLIER(x) \rightarrow$$
$$\neg AB(x) \equiv [out(x) = in1(x) \times in2(x)].$$

In this case the minimal conflicts become:

$$AB(A_1) \vee AB(M_1) \vee AB(M_2),$$

$$AB(A_1) \vee AB(A_2) \vee AB(M_1) \vee AB(M_3),$$

$$AB(A_2) \vee \neg AB(M_2) \vee AB(M_3),$$

$$AB(A_2) \vee AB(M_2) \vee \neg AB(M_3)$$

$$\neg AB(A_2) \vee AB(M_3) \vee AB(M_2),$$

and the kernel diagnoses become:

$$\neg AB(A_2) \wedge AB(M_1) \wedge \neg AB(M_2) \wedge \neg AB(M_3),$$

$$AB(A_2) \wedge AB(M_1) \wedge AB(M_3),$$

$$AB(A_1) \wedge \neg AB(A_2) \wedge \neg AB(M_2) \wedge \neg AB(M_3),$$

$$AB(A_1) \wedge AB(A_2) \wedge AB(M_3),$$

$$AB(A_2) \wedge AB(M_2), \quad AB(M_2) \wedge AB(M_3).$$

Note that because the positive minimal conflicts are unchanged, the set of minimal diagnoses remains unchanged.

In this example there are only a few more kernel diagnoses than minimal diagnoses (6 vs. 4). However, one possible disadvantage of this approach is that there may be exponentially more kernel diagnoses than diagnoses.

It is interesting to note that the set of minimal conflicts may be redundant. In Example 4b, the first and third minimal conflicts entail the second:

$$\begin{array}{l} AB(A_1) \vee AB(M_1) \vee AB(M_2) \\ AB(A_2) \vee \neg AB(M_2) \vee AB(M_3) \\ \hline AB(A_1) \vee AB(A_2) \vee AB(M_1) \vee AB(M_3) \end{array}$$

Therefore, the second minimal conflict is redundant. Such redundancy can only occur if there are non-positive minimal conflicts. Unfortunately, these observations do not seem to be of much practical use because there is no easy way to tell whether there are enough minimal conflicts without first finding them all.

**Definition 13** *A set of kernel diagnoses is irredundant iff it is a smallest cardinality set with the property that every diagnosis is covered by at least one of its elements.*

**Theorem 7** *If all minimal conflicts are positive there is exactly one irredundant set of kernel diagnoses, namely the set of minimal diagnoses.*

Note that a system can have multiple irredundant sets of kernel diagnoses.

**Example 5** Consider a circuit having three components A, B, C and the two minimal conflicts:

$$AB(A) \vee AB(B) \vee AB(C), \quad \neg AB(A) \vee \neg AB(B) \vee \neg AB(C)$$

These have six prime implicants (i.e., kernel diagnoses).

$$AB(A) \wedge \neg AB(B), \quad \neg AB(A) \wedge AB(C), \quad AB(B) \wedge \neg AB(C),$$

$$\neg AB(A) \wedge AB(B), \quad AB(A) \wedge \neg AB(C), \quad \neg AB(B) \wedge AB(C).$$

There are two irredundant sets of kernel diagnoses:

$$\{AB(A) \wedge \neg AB(B), \neg AB(A) \wedge AB(C), AB(B) \wedge \neg AB(C)\}$$
$$\{\neg AB(A) \wedge AB(B), AB(A) \wedge \neg AB(C), \neg AB(B) \wedge AB(C)\}.$$

Our analysis of kernel diagnoses corresponds exactly to the classical analysis in switching theory of so-called two level minimization of boolean functions (e.g., the Quine-McCluskey algorithm [14; 17]). The problem there is to synthesize a circuit realizing a given function as a disjunction of conjunctions of literals in such a way as to minimize the number of and-, or- and not-gates. Such circuits are characterized by irredundant sets of prime implicants of the given function. In the case of diagnosis, the given boolean function is specified by $\Pi$, the set of conflicts of $SD \cup OBS$. The kernel diagnoses are the prime implicants of $\Pi$, and the minimal sets of kernel diagnoses sufficient to cover every diagnosis are the irredundant sets of prime implicants of $\Pi$. It is well known from switching theory experience that the minimization problem is computationally intractable; there may be too many prime implicants, and even if there aren't, finding an irredundant subset of them is NP-hard. Designers of VLSI circuits have developed various approximation techniques [1]. Because of the exact correspondence with diagnosis, we can expect to profit from these techniques.

## 5 Prime diagnoses

Raiman [19] proposes a notion of prime diagnosis to characterize diagnoses. In his TRIAL architecture components are individually incriminated and exonerated. Therefore, he characterizes the diagnoses of a system in terms of the diagnoses involving its individual components. The following is a generalization of his definition.

**Definition 14** *Given (SD,COMPS,OBS), a prime diagnosis for $c \in COMPS$ is a minimal diagnosis for $(SD,COMPS,OBS \cup \{AB(c)\})$*

Prime diagnoses characterize all diagnoses as follows.

**Theorem 8 (Raiman)** *Suppose $\mathcal{D}(\Delta, COMPS - \Delta)$ is a diagnosis. Then for each $c_i \in \Delta$ there is a prime diagnosis $\mathcal{D}(\Delta_i, COMPS - \Delta_i)$ for $c_i$ such that $\Delta = \bigcup_i \Delta_i$.*

Unfortunately, Example 1 shows that not every union leads to a diagnosis. The prime diagnoses are:

$$P(I_1) = \{AB(I_1) \wedge \neg AB(I_2)\},$$
$$P(I_2) = \{AB(I_2) \wedge \neg AB(I_1)\}.$$

However, $AB(I_1) \wedge AB(I_2)$ is not a diagnosis. Thus, prime diagnoses are inadequate to characterize diagnoses.

Raiman [19] implicitly assumes all minimal conflicts contain at most one negative literal. In this case Raiman shows that the converse of Theorem 8 holds which makes prime diagnoses adequate for characterizing diagnoses. This useful property holds if $SD \cup OBS$ is horn, but we do not know of any more general practical condition on $SD \cup OBS$ which ensures it.

## 6 Restricting the system description

Our overall objective is to find methods of characterizing all diagnoses. We saw that minimal diagnoses were inadequate for this task in general and we examined kernel and prime diagnoses as alternatives. Another approach is to restrict the form of the system such that minimal diagnoses do characterize all diagnoses. We know from Theorem 4 that a necessary and sufficient condition ensuring that every superset of the faulty components of a minimal diagnosis provides a diagnosis is that all minimal conflicts be positive. Unfortunately, we are not aware of any simple necessary and sufficient condition on the syntactic form of a system which ensures that all minimal conflicts are positive. Clearly both $OBS$ and $SD$ need to be restricted because definition 1 allows non-positive AB-clauses to be part of $OBS$ and $SD$. In the extended paper we explore some commonly used practical restrictions on $OBS$ and $SD$ that suffice to ensure that the minimal diagnoses are adequate to characterize all diagnoses.

## 7 Summary

The notions of minimal and prime diagnosis are inadequate to characterize diagnoses generally. We argue that the notion of kernel diagnosis which designates some components as normal, others abnormal, and the remainder as being either, is a better way to characterize diagnoses. We avoid significant complexity if kernel diagnoses contain only positive literals (i.e., all minimal conflicts are positive). This can be achieved by limiting the description of the system to ensure this. Most current model-based techniques take this approach[10].

There are usually a large number of minimal conflicts and kernel diagnoses (or minimal diagnoses). Therefore, the brute-force application of the techniques suggested in this paper is not practical. The contribution of this paper is that it provides a clear logical framework for characterizing the space of diagnoses in the general case. It thus provides the specification for an ideal diagnostician. In practice, some focusing strategy must be brought to bear. One approach is to exploit hierarchical information as in [13]. Another approach is to focus the reasoning to identify the most relevant conflicts in order to find the most probable diagnoses [8; 11]. However, both of these approaches require additional information: the structural hierarchy and probabilistic information.

## 8 Acknowledgments

The contents of this paper benefitted from many discussions with Olivier Raiman. Daniel G. Bobrow, Brian C. Williams, Vijay Saraswat and Jeffrey Siskind provided extensive insights on early drafts.

## References

[1] Brayton, R.K., Hachtel, G.D., McMullen, C.T. and Sangiovanni-Vincentelli, A.L., *Logic minimization algorithms for VLSI Synthesis*, (Kluwer, 1984).

[2] Brown, J.S., Burton, R. R. and de Kleer, J., Pedagogical, natural language and knowledge engineering techniques in SOPHIE I, II and III, in: D. Sleeman and J.S. Brown (Eds.), *Intelligent Tutoring Systems*, (Academic Press, New York, 1982) 227–282.

[3] Davis, R., Diagnostic Reasoning based on structure and behavior, *Artificial Intelligence* **24** (1984) 347-410.

[4] Davis, R., and Hamscher, W., Model-based reasoning: Troubleshooting, in *Exploring artificial intelligence*, edited by H.E. Shrobe and the American Association for Artificial Intelligence, (Morgan Kaufmann, 1988), 297–346.

[5] de Kleer, J., An assumption-based truth maintenance system, *Artificial Intelligence* **28** (1986) 127–162. Also in *Readings in NonMonotonic Reasoning*, edited by Matthew L. Ginsberg, (Morgan Kaufmann, 1987), 280–297.

[6] de Kleer, J., Extending the ATMS, *Artificial Intelligence* **28** (1986) 163–196.

[7] de Kleer, J. and Williams, B.C., Diagnosing multiple faults, *Artificial Intelligence* **32** (1987) 97-130. Also in *Readings in NonMonotonic Reasoning*, edited by Matthew L. Ginsberg, (Morgan Kaufmann, 1987), 372–388.

[8] de Kleer, J. and Williams, B.C, Diagnosis with behavioral modes, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 1324–1330.

[9] de Kleer, J., A comparison of ATMS and CSP techniques, *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI (August 1989) 290–296.

[10] de Kleer, J., Mackworth, A.K. and Reiter, R., Characterizing Diagnoses and Systems, SSL Paper P89-00193, Xerox PARC, 1990. Also available as University of British Columbia Department of Computer Science TR90-8.

[11] Dressler, O., and Farquhar, A., Focusing ATMS-based problem solvers, Siemens Report INF-2-ARM 13, 1989.

[12] Genesereth, M.R., The use of design descriptions in automated diagnosis, *Artificial Intelligence* **24** (1984) 411–436.

[13] Hamscher, W.C., Model-based troubleshooting of digital systems, Artificial Intelligence Laboratory, TR-1074, Cambridge: M.I.T., 1988.

[14] Hill, F.J. and Peterson, G.R., *Introduction to Switching Theory and Logical Design* (John Wiley and Sons, New York, 1974).

[15] Kohavi, Z., *Switching and Finite Automata Theory* (McGraw-Hill, 1978).

[16] Kean, A. and Tsiknis, G., An incremental method for generating prime implicants/implicates, University of British Columbia Technical Report TR88-16, 1988.

[17] Kohavi, Z., *Switching and Finite Automata Theory* (McGraw-Hill, 1978).

[18] Poole, D., Representing knowledge for logic-based diagnosis, *Proc. Int. Conf. on Fifth Generation Computer Systems* (1988) 1282–1290.

[19] Raiman, O., Diagnosis as a trial: The alibi principle, IBM Scientific Center, 1989.

[20] Reiter, R., A theory of diagnosis from first principles, *Artificial Intelligence* **32** (1987) 57–95. Also in *Readings in Non-Monotonic Reasoning*, edited by Matthew L. Ginsberg, (Morgan Kaufmann, 1987), 352–371.

[21] Reiter, R. and J. de Kleer, Foundations of Assumption-Based Truth Maintenance Systems: Preliminary Report, *Proceedings of the National Conference on Artificial Intelligence*, Seattle, WA (July, 1987), 183–188.

[22] Slagle, J.R., C.L. Chang, and R.C.T. Lee, A new algorithm for generating prime implicants, *IEEE Transactions on Computers* **C-19**(4) (April 1970) 304–310.

[23] Struss, P., Extensions to ATMS-based Diagnosis, in: J.S. Gero (ed.), *Artificial Intelligence in Engineering: Diagnosis and Learning* (Elsevier, Amsterdam, 1988) 3–28.

[24] Struss, P., and Dressler, O., "Physical negation" — Integrating fault models into the general diagnostic engine, in: *Proceedings IJCAI-89* Detroit, MI (1989) 1318-1323.